Analysis of Non-imaging Data

Nicolas Duchateau, Oscar Camara, Rafael Sebastian, Andrew King

DOI = https://doi.org/10.1007/978-3-031-05071-8_10

10 Analysis of Non-imaging Data

Dr Nicolas Duchateau[a,b,*]

Prof Oscar Camara[c]

Prof Rafael Sebastian[d]

Dr Andrew King[e]

[a] *Univ Lyon, Université Claude Bernard Lyon 1, INSA-Lyon, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69621, Lyon, France.*

[b] *Institut Universitaire de France (IUF), France.*

[c] *BCN MedTech, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain.*

[d] *Department of Computer Science, Universitat de Valencia, Valencia, Spain.*

[e] *School of Biomedical Engineering and Imaging Sciences, King's College London, United Kingdom.*

[*] *Corresponding author.*

Authors' contribution:

- Main chapter: ND, OC, RS, AK.

Abstract

Whilst most of this book has focused on imaging data because of the key role it plays in cardiology, non-imaging data also has an important role to play. This chapter reviews some of the most relevant non-imaging data sources and how they can be used by AI to positively impact patient management. Electrophysiology data, electrocardiograms and electronic health records are all discussed in detail and potential and existing applications for artificial intelligence are discussed with practical examples.

Keywords:

non-imaging data, electrophysiology, EP, electrocardiogram, ECG, electronic health record, EHR

Learning Objectives:

At the end of this chapter you should be able to:

- *10.A Explain the potential role of AI in analysis of electrophysiology data*
- *10.B Describe some applications of AI-based analysis of electrocardiograms (ECGs) and outline some of the difficulties and challenges that must be addressed*
- *10.C Explain how AI can be used in the analysis and automated production of electronic health records (EHRs)*

Introduction

Imaging data play a central role in cardiology, and much of the recent research activity in AI for cardiology has focused on cardiac imaging. For this reason, imaging-based AI has been the main focus of this book. However, there are a range of other data sources that are of importance in clinical decision making in cardiology. In this chapter we review the most relevant of these, with a focus on the ways in which AI has been proposed for use to streamline and improve clinical workflows.

Electrophysiology

Cardiac electrophysiology deals with the diagnosis and treatment of the electrical function of the heart. In general, it involves the analysis of electrical phenomena by means of different sources of information such as the ECG, body surface potential maps (BSPMs), or the more invasive means of intracardiac catheter recordings. Its main area of work is the analysis and treatment of rhythm disorders (arrhythmias), which are managed by cardiac electrophysiologists, who acquire and analyze electrophysiology studies that aim to elucidate symptoms, evaluate abnormal ECGs and assess the risk of arrhythmias in the present and future. Among the different therapeutic options available for cardiac arrhythmia, we can highlight drug therapy, surgical implantation (pacemakers, implantable cardioverter-defibrilators or ICDs), and cardiac ablation (radiofrequency ablation, cryoablation). Due to the complexity to plan and optimize cardiac therapies, several novel approaches and technologies have grown in popularity during the last decades to aid electrophysiologists. Among them, it is worth mentioning precision cardiology that involves the construction of patient-specific representations of an individual heart to perform electrical simulations [93] (see Chapter 9, page 227). In the area of cardiac electrophysiology, the advent of machine learning is having a major impact at different levels in several applications, from the automatic interpretation of ECGs to basic research on arrhythmia mechanisms, both experimental and computational [389, 277, 301].

Precision Cardiology

The goal of precision cardiology is to come up with methods and tools that allow doctors to develop and provide personalized treatments to each in-

