Data–Driven, AI–Based Clinical Practice: Experiences, Challenges, and Research Directions

Davide Ferrari[1,2], Federica Mandreoli[3], Federico Motta[3] and Paolo Missier[4]

[1]King's College London, London, UK
[2]Guy's and St. Thomas' NHS Fundation Trust, London, UK
[3]Università di Modena e Reggio Emilia, Modena, Italy
[4]Newcastle University, Newcastle upon Tyne, UK

Abstract

Clinical practice is evolving rapidly, away from the traditional but inefficient detect-and-cure approach, and towards a Preventive, Predictive, Personalised and Participative (P4) vision that focuses on extending people's wellness state. This vision is increasingly data–driven, AI–based, and is underpinned by many forms of "Big Health Data" including periodic clinical assessments and electronic health records, but also using new forms of self–assessment, such as mobile–based questionnaires and personal wearable devices. Over the last few years, we have been conducting a fruitful research collaboration with the Infectious Disease Clinic of the University Hospital of Modena having the main aim of exploring specific opportunities offered by data–driven AI–based approaches to support diagnosis, hospital organization and clinical research. Drawing from this experience, in this paper we provide an overview of the main research challenges that need to be addressed to design and implement data–driven healthcare applications. We present concrete instantiations of these challenges in three real–world use cases and summarise the specific solutions we devised to address them and, finally, we propose a research agenda that outlines the future of research in this field.

Keywords

Artificial intelligence, Machine learning, P4 medicine, High stake domains

1. Introduction

The promise of data–driven healthcare is underpinned by the availability of "Big Health Data" to feed machine learning (ML) and artificial intelligence (AI) models to achieve *prevention by prediction*. These datasets traditionally include the individual medical history (known as EHR, for Electronic Health Records), including primary and secondary care (hospital) events as well as medicine prescription history. In the vision of Preventive, Predictive, Personalised and Participatory (P4) medicine, these are complemented by a rich "cloud" of additional data types, ranging from *omics data (genotypes, transcriptomes, proteomes, ...), but also new forms of self–assessment, such as mobile–based questionnaires and automated self–monitoring logs from personal wearable devices.

SEBD 2022: The 30th Italian Symposium on Advanced Database Systems, June 19–22, 2022, Tirrenia (PI), Italy
davide.ferrari@kcl.ac.uk (D. Ferrari); federica.mandreoli@unimore.it (F. Mandreoli); federico.motta@unimore.it (F. Motta); paolo.missier@newcastle.ac.uk (P. Missier)
0000-0003-2365-4157 (D. Ferrari); 0000-0002-8043-8787 (F. Mandreoli); 0000-0002-5946-0154 (F. Motta); 0000-0002-0978-2446 (P. Missier)

Realising this vision requires a strong alignment between clinical research questions, data science and AI methods on one side, and data collection, curation, and engineering practices, on the other. Out of these three elements, in this paper we focus specifically on the data issues, reflecting on the main data challenges that need to be overcome to enable AI–based healthcare.

Drawing from our own recent experience working with prospective and retrospective patient cohort studies to learn a variety of predictive models, we suggest that data–driven healthcare applications are unique in terms of the challenges generated by the constantly–evolving, and often poorly controlled and even chaotic environment within which they are developed. For instance, clinical data is typically highly sensitive, costly to acquire and curate, and subject to complex governance policies.

The paper describes the data challenges we faced in three case studies coming from the Infectious Disease Clinic of the University Hospital of Modena, articulates how these were addressed in an *ad hoc* fashion to make modelling possible, and finally suggests a research agenda aimed at making the data engineering approach more principled and systematic.

1.1. Case study: Predicting functional ability in long–term HIV patients

The international study My Smart Age with HIV (MySAwH) is a multi–centre prospective project aimed at studying and monitoring healthy ageing in older people living with HIV. The cohort included 260 patients with 3 longitudinal follow–ups over 18 months, consisting of standardised clinical assessments, but also including an innovative element of patient self–monitoring, achieved using mobile smartphone apps and commercial–grade activity loggers. These were used to collect Patient Related Outcomes (PROs), combining questionnaires delivered with daily physical activity reports (limited to hours of sleep and step counts).