dividual, taking into account inter-individual variability. It is an innovative approach that aims at improving risk stratification and at identifying personalized management through targeted diagnostic and therapeutic strategies. This is perfectly represented by the concept of a 'digital twin' (see Chapter 9, page 227), which aims to define patient-specific virtual hearts that dynamically integrate the clinical data acquired over time for an individual combined with previous observations from experiments and multi-scale simulations [93]. Such a virtual model can be used to aid doctors to make diagnoses and prognoses, tailor treatments to individual patients and make predictions of patient health evolution [340]. Biophysical simulations are successful at integrating multiscale, multiphysics information with the aim of uncovering mechanisms that can explain functions [69]. For instance, a digital twin equipped with physics-based models could be used to predict the response of a patient to a specific medical device, such as a cardiac pacemaker, or even to personalize the configuration of the device to its particular anatomical and functional properties (ventricular wall morphology, location of coronary veins, presence and location of scar tissue). Although, at first glance, the relationship between machine learning and multiscale biophysical simulations does not seem obvious, they can benefit from each other in a number of applications [21], such as the integration of physics-based knowledge in the form of governing equations (learning the underlying physics), or constraints to manage ill-posed problems (e.g. electrocardiographic imaging (ECGI) inverse problems) [257] or handle sparse and noisy data [336]. Another important use of machine learning in precision cardiology is the definition of surrogate models that can predict the response of a complex biophysical model from a reduced number of clinical inputs. This is possible due to the ability of machine learning to reveal correlations between different features that can be exploited by biophysical models to, for instance, classify or stratify patients. Since creating a personalized model is time consuming and requires expert input and many different types of data, machine learning techniques such as transfer learning are good alternatives to make predictions in a fast and reliable way without the need to create a full detailed model from scratch [301].

Machine Learning in Cardiac Computational Modeling

Digital twins must include the particular properties of an individual, so that simulations on the model are able to predict outcome of antiarrhythmia treat-

ments, or stratify patients. To build a digital twin, the first step is to reconstruct the patient-specific 3-D anatomy of the patient's heart. For the case of the geometry of the atria and ventricles, the use of deep learning based methods, and the proliferation of some particular models, such as the U-Net [50] has opened up new possibilities to build detailed models from clinical data with very little user interaction. However, if one wants to incorporate other physiological properties into the model to be able to perform biophysical simulations of cardiac electrophysiology, many additional features have to be extracted from the patient's clinical records, imaging data, and electrophysiological measurements to personalize the model [243]. For instance, the underlying organization of cardiac tissue, so-called fiber orientation, that determines the principal direction of the depolarization wavefront in cardiac tissue has to be incorporated into the model, but this cannot be obtained *in vivo* using imaging techniques. Physics-informed neural networks (PINNs) have been developed to learn properties such as the fiber orientation from *in vivo* anatomical maps (e.g. FiberNet [163]). PINNs are variants of machine learning based methods that are used to solve inverse problems governed by partial differential equations, and do not typically need large amounts of labeled data to make accurate predictions thanks to the incorporation of physical laws into their loss functions [325]. Other studies have focused on personalizing parameters of a simplified electrical model, for example activation onset location and tissue conductivity from patients that presented premature ventricular contractions, using Kernel Ridge Regression [136]. In this work, the authors were able to personalize the cardiac electrophysiological model and predicted new patient-specific pacing conditions.

Biophysical simulations of the heart have also been used as tools to generate synthetic datasets that include detailed anatomical and electrical information to train machine learning systems for different applications [312, 123, 95, 139, 111].

Personalization of models to reproduce the electrical activation sequence of the heart is also an active area of research. Several sources of data have been employed to adapt the model to the patient, such as electro-anatomical maps (EAMs, acquired invasively with a catheter), to BSPMs and ECGs. EAMs are created sequentially by acquiring random discrete samples from different heart beats that are scattered all over the heart's endocardial cavity. As a result, EAMs often present large errors and inconsistencies that can affect the decision taken during the radiofrequency ablation intervention. Recently,

a PINN has been proposed (EikonalNet) to overcome these limitations, imposing wave propagation dynamics to the estimated EAM, and adding a quantification of the uncertainty [336]. This is possible thanks to the current understanding of the system, which could be used to constrain the design space using the known underlying wave propagation dynamics. In the same work, an active learning algorithm was proposed to guide the electrophysiologist in the data acquisition process during the intervention. Similar works have used machine learning to estimate the sequence of activation from motion patterns (using Kernel Ridge Regression) [313], or directly from images (using least-squares SVM) [312], since there exists a relationship between the electrical activation and mechanical contraction. Non-invasive ECGI has also been employed as a source of information to personalize cardiac electrophysiology models when combined with machine learning algorithms, such as the Time-Delay Artificial Neural Network (TDANN) [257], transfer learning [135] or support vector regression (SVR) [176].

Machine Learning in Cardiac Arrhythmia Mechanisms

Biophysical models can provide insight into the heart as a system at a high level of resolution and precision. They can systematically probe various pathological conditions and treatments, and they can do this faster, more cost effectively and go beyond what is experimentally possible. The massive datasets produced by these simulations are suited to machine learning analysis to uncover hidden relationships between parameters.

At the cellular level, machine learning has been employed in ion channel modeling to i) predict functional changes in channels due to mutations [88]; ii) identify the structure/function relationship in voltage potassium channels [228]; or find relationships between kinetic properties of ion channel recovery and dynamics of arrhythmias [217]. It is also worth mentioning its application to investigating drug cardiotoxicity by predicting hERG (ether-a-go-go-related gene) related cardiotoxicity of a given compound, which is a surrogate marker of pro-arrhythmic risk [402].

At the organ level, machine learning has been applied to the investigation of reentrant activity. In particular, Muimani et al. [275] developed a deep learning method (a CNN, see Chapter 3, page 74) for the detection of unbroken and broken spiral waves, which are analogs of life-threatening cardiac arrhythmias, and their efficient elimination by targeted delivery of low am-

plitude current. Other studies have focused on predicting the effect of the fibrosis density and entropy on the maintenance of reentrant drivers by using patient specific computational models of the atria and SVMs with second degree polynomial kernels [433].

Machine Learning in Therapy Guidance

The combination of machine learning and digital twin technology could also be a powerful tool for therapy guidance, with a large potential to be transferred to electrophysiology labs. Currently, most common arrhythmias are treated by catheter-based ablation, which destroys the ability of cardiac tissue to trigger and conduct electrical signals, and can stop several types of arrhythmias, such as ventricular tachycardia (VT) or atrial fibrillation (AF). An important area of study that combines biophysical modeling and machine learning has focused on predicting the location of arrhythmic sources, such as ectopic foci or rotor drivers, in the atria and ventricles. In [123] a SVM classifier was built to determine non-invasively from the virtual BSPM of a patient, the location (region based) of the ectopic focus that was triggering the atrial tachycardia, with an accuracy over 90%. Yang et al [424] used CNNs to detect the exit site of postinfarction VT on the basis of the 12-lead ECG, which was subsequently validated by computer simulations. In [149], the use of sequential factorized autoencoders (a type of deep CNN) was proposed to find the location of VT exit sites, taking into account differences in 12-lead ECG due to patient variability at electrical (source of VT) and anatomical (heart anatomy) levels.

Regarding ablation of AF, there has been a large number of studies that aim to predict ablation success or recurrence after ablation based on clinical recordings, which analyze the 12-lead ECG, patients' anatomy, or distribution of fibrosis. Computer simulations on patient-specific geometries including fibrosis segmented from LGE-MRI were conducted to pre-operatively predict recurrence of AF after ablation together with a machine learning based classifier [352].

Limitations

Although there are big expectations and optimism for the potential applications of machine learning techniques to physics-based modeling, it is important to be aware of its limitations. In general, machine learning techniques

identify correlations but are agnostic as to causality, while multiscale modeling can find causal mechanisms. Besides, it is very common to see cases in which machine learning systems do not generalize well, i.e. the system is not really learning from the samples, but memorizing them (i.e. the model overfits, see Chapter 2, page 33). In addition, in many studies it is assumed that the distributions of the training and the test data are the same, which may be not true. Finally, another recurrent problem in many cases is that there is class imbalance, i.e. a particular class is over represented compared to others.