The study [1] used these combined datasets that refer to the notion of intrinsic capacity (IC), i.e. the combination of all the individual's physical and mental capacities, as proposed by the World Health Organization (WHO)[1], to predict individual health outcomes. The main data challenges associated with the study include ensuring the reliability, consistency, and completeness of the data collected from self–monitoring individuals.

1.2. Case study: predicting respiratory crisis in hospitalised Covid–19 patients

When Covid–19 hit the world in 2020, hospitals and researchers were not ready to tackle the emergency and adapted themselves day by day based on the unfolding of new necessities. Covid–19 brought unexpected complications to patient management, including managing limited ICU (Intensive Care Unit) resources in local hospitals. One of these was the University Hospital of Modena, where around 200 patients admitted between February and April 2020, at the start of the first wave of Covid–19 crisis in Italy, contributed to generate 1068 usable observations. Like in many other professional healthcare settings around the world, here the clinical staff started to collect new types of data for the inpatients, through standard blood tests but also specifically to monitor respiratory efficiency and to track their trajectory through stages of oxygen treatment and eventual outcome (discharge or death). Many hospitals used similar datasets in combination with ML algorithms to model mortality risk. In [2] we faced a problem related to resource

[1] Ageing and health, WHO, 2018. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

management; in detail, we developed a model to estimate the probability that a patient would experience a respiratory crisis within the next 48 hours. A correct estimate would be able for instance to prevent the early discharge of patients at risk. This was approached as a probabilistic classification problem (i.e. estimating the probability that a patient will or will not undergo a respiratory crisis).

1.3. Case study: predicting oxygen therapy progression in hospitalised Covid–19 patients

This latter case study shared the same domain of application of the previous one and started temporally after it; for these reasons several characteristics from its ancestor were actually inherited. In [3] too the faced problem was related to resource management; in detail we aimed at predicting a patient's transition from one type of oxygen therapy to another, which in turn would have helped to manage limited ICU resources in hospitals. This required a process modelling approach, based on Hidden Markov Models (HMMs), aimed at estimating the transition probability between any two states representing therapy regimes.

As anticipated, both the studies [2, 3] were underpinned by the same hospital dataset, but they used different subsets of variables at different time points. These were EHRs consisting of a combination of routine clinical tests as well as more specialised types of tests and observations. The experimental nature of these data, collected in a time of emergency, resulted in high instability (physicians constantly added or removed variables, effectively changing the schema on a daily basis) and imbalance, as the outcome of interest such as respiratory crisis or death were inevitably (and fortunately) the minority classes. Critically, modelling for both tasks had to contend with very small data sizes (in the order of thousands) and the problem of selecting significant predictors amongst a set of about a hundred ones.

2. Recurring research issues

In this section we present a catalogue of the main research issues that emerged in the three case studies introduced in Section 1, most concerning data quality. In the next Section we will articulate how these have been addressed in our case studies.

Data sparsity and scarcity. Individual medical histories like EHR data can be viewed as an irregular collection of time series, one per–patient, where each data point in the series is a patient event, consisting of a vector of variables. The collection is irregular because the time series have different event density for different patients and for different variables. Indeed, most variables are typically collected on a as–needed clinical basis, especially when involving expensive instruments or invasive examinations. Other variables, instead, are collected regularly but with varying frequency over time. Moreover, novel scientific evidence may induce changes in the data collection protocol.

For example, during the Covid–19 pandemic, routine data was collected from all the patients at University Hospital in Modena from the beginning; whilst some specific pieces of information, e.g. interleukin–6, from the blood laboratory analysis, were regularly collected only after

scientific evidence of their relevance in the diagnosis of patients affected by Covid–19–induced pneumonia, and thus they were absent for a substantial proportion of the patients. Similar problems were present in the data collected from self–monitoring individuals for the MySAwH study because activity loggers' data were available daily whilst Emotional Momentary Assessment data was collected through a smartphone app monthly.

When data are represented in a tabular format as needed to feed ML models, data sparsity can lead to large amount of missing values in the already collected data, which in turn results in fewer usable records, when missingness in important variables is not tolerated by the chosen learning algorithms. This problem is exacerbated when patients are few as it happens, for instance, in prospective studies like MySAwH or in emergency situation, like Covid–19, when there is the urgency of developing clinical decision making support systems.