ECG Analysis

Transition to the Digital Era

The electrocardiogram (ECG) is a central tool in the assessment of a patient's condition and their follow-up. It is non-invasive and inexpensive compared to other devices, available in a large variety of clinical environments and used by a large array of healthcare professionals with varying knowledge on cardiology, an important point which can hamper ECG interpretation. Over recent decades, computational techniques have substantially improved the quantification and analysis of ECG signals [368], and the use of machine learning has further increased the efficiency and robustness of these tasks [130, 166]. As access to data is key to developing high performing machine learning algorithms, the entrance of the field of ECG analysis into the digital era has clearly boosted the use of machine learning models, as visible from the publications registries[33] and continuously increasing industrial investment.

However, the route to the digital world is not straightforward for ECG data. Many hospitals still rely on paper-printed ECG records, which requires addressing a large amount of issues before their computational analysis: digitization of the printed records, extraction of the signals from the background, standardization of the traces, etc. Many efforts have been made to properly standardize the existing data, but still heterogeneity between the proposed formats hampers the interoperability of analysis tools [58, 391, 35]. Besides, even for a given data format, many differences can remain in the stored data, as illustrated in Figure 10.1. For example, the duration and number

[33]The query "ECG machine learning" in Pubmed returns 350+ papers for 2021 against around 75 and 20 papers ten and twelve years previously.

of cardiac cycles considered actually depends on the underlying disease and the type of acquisition (for example, in 12-lead ECG as opposed to Holter acquisitions). Given this context, it is evident that despite being 1-D, the computational analysis of ECG signals is not at all easier compared to 2-D images.

One also needs to remember that given the scarcity of large standardized databases of digital ECG signals, such computational analysis started much before the advent of machine learning with many efforts for community-based post-processing tools using standard signal processing [368]. Among these, popular methods largely relied on smart signal analysis (e.g. wavelet-based methods that are able to represent the multi-scale structure of signals) [259] and generic-but-relevant decisions (rule- or threshold-based, using relevant features extracted from the signals). Highly curated databases have now started to emerge (see database reviews in [248, 166, 366]) to drive the whole community around data analysis challenges [305], which open the path to applying machine learning models but also to compare them to common ground truth data.

Machine learning naturally has the potential to move this automated analysis forward, with methods better suited to the data under study. In the following, we will discuss how two main tasks of ECG analysis are handled: automatic feature extraction and automatic diagnosis. We will pay specific attention to issues that highly condition the performance of machine learning, such as the database size, the quality and variability of annotations as well as the interpretability of the results.

Automatic Quantification

A first task for the computational analysis of ECGs with machine learning consists of automatic quantification, namely the automatic extraction of features of interest in the signals. As discussed in the Clinical Introduction to Chapter 4 (see page 86), typical measurements from ECG signals consist of the onset/offset of each cycle, and complementary markers of the cardiac cycle such as the events of the QRS, P and T waves, and the duration of the cardiac phases that can be derived from these events, since they are biomarkers of different cardiac diseases (e.g. enlarged QRS as a surrogate of electrical dyssynchrony, or elevated ST segment for infarction). From a machine learning perspective, extracting these events can be formulated as

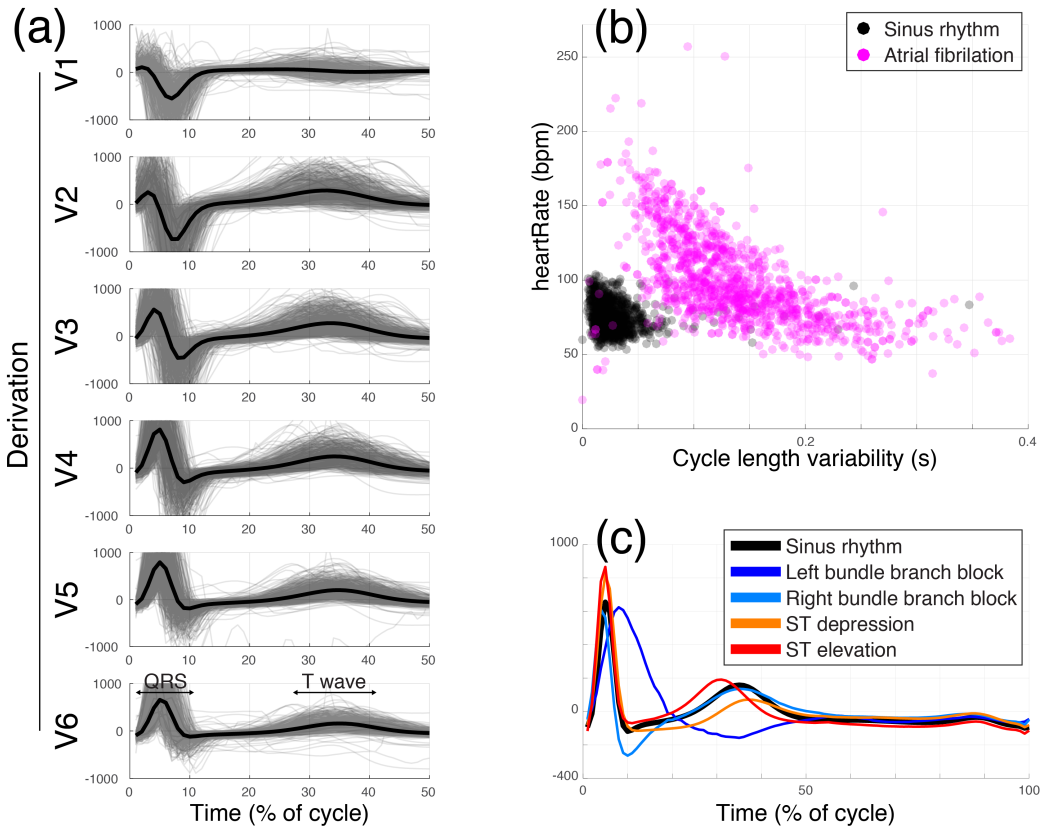

Figure 10.1: Variability in the ECG signals from the CPSC2018 database [238] used in the PhysioNet 2020 challenge [305], consisting of 12-derivation signals from 6877 subjects. For visualization purposes, all signals were temporally resampled to 100 instants with the beginning and end of the cycle normalized to 0 and 100%, respectively, and averaged across all the cycles of a given subject. (a) V1 to V6 derivations (half of the cycle) for the 917 subjects labeled as normal sinus rhythm (the thick black trace corresponds to the average of all signals). Despite all being "normal", we observe a large variability in the QRS complex amplitude, and in the timing and amplitude of the T wave. (b) Comparison of normal sinus rhythm and atrial fibrillation subjects based on two features extracted from the ECG signals using standard signal processing. Two clusters are easily visible, indicating that these two features may be enough to classify most subjects, but the presence of some subjects near the other cluster indicates that more advanced features or signal analyses are required to improve diagnosis. (c) Representative ECG (average across a subgroup, V6 derivation displayed) for five subgroups for which specific QRS and T wave changes are visible depending on the subgroup, motivating the use of a more sophisticated analysis of ECG patterns.

a supervised problem, where the training labels come from ECG signals in which the events have been annotated by experts. Naturally, their identification may be more or less challenging depending on the quality of the signals, the derivation, and the disease under study. Testing machine learning models of different complexities and increasing the database size and richness are ways to prevent this, although the latter may not be possible in all situations.

Although automatic diagnosis (discussed in the next subsection) attracts most of the attention in ECG analysis, several works have attempted to match or exceed the performance of standard ECG quantification methods by using machine learning. Convolutional neural networks (CNNs, see page 74) are attractive compared to fully-connected networks (FCNs, see page 88), as they use convolutions that both reduce the number of network connections (the number of parameters to optimize) and better take into account the structure of the input data (the spatial arrangement of pixels for images, and the temporal sequence of values for signals), as demonstrated on ECG data by, for example, [365, 62]. Inspired by its success in image segmentation tasks, a variant of the U-net architecture has been recently adapted for ECG quantification [178, 179, 274, 383]. Another branch of works has considered Recurrent Neural Networks (RNNs, see page 88), which are tailored for analyzing temporal sequences of data, and in particular the long short-term memory (LSTM) architecture, which partially addresses some computational issues of RNNs [10, 300].