For instance, the second Covid–19 study mentioned above used a later version of the same dataset with data collected for a longer period (March 2020–May 2021); where, as depicted in Figure 1, only between 23% and 60% of the available records were used, against the 56% of the original study.

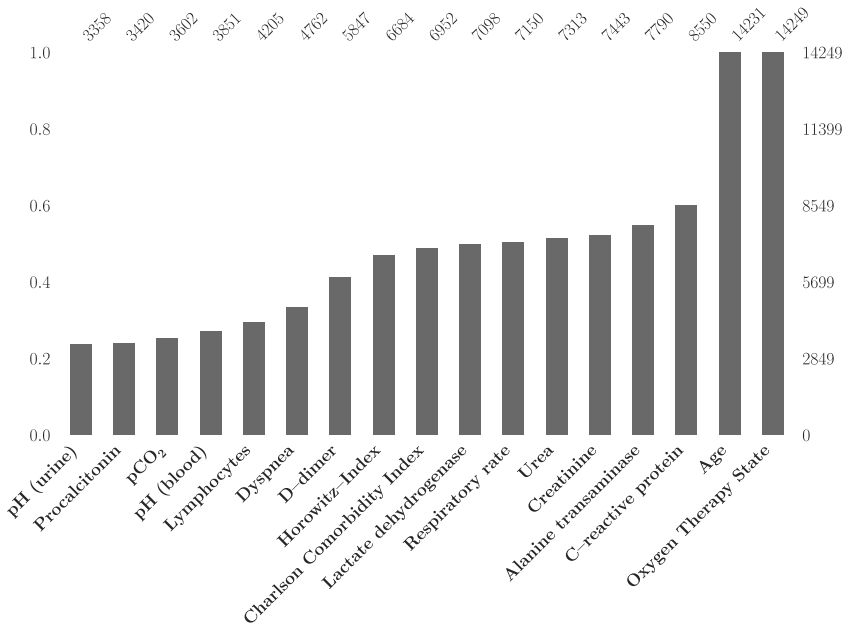

Figure 1: Percentage variable completeness of the Covid–19 dataset used in [3]

While missing data can sometimes be inferred, or imputed, from available value distributions, this is not an option when dealing with critical patients vital parameters, which by their own nature are subject to abrupt changes and thus should not be extrapolated from known distributions. In fact, one may argue that the value of the data in this context is the change in data values, signalling an impending crisis.

Data imbalance Data imbalance is a well–known problem when trying to learn a classifier predicting a rare event. Yet, these are often precisely the classes of interest, for instance a respiratory crisis, which is rare relative to the balance of other more positive outcomes. Similarly, when the focus is on patient's oxygen state transition, the intubation therapy is (fortunately) a rare state and, as such, it could perhaps be disregarded. However, as the low frequency matches of these events are matched by correspondingly low ICU capacity, predicting such events remains a priority.

Data inconsistency and instability Prospective datasets that are collected for research purposes, such as for [1], tend to be stable as they are the result of an agreed upon protocol. Interesting research insights, however, are often derived from retrospective datasets, which may include a broad variety of observations and are often collected in an opportunistic fashion over unanticipated periods of time. This has been the case for the Covid–19 datasets used in [2, 3], where not only the content were updated at irregular intervals, but also the set of variables and their use and indeed the overall schema evolved over time, driven by the discovery of new variants, changes in diagnostic strategies, and changes in hospital management policies. For example, the information about *tocilizumab* administration was initially collected in a column containing free text notes, while after a while it became an *ad hoc* boolean column due to the fact that patients who received it started being a considerable amount of the total ones. Another example regards the O_2 therapy setup: it has always been a free text entry, so the inconsistencies, mistyping and individual interpretation while inputting made the programmatic analysis of that data very complicated (therapies regimes were reported as percentage of oxygen in breathing air, liters per–minute delivered by the mask or the name of the mask itself, all in the same database field). A subsequent adaptation and correction of previous values was needed to harmonise the reporting of such information. In [3] too, we faced similar issues, this time directly involving the outcome; in fact the oxygen therapy states were initially only 4, i.e.: *No* O_2, O_2, *NIV*, *Intubated*, plus the two final ones *Deceased* and *Discharged*, respectively. Then, when the second wave of patients began, they were so many that the previously known O_2 state, in which patients used to breath air enriched with oxygen through a venturi mask, had to be partitioned into two states: a first one with the same name/ventilation support and a second one named *HFNO*, providing a High Flow of Oxygen through a Nasal cannula. This change forced the first Hidden Markov Model [4] to be retrained and pushed towards adopting a more robust ensemble solution.