Automatic Diagnosis

Once the features of interest have been extracted, these can be fed into subsequent models for the characterization of populations (e.g. examining statistical differences between two subgroups), or automatic diagnosis. In theory, as for 2-D images, neural networks may also offer an all-in-one approach that avoids the need to extract pre-identified (i.e., 'hand-crafted') features from the signals, and instead performs both feature extraction and diagnosis at once. However, more complex models mean many parameters to optimize and present the risk of overfitting if not enough data are available, which can be critical for ECG signals due to their potentially large variability and the limited amount of well-curated databases for training. Thus, for automatic diagnosis, the use of a separate feature extraction step can be a way

to reach more powerful and simplified data representations based on expert prior knowledge. Given the potential amount of features and their partial redundancy, feature extraction can be coupled with dimensionality reduction to reach more robust representations for use by machine learning models for automatic diagnosis.

Given the abundance of publications on this topic, we refer the interested reader to reviews of the literature addressing this question [248, 268], including some specific to deep learning [366, 116, 297, 166] which mostly rely on CNN and RNN architectures. We include a brief summary of this body of work below.

A first group of works focuses on heart beat classification, for which very high performance (more than 95% accuracy) has been achieved in much of the recent literature. A second group of works is aimed at automatic diagnosis of patients based on complete ECG recordings; the performance of these methods highly depends on the disease. This is clearly illustrated in the 2020 PhysioNet challenge [305], which provided 66,405 ECG recordings (43,101 with labels for training) and evaluated the results from 217 teams who attempted to automatically classify the ECG signals. Interestingly, the organizers designed a specific metric to compare the outputs of the competitors, using a reward process that softens some misdiagnoses depending on the severity of the disease or potentially different labelling of variants of a disease (e.g. "Complete right bundle branch block" vs. "Right bundle branch block"). A more recent paper focused on the PTB-XL database [372], which was part of the 2020 PhysioNet challenge, and provided a complementary view on deep learning methods for diagnosis on this database, with an interesting hierarchical organization of the diagnostic labels and some insights on the uncertainty and interpretability of such models.

Current Open Questions

As briefly summarized above, there are reasons to believe that high performing machine learning-based analysis of ECGs will become a reality for several applications in the near future, with the proviso that learning to cope with real-world data may present challenges.

As most of the methods involve supervised learning, the availability of large datasets with high quality annotations is crucial. The uncertainty in the man-

ual ECG annotations from a single expert can already be dramatic, and consensus in the annotation of events by different experts may be hard to reach [173, 385]. In addition, carefully and consistently annotating large series of signals is not feasible on any local database. The scarcity of well-annotated databases probably explains why a lot of focus is on the classification of ECG signals, and much less on delineation and feature extraction. One promising area for future work lies in the generation of realistic synthetic data, which by definition comes with ground truth annotations. This strategy has been successfully demonstrated in computer vision [328] and medical imaging [161] applications, and has started to be adapted to electrophysiology data [110, 177].

Developers and users of machine learning tools also need to keep in mind that 1-D (i.e. signals) does not necessarily mean simpler than 2-D (i.e. images). There exists a lot of variability in the signals due to noise, acquisition factors, or disease, which makes the detection of subtle events very challenging. Besides, the temporal dimension contains much of the useful information in ECG analysis where several cycles are often considered, compared to image analysis where a single image (for static data) or a single cardiac cycle (for temporal sequence analysis) is generally considered representative of the patient under study.

Also, although experienced users of neural networks tend to understand the role of subparts of the network and specific architecture choices, the path to the decision taken by the network is still hard to interpret. As described in other parts of this book (see footnote, page 29), interpretability is crucial for the transfer of these technologies to the clinic and this issue has started to be addressed by the machine learning community. A simple approach can be to produce attention maps that highlight the specific regions of the signal that led to the decision. For ECG analysis, this has been demonstrated on 2-D pictures of ECG signals, therefore borrowing the concept of attention maps from 2-D CNN and image analysis [394], and on actual 1-D signals [372].

Despite these issues, the advent of machine learning brings many hopes to the field of ECG analysis. The role of data analysis challenges will likely play an important role in realising these hopes, since they provide well-curated large databases and a specific question to address each year. They also serve to closely follow the evolution of the state-of-the-art and compare existing meth-

ods in a standardized manner. In this sense, the annual PhysioNet / Computing in Cardiology challenges have to be commended as they encourage focus on evaluation of performance specifically for ECG applications. In addition, their recent versions [305] comply with the good practices highlighted in recent meta-analyses of health data challenges [255]. This is surprisingly not the case for a large amount of data analysis competitions, although it should be of prime importance given the trend to use such public databases for local training or even for validation purposes. This is especially important given the amount of industrial investment in ECG analysis solutions, in particular for very precise applications such as diagnosing atrial fibrillation but also widening the spectrum of available signals (e.g. from wearables, smart watches, etc.), which bring complementary memory and speed issues that researchers will need to address.

Electronic Health Records

Transition to the Digital Era

Due to the boom of electronic devices and analysis techniques such as AI, but also the wide adoption of digital technologies within hospitals, the use of Electronic Health Records (EHR) has drastically increased in the last decade. They encompass a centralized collection of a patient's data followed along time through hospital visits or remote monitoring, and facilitate analyses and reporting at the scale of an individual patient or at population level [187].

The transition to digital technologies and big data raises a question that is not specific to EHRs, but is certainly shared by their users: how to properly manage this data deluge, a question which covers issues around acquisition, storage, maintenance, and access to these data.

More specific to the EHR, moving to such data requires a careful digitization of handwritten notes and voice records, for which a first set of AI methods from computer vision and speech processing are very relevant. This process is generally seen as *supervised*, in the sense that the inputs are mapped or tagged to given categories or values through classification or regression (for example, the concept "heart failure" written several times in a report should be tagged as a single "heart failure" item, which means recognizing the words "heart" and "failure", and considering them jointly).

Natural Language Processing (NLP) is a family of methods that are highly relevant for examining structured texts and enabling the machine to "understand" them. Among the AI methods it relies on, Recurrent Neural Networks (RNN) are suited for data that are sequentially ordered (typically, the text in a written document) as they can include long-range dependencies between the feature representations (the hidden states of the RNN). As nicely summarized in [356], they should not only address the extraction of single concepts, but also be able to spot the temporality of these events (namely how to convert a sometimes vague period of time into data that can be analyzed), and the relations in the text (for example, causes or conditions).

One important issue is that the huge amount of information contained in EHRs is currently insufficiently standardized across hospitals, clinicians, diseases, several visits, etc. The EHR scientific community is progressively moving towards more standardized formats, as done previously with the DICOM format for medical images. For this purpose, public datasets are of high value as they structure the community around a common task or challenge. A widely recognized example is the MIMIC-III dataset (Medical Information Mart for Intensive Care [180], which consists of deidentified EHRs from around 60K intensive care unit admissions.

To further structure the contents of these data, the combination of NLP and ontologies defined *a priori* can be useful, although these may be challenging to define. A broader view can also be adopted to better exploit the available data. For medical images this means, for example, considering both the image contents and the associated metadata either from the DICOM file or available in the EHR, as explicitly reviewed in [168].

Disease Perspective

Once the information has been extracted and structured, AI techniques and in particular machine learning and deep learning are now a "must-have" for the analysis. As for images and signals, machine learning can be used to address many challenges related to disease analysis with EHRs [155, 356], both in a retrospective or a prospective way:

- Detection [83, 235, 215]: for example, diagnosis (supervised) or abnormality detection (which can be unsupervised). See overview in Chapter 5.

- Prediction [81, 270, 432, 408]: for example, prognosis or evolution of specific values (either using a single timepoint, or through methods that explicitly address the temporality of events, such as RNNs or regression models). See overview in Chapter 6.
- Phenotyping [214, 43] (partially discussed in Chapter 8) to discover new concepts or confront existing ones with the data, for which unsupervised learning techniques are interesting as they can aggregate patients with similar data or conditions (clustering) or highlight the main characteristics of a dataset.
- Better representing a dataset [388, 82, 84, 211], which encompasses the previous item, and for which a specific family of representation learning algorithms exists [358], either using classical machine learning (manifold learning) or neural networks (autoencoders, see Chapter 4, page 91, and Chapter 5, page 128). The review in [356] nicely distinguishes between the objective of representing a medical concept across a population, or the data associated to a single patient.