Changes in data format and units are as common as they are insidious, as they tend to break the data pre–processing pipelines designed to wrangle the raw data into a training–ready format, requiring lengthy repairs. These changes are not limited to emergency situations such as the one described. Indeed, data acquisition and curation practices are also affected by changes in public policy, hospital resources, collection technology, as well as the ability to link out and integrate with other data sources.

Importantly, in a scenario where data is used to train ML models, this instability, in turn, translates into instability of the models trained on these datasets. While simple re–training is sufficient when the data grows with a stable schema, an evolving and unstable schema affects the choice of learning algorithm and of its hyper–parameters, as well as variable ranking and overall model performances, requiring constant maintenance of the models and thus propagating the

instability problem into the deployment stage.

Not all errors are equally wrong In binary classification problems, predictive performance is routinely measured by counting false positives and false negatives. When the relative cost of these errors is the same, standard measures such as F–score and AUROC represent an efficient way to summarise predictive performance.

This is hardly ever the case with high–stakes medical applications, where a bias towards one type of error is often preferable. For instance, when predicting respiratory failure (and generally when predicting a class that represents the undesirable outcome), a conservative stance where false positives are preferred to false negatives ensures that no unnecessary risk (i.e., of early discharge) are taken, possibly at the cost of extra attention to patients. In the next Section we will reflect on how such deliberate bias was introduced in both of our Covid–19 case studies.

Human-in-the-loop The closer predictive models come to being adopted as part of clinical practice, the more pressing the need becomes to ensure that the models are explainable, on the assumption that explanations engender trust in the models. What this means in practice, however, is not always clear. For instance, it is becoming increasingly evident in the health data science community that trust should include not only the clinician but also the patient.

Thus, even the simplest type of explanations, namely a weight–ranked list of features used in a linear model, is questionable as those variables mean little to the patient. More sophisticated technical explanations are now available for non–linear models, too, as well as for the interpretation of histology images, for example. However, these still do not address the patient side of the "dialogue", and there is little agreement that they would be sufficient for the physician, too.

Even assuming the model can "explain itself", this only addresses half of the problem. In a true *human-in-the-loop* AI scenario, it should be possible to provide feedback to the learning algorithm, reflecting agreement or disagreement with the prediction, or perhaps to force a bias (like discussed briefly in the previous point). While this is technically possible, for instance by changing ground truth annotations or using one of the many available penalty–based models, this is again not a level at which clinicians are comfortable to operate.

Data science as a translational science Statistics has been at the basis of clinical practice for decades, however in the last years also the communities built around the development of AI and ML techniques have came in touch with medicine; these two fields are nevertheless evolving at different speeds, as well as some physiological diffidence and resistance are slowing down this cross–domain integration too. To bridge this gap between physicians and data scientists, a common language and shared efforts are inevitably needed: clinicians need to better understand the rationale behind mathematical models in order to allow improving them by providing useful insights from their domain of expertise, to better overcome the issues arisen by the data; on the other hand computer scientists and mathematicians need to understand the clinical meaning of the data they deal with, in order to build really useful and trustworthy applications. In our experience, the closer this relationship is cultivated, such as in a daily interaction, the quicker this iterative methodology will converge towards results suitable for both the research fields,

because the former need new and more powerful tools, the latter real–world problems to tackle and improve well–known methods.

3. Addressing the challenges, one case study at a time

In this section we are presenting concrete examples of real–world research projects in which those challenges occurred. We will briefly contextualise the problem in the clinical scenario and explain the strategies that were put in place to adequately produce functioning data–driven pipelines. Table 1 provides a summary of such challenges associated with the data used in each of the studies.

Table 1
Challenges faced by each study

Challenge	MySAwH	Covid–19 PaO_2/FiO_2	Covid–19 O_2
Data imbalance			✓
Data inconsistency and instability		✓	
Data sparsity and scarcity	✓	✓	✓
Human-in-the-loop	✓	✓	
Not all errors are equally wrong		✓	✓

3.1. Use case 1: My Smart Age with HIV

The goal of this project was to create ML models using the patient–generated data to predict three relevant clinical outcomes at the following 9–months visit in hospital; namely, Quality of Life (QoL), physical activity capability, and the risk of having a fall. Available variables included (i) *physical activity* (steps count, sleep hours and calories); (ii) 56 *Patient-reported-outcomes (PROs) on QoL, collected using a smartphone*; and (iii) *clinical variables*, including HIV–specific variables. Clinical variables built a comprehensive geriatric assessment and were collected by healthcare workers during hospital visits at time 0, 9 and 18 months generating 2 inner time windows; 37 of these variables were also used to measure the Frailty Index (FI) as defined in [5]).

Data heterogeneity and sparsity emerged since the protocol was designed to collect different type of data at different frequency. Gaps of up to 17 consecutive missing observations were found in PRO variables, with 108 gaps per–patient on average. Our approach was to impute by interpolating missing data points in the time series modulating the maximum number of consecutive missing values imputed to not compromise the model performance. Still, the remaining missing data resulted in a loss of usable observations. Given the different granularity of the collected data and the missing data left after the interpolation, we also needed to re–sample and aggregate the three data sources to a monthly frequency. The two time windows were used as reference for the prediction task as for each monthly data point, the prediction target was the clinical outcome at the end of the respective 9–months period.

Despite data scarcity, we were able to compare expert–provided, or "knowledge–driven" (KD) clinical risk scores to data–driven (DD) scores obtained using the training set. The results,

indicating superiority of DD, are shown in Figure 2, where we also tested the relevance of a Frailty Index FI [5] as an additional predictor.

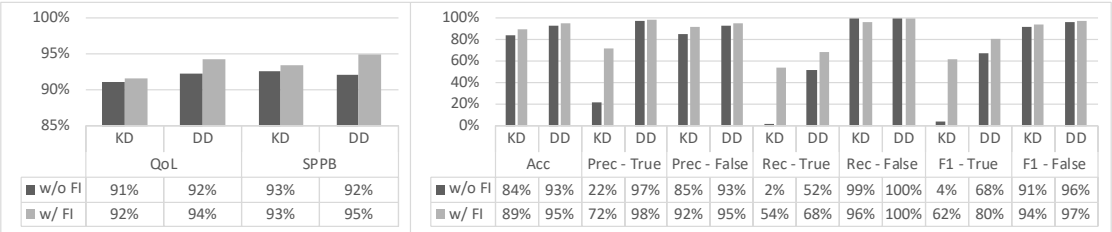

Figure 2: Performance measures of the three clinical outcomes predictors: QoL and SPPB (left) and Falls (right)

Shapley values [6, 7] were then used to provide an interpretation of model prediction at individual level. These provided both a quantitative as well as qualitative view of the relative importance of the variables in prediction. Figure 3 shows different interpretations for the same predicted value but for two different patients.

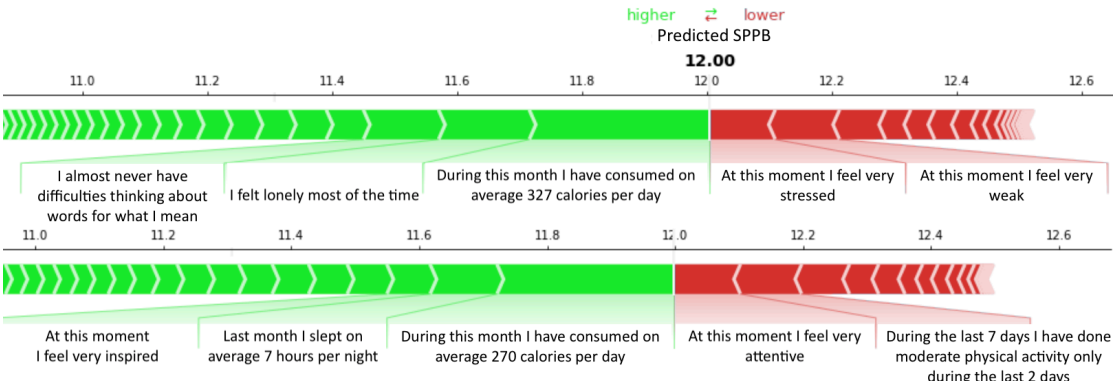

Figure 3: Shapley Values interpretation for two different patients with the same prediction of Short Physical Performance Battery (SPPB)

3.2. Use case 2: Covid–19, predicting respiratory failure

In [2], respiratory failure was tested by using blood gas analysis, namely when the PaO_2/FiO_2 ratio falls below the 150 mmHg threshold, within two days from admission to hospital. Thus, its prediction translated into a binary classification problem.

Data instability resulted from the irregular and experimental collection of 91 variables, including 39 from blood and urine tests, 7 from the blood gas analysis, 29 different disease specific symptoms, 14 co–morbidities and demographics.

Our milestone data extractions of 2,454 and 2,888 data points each provided an average completeness of $62\% \pm 22$ and $57\% \pm 22$ respectively. Some of the variables were collected on–demand based on clinical needs and few were introduced in the data collection only after their need was proven by scientific literature. For example, *lymphocytes* were present only in 7.5% of the samples in a first data extraction and 7.8% in a subsequent one; *interleukin–6* instead was collected in about 20% of the daily samples and given its rapid variability in time, imputation was not a reliable strategy. To handle this inevitable lack of data we created a model trained using the robust Python implementation of LightGBM [8] which supports missing values without the need of imputing them.

Data inconsistency was also introduced, as the information systems used by the hospital evolved almost daily. For example, oxygen therapy measures were progressively refined, but the units of measure changed in the process from liters per–minute to percentage of oxygen in breathable air, making it difficult to discern automatically with which the clinician recorded the value.

In this high–stake domain we want to prevent as much as possible false negative (FN) predictions because they are extremely more dangerous than false positives (FP) and they can imply the discharge of a patient at high risk; to address this issue, a bias was introduced into the binary cross entropy in order to penalise FP, possibly at the expense of additional FN. The employed equation was the following:

$$L(y, p\,(x)) = -\beta \cdot y \ln p(x) - (1 - y) \ln\,(1 - p\,(x)) \tag{1}$$

where β parameter was used to modulate the increasing penalty given to FN and we experimentally found the best balance with $\beta = 2$ (i.e., the penalty for a FN prediction is double of the penalty for a FP).

As in the previous case study, we relied on Shaply values [7] to provide per–individual explanations of the predictions in terms of the variables. Moreover, the global Shapley value ranking was used to identify the top features out of the initial 91, resulting in a more parsimonious model with no appreciable performance loss (AUROC 84% for the leaner model, compared with 85% of the one using the full feature set). More precisely, we removed the less relevant features until the performance started to detriment significantly; this produced a list of 20 final variables which accounted for the vast majority of information delivered by the dataset for this learning task.

3.3. Use case 3: Covid–19, predicting oxygen therapy states

In [3] the aim was to predict the patients' state transitions, where each *state* represented an oxygen therapy state, which could change for each patient on a daily basis.

The initial approach involved training an HMM over a set of 17 observable variables, 2 cross–sectional (*age* and *Charlson Comorbidity Index*) and 15 longitudinal, including oxygen therapy. This was complicated by data sparsity, as shown in Figure 1, which was addressed using a library robust to missing data [9]; and more importantly, by a strong imbalance on the state transitions, whereby the most common state [4] was also the clinically least interesting.

Table 2
Per–state performance in terms of E–measure($\beta = 0.5$)

Performance measure (ε < 0.01%)	Global accuracy	No O_2	O_2	HFNO	NIV	Intubated	Deceased	Discharged
a) Single HMM	38.7	–	–	–	–	–	–	–
b) Majority voting ensemble	–	ε	94.7	ε	ε	ε	ε	ε
c) HMM–ensemble	–	ε	97.3	95.1	98.5	98.1	37.5	25.9

As standard re–sampling does not work well on such small datasets ($14,249$ EHRs for about $1,040$ patients), we implemented an ensemble–based approach based on two main strategies. Firstly, we made the model overfitting aware, which made it possible to prune out some models/outcomes *a priori*. And secondly, we gave it the ability to compare and eventually combine outcomes coming from each one of the models, by "cherry–picking" simple pieces of solution from a partition of the explored space, in order to mitigate and better face the imbalance problem in a *divide et impera* manner; in fact each sub–problem had more balanced outcomes, which allowed the respective HMMs to overfit less.

The elements of novelty were the choice of two hyper–parameters, able to tune the aggressiveness of the aforementioned overfitting–aware pruning mechanism; as well as the choice of the functions used to measure the *degree of support* of each model for each outcome. The latter ones are in facts, the key point around which the pieces of solution are compared, i.e. ranked, and combined; in our experience they produced better performance when including among the terms: a model performance metric (e.g. F_1–score, Recall) and either the number of classifiers which did not vote for the outcome, rather than the outcome inverse probability or its logarithm. As visible in Table 2, our algorithm (c) had remarkable performance ($> 95\%$) in terms of E–measure$_{\beta=0.5}$ in all the severe/critical states, as well as it better performed than state of the art approaches (a, b) with regard to the final ones (i.e. *deceased* and *discharged*).

4. Concluding remarks and future research agenda

In this paper we illustrated three examples where data quality challenges that are common to many problems in Health Data Science have been addressed in an *ad hoc* fashion, by customizing out-of-the-box ML algorithms to meet specific requirements.

Here we advocate a more systematic and principled approach, possibly leading to an interesting research agenda. Such approach should be centred on data–centric AI, i.e. the systematic engineering of the data used to build an AI system [10]. This consists of a number of techniques that are perhaps less known in ML. Firstly *data augmentation*, a data–centric AI technique that can be employed to overcome real–world data challenges including data sparsity, scarcity, and unbalance.Data augmentation involves combining limited labeled data with synthetic data. While this line of research is still in its infancy, interesting advances are found in the development of generative models [11, 12], a promising direction specifically when dealing with

clinical data.

Secondly, *self–supervision* and *semi–supervision* provide a way to overcome the scarcity of labelled data, by combining it with existing and more abundant unlabeled data. However, current solutions like the VIME system [13] have limited applicability in clinical practice as they do not tolerate the large portion of null feature values that characterises this type of data, whilst at the same time state-of-the art data imputation approaches, applied to our clinical datasets, reduce ML model performance to unacceptable levels.

In the data–centric AI vision of system production, most of the complexity of ML systems is tied to data processing, handling and monitoring. Preparing data pipelines in clinical practice currently requires considerable human effort and is very time consuming. Extensive domain knowledge is necessary to address the problem of inconsistency in such kind of data. A more comprehensive approach to improve the quality of the data itself, along with the various dimensions discussed in this paper, are therefore also required.

However, when trying to automate data cleaning and preparation using scripts or workflows, we see that data instability, caused by successive data extractions from a source, runs counter to these efforts, as changes in data format and schema may easily break the data pipelines. One possible direction to alleviate this problem involves adding more robust debugging facilities to the pipeline. Capturing and querying the provenance of the transformations produced by the pipeline may just be the tip of the solution.

Another research field which could contribute in making the P4 medicine vision concrete in clinical practice is the human-in-the-loop ML. Indeed, a real iterative process where the model provides comprehensible explanations and in turn is incrementally improved based on user feedback is not yet part of routine health data science practice. Nevertheless, according to the physicians we work with, this is one of the most expected revolutions in the near future.

As far as explainability is concerned, although Shapley values represent a valuable concrete option in medicine [14], it will become important in the next future to study more intuitive forms of explanation that better meet physicians' and patients' needs like example–based, image, and textual explanations [15]. Last but not least, future research investment should be devoted to the development of ML models that are able to alter their decision [16] and act on modifiable variables [17] according to external inputs like expert domain knowledge and user feedback.

According to the clinicians we reached out and us, both our communities should push towards the implementation of these human-in-the-loop aspects; which may contribute in closing the gap in the cross talk between the patients, the physicians and the data scientists, finally building mutual trust between them and providing qualitatively better healthcare solutions.

Acknowledgments

The authors would like to thank Prof. Giovanni Guaraldi for providing the datasets which made these studies possible, but also all the healthcare workers of the Infections Diseases Clinic of the University Hospital of Modena which contributed to their collection and constantly provided constructive feedback during the designing, developing and testing phases of the models built for the three use cases presented in this paper.

References

[1] D. Ferrari, G. Guaraldi, F. Mandreoli, R. Martoglia, J. Milić, P. Missier, Data–driven vs knowledge–driven inference of health outcomes in the ageing population: A case study, in: CEUR Workshop Proc., 2020. URL: http://ceur-ws.org/Vol-2578/DARLIAP8.pdf.

[2] D. Ferrari, J. Milić, F. Mandreoli, P. Missier, G. Guaraldi, et al., ML in predicting respiratory failure in patients with COVID–19 pneumonia: challenges, strengths, and opportunities in a global health emergency, PLOS ONE (2020)1–14. doi:10.1371/journal.pone.0239172.

[3] F. Mandreoli, F. Motta, P. Missier, An HMM–ensemble approach to predict severity progression of ICU treatment for hospitalized COVID–19 patients, in: 20th IEEE Int. Conf. on ML and Appl., 2021, pp. 1299–1306. doi:10.1109/ICMLA52953.2021.00211.

[4] F. Motta, Hidden Markov Models to predict the transitions of SARS–CoV–2 patients' state, Master's thesis, Università degli studi di Modena e Reggio Emilia, 2020.

[5] I. Franconi, O. Theou, L. Wallace, A. Malagoli, C. Mussini, K. Rockwood, G. Guaraldi, Construct validation of a Frailty Index, an HIV Index and a Protective Index from a clinical HIV database, PLOS ONE (2018). doi:10.1371/journal.pone.0201394.

[6] S. M. Lundberg, et al., Explainable ML predictions for the prevention of hypoxaemia during surgery, Nature Biomedical Eng. (2018) 749–760. doi:10.1038/s41551-018-0304-0.

[7] S. M. Lundberg, et al., From local explanations to global understanding with explainable AI for trees, Nature Machine Intell. (2020) 56–67. doi:10.1038/s42256-019-0138-9.

[8] G. Ke, et al., LightGBM: A highly efficient gradient boosting decision tree, in: Adv. in Neural Inf. Process. Syst., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[9] J. Schreiber, Pomegranate: fast and flexible probabilistic modeling in Python, Journal of ML Research (2018) 1–6. URL: http://www.jmlr.org/papers/volume18/17-636/17-636.pdf.

[10] N. Polyzotis, M. Zaharia, What can data–centric AI learn from data and ML engineering?, in: Adv. in Neural Inf. Process. Syst., 2021. doi:10.48550/arXiv.2112.06439.

[11] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, X. Xiao, Privbayes: Private data release via bayesian networks, ACM Trans. Database Syst. (2017). doi:10.1145/3134428.

[12] J. Jordon, J. Yoon, M. van der Schaar, PATE–GAN: Generating synthetic data with differential privacy guarantees, in: 7th Int. Conf. on Learning Representations (ICLR), 2019. URL: https://openreview.net/forum?id=S1zk9iRqF7.

[13] J. Yoon, M. van der Schaar, et al., VIME: Extending the success of self and semi–supervised learning to tabular domain, in: Adv. in Neural Inf. Process. Syst., 2020. URL: https://proceedings.neurips.cc/paper/2020/file/7d97667a3e056acab9aaf653807b4a03-Paper.pdf.

[14] K. D. Pandl, F. Feiland, S. Thiebes, A. Sunyaev, Trustworthy ML for health care: Scalable data valuation with the shapley value, in: Proc. of the ACM Conf. on Health, Inference, and Learning, 2021, pp. 47–57. doi:10.1145/3450439.3451861.

[15] G. Vilone, et al., Classification of explainable AI methods through their output formats, Machine Learning and Knowledge Extraction (2021) 615–661. doi:10.3390/make3030032.

[16] N. Boer, T. Milo, et al., Personal insights for altering decisions of tree–based ensembles over time, Proc. VLDB Endow. (2020) 798–811. doi:10.14778/3380750.3380752.

[17] A. Hall, et al., Identifying modifiable predictors of patient outcomes after intracerebral hemorrhage with ML, Neurocritical Care (2021) 73–84. doi:10.1007/s12028-020-00982-8.