Nonetheless, users should carefully balance the sophistication of the techniques used against the actual gain for the medical application. Indeed, a recent review [46] analyzed the evolution of an algorithm's performance in longitudinal EHR studies, where neural networks did not necessarily bring a clear gain in the last years. This report has to be tempered against the potentially increasing complexity of the databases, but the authors remind us that the difficulty of the medical questions and the variety of outcomes are clear bottlenecks for computational techniques using EHRs. They also argue for better standardization and organization in the EHR scientific community, referring to the good example of the ECG analysis community, which both structured the data formats, the feature extraction algorithms, and even databases including yearly data challenges[34].

Hospital and Patient Perspective

The wide spectrum of data covered by the EHR and their rather global acceptance also opens up new perspectives beyond disease specific studies.

Having patient records in a centralized and somehow standardized format first benefits the managing of resources by the hospitals. In this context, AI

[34] http://physionetchallenges.github.io/

techniques can be very valuable for comparisons and management at the scale of a whole hospital. When replaced in a temporal perspective, they can go beyond the prediction of mortality and estimate the length of stay of patients [46]. Automated reporting techniques are being developed in an attempt to speed-up the cumbersome processes by clinicians and the hospital staff [263], which can be seen as the process of generating contents in a structured way, and therefore encompasses AI generative models.

However, working at the scale of a whole hospital or even a network of hospitals brings additional challenges. Algorithms should target near-to-real-time access, or at least provide rapid information retrieval tools. *Federated learning* (see Chapter 5) is a framework that can be very useful to move beyond the limited point-of-view of a given hospital [329]: it involves training algorithms across multiple data warehouses without explicitly exchanging data. In the context of healthcare, this is highly desirable to develop more robust models with much better generalization ability, avoiding bias to some populations, and achieving better performance for rare diseases, while being safe in terms of privacy and security issues.

EHRs still come with many challenges around the standardization and fusion of many heterogeneous and time-varying data. However, the dynamism of EHR analysis with AI opens up promising perspectives to better contextualize the patients' data, including exploitation of external factors that are available in EHRs but not necessarily included in current analyses, accompanied by a much more regular follow-up and traceability that can benefit both the patient and the clinical institutions.

Closing Remarks

Cardiologists routinely make use of non-imaging data when making clinical decisions, so it seems inevitable that such data will play an important role in the future of AI in cardiology. In particular, some types of non-imaging data are routinely and widely available (i.e. ECGs and EHRs), so if techniques could be developed to better exploit the richness of these data this would be very attractive in terms of incorporating AI into current clinical workflows. However, non-imaging data sources are not immune from the difficulties and challenges associated with imaging data, such as standardization of formats, missing/corrupted data and privacy concerns. These issues must be satisfactorily addressed before AI techniques based on non-imaging data sources can

be translated into the clinic. In addition, challenges will likely be faced when using AI to combine features learnt from imaging and non-imaging data. Such an approach mimics the way in which cardiologists consider multiple sources of information when making decisions about patient management, and so has great potential, but it does increase the complexity of the models and of the data curation process. These are not concerns to be taken lightly, and further work is required before AI can truly emulate the way in which cardiologists are able to deal with such complexity in a seemingly effortless way.

Acknowledgements

ND was supported by the French ANR (LABEX PRIMES of Univ. Lyon [ANR-11-LABX-0063] within the program "Investissements d'Avenir" [ANR-11-IDEX-0007], and the JCJC project "MIC-MAC" [ANR-19-CE45-0005]).

RS was supported by Generalitat Valenciana Grant AICO/2021/318 (Consolidables 2021) and Grant PID2020-114291RB-I00 funded by MCIN/ 10.13039/501100011033 and by "ERDF A way of making Europe".

AK was supported by the EPSRC (EP/P001009/1), the Wellcome/EPSRC Centre for Medical Engineering at the School of Biomedical Engineering and Imaging Sciences, King's College London (WT 203148/Z/16/Z) and the UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare.