Data-driven vs knowledge-driven inference of health outcomes in the ageing population: a case study

Davide Ferrari
University of Modena and Reggio Emilia
Modena, Italy
162996@studenti.unimore.it

Giovanni Guaraldi
University of Modena and Reggio Emilia
Modena, Italy
giovanni.guaraldi@unimore.it

Federica Mandreoli
University of Modena and Reggio Emilia
Modena, Italy
federica.mandreoli@unimore.it

Riccardo Martoglia
University of Modena and Reggio Emilia
Modena, Italy
riccardo.martoglia@unimore.it

Jovana Milić
University of Modena and Reggio Emilia
Modena, Italy
jovana.milic@gmail.com

Paolo Missier
Newcastle University
Newcastle upon Tyne, UK
paolo.missier@ncl.ac.uk

ABSTRACT

Preventive, Predictive, Personalised and Participative (P4) medicine has the potential to not only vastly improve people's quality of life, but also to significantly reduce healthcare costs and improve its efficiency. Our research focuses on age-related diseases and explores the opportunities offered by a data-driven approach to predict wellness states of ageing individuals, in contrast to the commonly adopted knowledge-driven approach that relies on easy-to-interpret metrics manually introduced by clinical experts. This is done by means of machine learning models applied on the My Smart Age with HIV (MySAwH) dataset, which is collected through a relatively new approach especially for older HIV patient cohorts. This includes Patient Related Outcomes values from mobile smartphone apps and activity traces from commercial-grade activity loggers. Our results show better predictive performance for the data-driven approach. We also show that a *post hoc* interpretation method applied to the predictive models can provide intelligible explanations that enable new forms of personalised and preventive medicine.

1 INTRODUCTION

Medical 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, with particular focus on ageing individuals [19]. This vision is increasingly data-driven, 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. With these premises, P4 medicine has the potential to not only vastly improve people's quality of life, but also to significantly reduce healthcare costs and improve its efficiency.

Our research explores specific opportunities offered by data-driven approaches to predictive care, in contrast to traditional, knowledge-driven approaches that rely purely on clinical expertise. Our focus is on age-related diseases, an emerging issue for health care systems. The World Health Organisation estimates that the proportion of people over 60 years of age will reach 2 billion by 2050 [17]. Ageing is associated with increased prevalence of co-morbidities that accumulate in the complex of multi-morbidity in older people [1].

1.1 Background

We focus on two easy-to-interpret metrics that clinical researchers have proposed to succinctly express the health status of patients at a given point in time. The first is a measure of *frailty*, designed to quantify the reduction of *homeostatic reserves* available to an individual. High frailty is indicative of higher risk of negative health outcomes, but it is also a potentially reversible condition [21] In practice, frailty is measured using a variety of specific *Frailty Index* metrics (FIs). These are calculated using at least 30 directly assessed health variables, which include signs or symptoms, biochemical parameters, various comorbidities or socio-demographic data [6, 22]. The choice of the specific variables, as well as their sources, may vary depending on data availability, leading to different specifications for the FI. Given their complexity and multidimensionality, FIs reflect the biological age of an individual rather than their chronological age, making them reliable prognostic tools that can be used in different settings for clinical decision algorithms [6]. In particular, the dataset used in this research concerns a cohort of HIV patients. This is important, because long-lived HIV patients exhibit a form of accentuated ageing [3],such that they can successfully be used to study frailty, i.e., where the duration of the condition (number of years since infection) is used as a proxy for chronological age.

The second measure of health considered in this work directly reflects the more positive notion of healthy ageing [16] that is becoming prevalent especially in public health settings. In contrast to frailty, which is designed to measure decay, the term *healthy ageing* (HA) has been proposed to promote a positive approach to ageing that relies on reserves and preserved capacities in an individual, rather than accumulation of deficits. In the World Health Organization Guidelines on Integrated Care for Older People (ICOPE), HA is based on concrete measures of *Intrinsic Capacity* (IC). These are defined as a composite of all the physical and mental capacities of an individual, divided into five domains: locomotion, cognition, psychological, vitality and sensory capacity [16].

As suggested by Belloni and Cesari [2], frailty and IC should not be considered as two opposed constructs, but rather two constructs that share a common biological background. The IC should be considered as an evolution of the frailty concept, taking into special consideration the functional reserve expressed by the vitality domain, the need for a worldwide implementation of prevention, the continuum of the ageing process, and the opportunities offered by novel technologies [8].

Key to a successful operational definition of IC is the choice of the variables used to characterise healthy ageing. [16] argues that integrated care is crucial to incorporate intrinsic capacity assessment in new care models and encourages the adoption of

wearable devices and mobile apps for longitudinal data collection. In [9], the authors proposed a self-generated health measure called Intrinsic Capacity Index (ICI) which relies on physical function data collected through fitness tracking wearable devices and a set of electronic *Patient-Related Outcomes* (PRO) collected through a dedicated smart-phone app (MySAwH App). In the study, these variables have been collected longitudinally over a period of 18 months for a cohort of HIV patients as part of the *My Smart Age with HIV* (MySAwH) study. This is a prospective multi-center international case-only study, designed to empower *Older People Living With HIV* (OPLWH) to achieve healthy lifestyles and improvement in quality of life.

Similarly to FI, ICI is computed from a subset of the available variables, by manually selecting a cutoff point for each variable and simply counting, for a given patient, the variables with value higher than the cutoff for that patient. We have referred to this approach that represents the common practice to assess health condition in geriatric medicine as *"knowledge-driven"* (KD for short), as it relies on easy-to-interpret metrics where the choice of variables and of their cutoff points is defined manually by clinical experts. The usefulness of the ICI is demonstrated in [9] by showing experimentally that it displayed higher sensitivity than FI to predict one central indicator of "wellness state" of ageing individuals, the Quality of Life (QoL in short).

1.2 Contributions

In contrast to the dominant KD methods as described, in this paper we explore a complementary data-driven approach (DD) to predicting wellness states for long-term patients, using a combination of clinical, self-monitoring, and PRO (self-reporting) longitudinal observations that refer to the five IC domains. As we will see, while this approach removes the need for the clinical experts to directly define metrics such as the ICI, it also empowers them by deploying machine learning techniques that make the predictions easily interpretable.

Specifically, we focus on three dependent variables to characterise the "wellness state" of ageing individuals, namely, (i) Falls, indicating whether or not a person has experienced any falls within a given time period; (ii) SPPB, or Short Physical Performance Battery, measuring movement of the lower limbs, and (iii) Quality of Life (QoL). Using the MySAwH dataset for training, we can directly contrast the DD and KD approaches, and we show that we can achieve better predictive power without the need to rely on manually formulated ICI metrics. Furthermore, we also show that the models' performance increases if we also include a single Frailty Index value for each patient, representing a "baseline" clinical assessment, in addition to the PRO and activity variables.

Finally, we combine well-established machine learning models with a post-hoc interpretation method and we show that satisfactory model performance can be achieved together with *intelligible* explanations, which provide the additional benefit of ranking the variables with respect to a prediction. Indeed, we show that the relative importance of the variables differs for each patient, indicating the important role played by intelligible models towards personalising healthcare. This capability also provides critical feedback to the clinicians, who can use this information in combination with their expertise to moderate the model's predictions.

2 RELATED WORK

The idea of exploiting machine learning toward a general "health" focus is certainly gaining popularity, also thanks to the possibility of continuously monitoring well-being through wearable activity trackers' data. For instance, passive sensing techniques have been recently exploited to assess both mental and physical health [20] or to predict weight objectives for users of smart connected devices [26]. Some researchers also successfully exploited this kind of data in more traditional disease monitoring scenarios, e.g., to track Multiple Sclerosis [24], depression [29] and schizophrenia [28] patients' health. In most of these studies mobile phones are only exploited as a passive sensing device, while in our case we combine the use of passive sensing (wearable activity tracker) data with self-reporting EMA data collected through mobile phone apps.

A critical aspect for the successful application of ML to medicine is the recent increased emphasis on the need for explanations of ML systems [7, 10]. This has lead to another important research trend, i.e. Interpretable Machine Learning in healtcare [14]. Even if recent researches have proposed new models which exhibit high performance as well as interpretability, e.g., GA2M[15] and rule-based models [25], the utility of these models in healthcare has not been convincingly demonstrated yet, due to the rarity of their application [14]. The interpretation method used in [11, 12], instead, is designed to work with existing and well established (even if less interpretable) ML methods, such as gradient boosting or deep learning, by extracting explanations through models that are applied post-hoc using Shapley Values [23]. This is one of the most advanced interpretation methods available, allowing for both global (entire study population) and local (instance level) explanations. In our study we exploit such model together with the XGBoost [4] gradient boosting algorithm, in order to aim to both interpretability and performance. A similar technique has been also proposed, only at conceptual level, in the Explainable AI framework discussed in [27], which is however based on clinical-only data analysis.

3 THE HIV COHORT DATASET

The experimental dataset used in this paper was obtained from the My Smart Age with HIV (MySAwH) [18] project. MySAwH is a multi-centre prospective ongoing study aiming at empowering OPLWH, i.e. 50+ years old, to develop healthy lifestyles. The project involves 261 patients from three clinics: 128 from Modena (Italy), 100 from Sydney (Australia) and 33 from Hong Kong (China). One novelty of the approach is the combination of clinical patient data, acquired during periodic scheduled assessments in the clinic, with *patient-oriented* longitudinal data about patients' behavioural, physiological and environmental health status, which is collected at higher frequency through smartphones and wearable devices.

Thus, the resulting dataset is highly heterogeneous as to the data type, the geographical origin of patients and the acquisition rate. It provides a comprehensive characterization of patients from the broad determinants of health that impact aging, complemented by those more specific to OPLWH, namely:

- **Activity tracking variables:** Step count, sleep hours and calories, which are collected daily using a commercial-grade wearable activity tracker;
- **PRO variables:** 56 categorical questions exploring functional abilities and Quality of life (QoL). These are collected monthly using a dedicated smartphone app;

Clinical variables: Comprehensive geriatric assessment and HIV-specific variables, which are collected by healthcare workers during study visits at time 0, 9 and 18 months. 37 of these variables were used to measure the Frailty Index (FI) as defined in [6]: 27 from blood tests, 3 about body composition, 7 HIV-related variables and patient-reported outcomes.

A preliminary analysis [8] introduced a new index for expressing IC, named ICI, and compared the performance of FI and ICI in predicting QoL. This is a quantitative measure of health as assessed by the individual respondents that is widely used on aging population. QoL was assessed using a standardised EQ-5D-5L questionnaire, based on the EQ Visual Analogue scale (EQ VAS) [5]. Through the MySAwH dataset, FI and ICI were shown to be performative tools that can be used in research and clinical setting to describe disease and health status in OPLWH. ICI score in comparison to FI displayed higher sensitivity to predict the QoL and self-perceived health in OPLWH.

Outcomes. In this work, we focus on the task of predicting significant healty ageing indicators and QoL is one of such indicators. In addition to QoL, two other indicators were selected: Falls, that is an adverse outcome included among the geriatric syndroms, and the Short Physical Performance Battery (SPPB), a group of measures about movement ability with lower limbs (Guralnik et al., 2000) that can aid in the monitoring of function in older people. These three outcomes are chosen because they widely cover all 5 domains of IC. In summary, these are as follows (Fig. 1 shows their distributions):

Quality of Life (QoL): assessed using the EQ-5D5L standard, with values between 0 and 1;

Falls: A binary outcome that evaluates to True if a patient has fallen at least once since the previous visit, and False otherwise;

SPPB Index: a discrete index that assumes integer values between 0 and 12.

Observational data and feature space. The observational dataset used to predict these outcomes covers 18 months and is broken down into two time windows, reflecting the clinical assessment schedule (months 9 and 18), when the selected clinical outcomes are assessed. For each time window, we draw two sets of samples from the related observations, which we are going to use as ground truth to train our models.

The first consists of the patient-centric longitudinal data, including the activity tracking and the PRO variables. To this end, we further aggregated the resulting PRO time series and the activity tracking time series at regular intervals of 1 month, resulting in two sets of data points, one for each of the two windows. The second sample set augments the first with the FI values computed from the clinical variables measured during hospital visits at the beginning of each window, namely at times 0 and 9. This added value is a physician's assessment that complements the patient-centric data point and can be interpreted as the baseline of the time series the data point comes from.

Formally, for each outcome $o \in \{QoL, SPPB, Falls\}$ we write $Sample_o$ to denote the sample set containing the monthly samples per patient. For each patient p, we denote a single sample in the set $Sample_o$ at month $m = i + (j-1) * 9$, corresponding to the i-th observation, $i \in [1, 8]$, in the j-th window, $j \in [1, 2]$, by a pair $s^p_m = (x^p_{i,j}, y^p_j) \in Sample_o$ where

- $x^p_{i,j}$ represents the feature vector for i-th observation;
- the feature vector $x^p_{i,j}$ contains the values of the PRO and activity tracking variables $\mathcal{V}$ for patient p at month m: 56 PRO answers provided by the patient during the corresponding month, i.e. plus 3 aggregated values computed as the mean of the daily wearable device data (step count, calories, number of sleep hours) collected during the same month. Given a variable V_k, $x^p_{i,j}[V_k]$ denotes the V_k value in $x^p_{i,j}$;
- y^p_j is the value of the outcome o measured during the hospital visit at the end of period j.

The second sample set, denoted as $Sample^{FI}_o$, is built by adding to each feature vector $x^p_{i,j} \in Sample_o$, the FI value computed from the clinical variables measured during the hospital visit at the beginning of period j, i.e. at month 0 when $j = 1$ and 9 when $j = 2$.

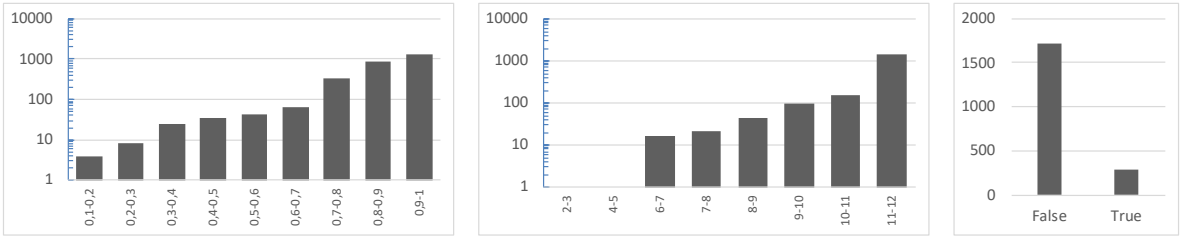
Quality Assurance. Lastly, we performed a quality assurance step on the resulting sample sets. Within each time window, observations of PRO variables are sometimes incomplete, resulting in sequence gaps. The size of the gaps is 5 consecutive missing observations on average, with a max of 17, and we found 108 gaps per patient on average, with a max of 284 gaps (regardless of size). We performed imputation by interpolating missing data points in the time series, with an aim to achieve a balance between the size of the gaps and the performance of the predictive model. Clearly, interpolating very large gaps produces spurious data in the training set. We experimentally determined the max size of gaps that could be safely interpolated (five missing steps), by assessing the predictive performance of each of the models resulting from training sets obtained from more or less "aggressive" interpolation. After adjusting for missing data, the final training set contains 2,250 data points, with an average of 8 per patient. The construction of the dataset results in at most 16 samples per patient, for a total of 4,176 records, considering each month for each patient.

4 THE LEARNING FRAMEWORK

The objective of the data-driven approach is to predict each outcome at the time of a visit using the samples referring to the time window before that visit, as shown in Fig. 2.

To this end, the data-driven learning framework we built is depicted on the left side of Fig. 3. Each outcome o is predicted by two learning models M_o and M^{FI}_o, one for each of the two sample sets $Sample_o$ and $Sample^{FI}_o$, respectively. We trained the two models separately and assessed the performance using standard KFold cross-validation (CV) on an 80% of the samples (χ_{train}) and a test phase on the remaining 20% samples (χ_{test}) of the corresponding sample set.

The DD approach is compared to the KD approach that represents the common practice in geriatric medicine. To this end, we built the KD learning framework depicted on the right hand side of Fig. 3. The approach aims at computing an ICI scores for each observation by manually selecting a subset $V = \{V_1 \dots V_n\} \subseteq \mathcal{V}$ of the set of PRO and activity tracking variables $\mathcal{V}$, specifying functions $s_i(x)$ to map each value x for variable $V_i \in \mathcal{V}$ to a score, and finally combining the individual scores $s_i(x)$ into a unique value. The variables are chosen to represent each of the five IC domains, namely locomotion, cognition, psychological, vitality and sensory capacity. For most of the variables V_i, a binary score is defined, i.e., $s_i(x) \in \{0, 1\}$, based on a single threshold, for instance when $V_i = stress\ level$ (from 1 to 10) the score is mapped



(a) QoL distribution (b) SPPB distribution (c) Falls distribution

Figure 1: Distribution of outcomes in the dataset

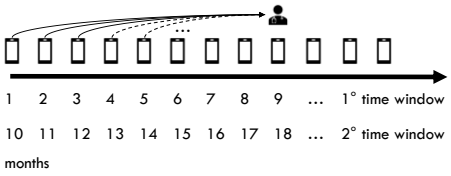


Figure 2: Prediction at next clinical visit using only patient reported outcomes

to 1 if the value is lower than 3 and 0 otherwise. Other variables are mapped to a score in the [0, 1] range, for instance the number of steps per day.

Then, given a feature vector $x^p_{i,j}$ for patient p, the corresponding ICI value $ICI(i, j, p)$ is computed as the sum of the $s_i(x^p_{i,j}[V_i])$ scores, normalised by the number of variables:

$$ICI(i, j, p) = \frac{\sum_{i:1}^{|V|} s_i(x^p_{i,j}[V_i])}{n}$$

Such an index is subject to an inevitable bias: the imposition of the physician's interpretation on the choice of the variables of the subset, as well as on the thresholds and the arithmetic formula to be used.

Also for this case, a dataset consisting exclusively of the ICIs, $Sample^{ICI}_o$, and one consisting of the ICIs and the FI at the most recent visit, $Sample^{ICI,FI}_o$, was isolated. This parallelism with what was done in the data-driven approach allowed to train 6 learning models with the datasets just described (M^{ICI}_o and $M^{ICI,FI}_o$ for each of the three outcomes o) and to compare the predictive performances between the two different approaches.

5 EXPERIMENTAL RESULTS AND MODEL INTERPRETATION

In this section we first discuss the performance of the predictive models, and then describe in detail our approach and results regarding *model interpretation*, which is reported as output to the DD approach in the left side of Fig. 3. The latter is a fundamental requirement in medicine, where the ability to provide medical doctors with an easy-to-understand interpretation of the model predictions is fundamental. This not only conveys confidence in the predictions, but also helps to make them actionable, i.e., in the form of recommendations to patients. We present examples of such interpretations, and their practical relevance, in Sec. 5.2.

The Gradient Boosting algorithm [13] proved to offer better predictive performance than other popular intelligible learning frameworks such as GA^2M [15], suggesting that separating model

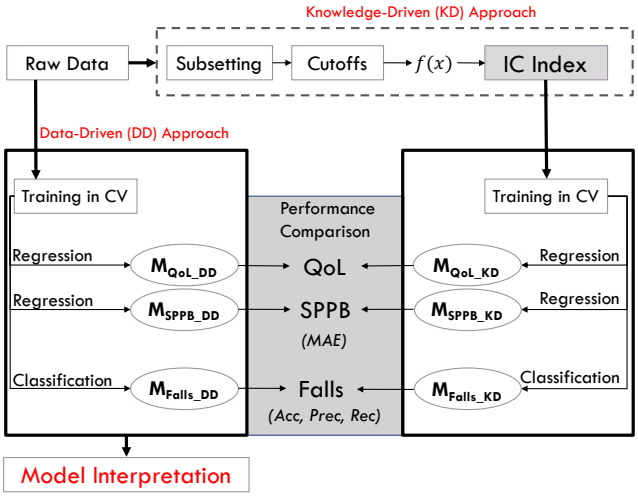


Figure 3: Comparison between Data-Driven and Knowledge-Driven approaches

performance from model interpretability would better suit our needs. Thus, our results are based on a combination of Gradient Boosting (the XGBoost implementation [4] for performance (Sec. 5.1), and Shapley Values [23], using the SHAP implementation [11]) to generate reports on the relative importance of each individual feature, both across the entire population and for individual patients (Sec. 5.2).

5.1 Predictive Performance

In our evaluation we compare our DD approach, illustrated in the left side of Fig. 3, with the KD approach based on a regression on the IC index (right hand side in the Figure). Fig. 4 shows the predictive performance of the models we tested, namely: the DD models trained with and without using FI as a feature, earlier referred to as M^{FI}_o and M_o, respectively; and the KD models, where again the expert may or may not consider FI. Results are presented using 1-MAPE (Mean Average Percentage Error) for the numerical outcomes QoL and SPPB on the left of the figure, and accuracy, precision, recall and F1 for Falls, on the right.

The results indicate a higher than 90% 1-MAPE for all cases in QoL and SPPB, while classification accuracy for Falls is higher than 84%. Further, the DD approach performs generally better than KD, and both benefit from using FI, with performance reaching 94.3%, 94.9% and 95% for QoL, SPPB, and Falls, respectively.

To note, in one case the KD approach returns a very low Recall when FI is not used. This can be explained by the strong

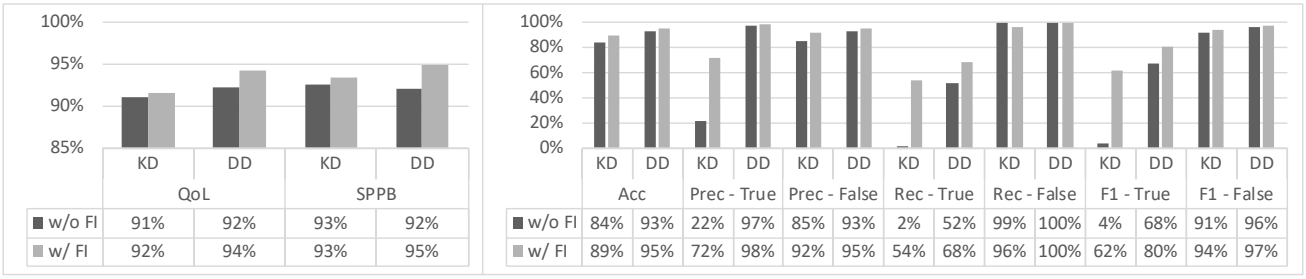


Figure 4: Predictive Performance. Left: 1-MAPE (Mean Average Percentage Error) for QoL and SPPB, right: classification effectiveness for Falls

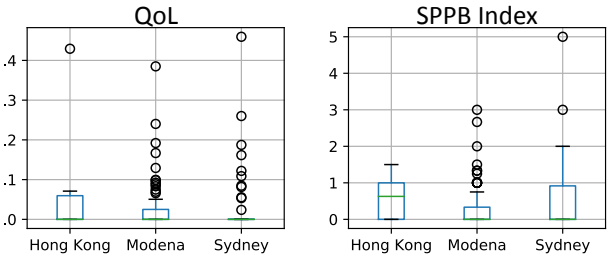


Figure 5: Regression MAE distribution per patient grouped by clinical center

imbalance of the majority "False" class (no Falls) relative to the small minority "True".

The training sets used to generate these models combine patients from all three clinics. To account for possible differences in data collection protocols between the clinics, we also created one separate model for each. The corresponding results are presented in Table 1 and are consistent with those presented above. Some anomalies appear in the Hong Kong models, and these are probably due to the small size of the training set.

Finally, Fig. 5 shows the MAE distribution grouped per clinical center for QoL and SPPB. This helps understanding the robustness of the models and to identify any non-homogeneity in the data. In particular, Hong Kong exhibits a higher number of outliers compared to Modena and Sydney, probably because of the small number of cases (33, compared to 128 in Modena and 100 in Sydney), which are also more homogeneous. These results suggest that developing separate models by stratifying across clinics and data collection centres may be beneficial for future, larger scale studies.

5.2 Model Interpretation

SHAP [11] is a framework for interpreting predictions from machine learning models. It is based on *Shapley values*, first introduced in 1953 in the context of cooperative game theory [23]. Briefly, the main goal of the framework is to rank the relative influence of each feature on a predictive model, both locally, that is, for a specific instance prediction, and globally, i.e., when considering the model predictions for an entire population.

In our medical setting, this means that for each patient, in addition to the predicted outcome the clinician also receives a list of features, ranked in order of their relative importance in achieving the prediction. Importantly, these orders may differ for any two patients. This means that, using SHAP, we enable forms of personalised medicine whereby similar outcomes are explained in terms of different behaviour, represented for instance by different EMA features. An example appears in Fig. 6, showing two different sets of positively contributing (green) and negatively contributing (red) features for two patients with the *same* SPPB index (note that, for SPPB, higher is better, as this indicates the patient's capacity of physical movement. In the case of Falls, for instance, the opposite would be true). Clearly, this added information may lead to different interventions for these two patients.

At the same time, SHAP provides global explanations, which characterise the contribution of each feature as a function of its range of values. For instance, Fig. 7 shows how the SV, indicating the overall contribution of one of these features (a PRO question), goes from negative to positive depending on the patients' responses to this question, with a definite threshold of $\geqslant 3$.

We note that this capability essentially mimics the KD approach in that it identifies thresholds for the variables. While these are similar to the manually selected cutoffs, in our DD approach these are automatically identified from the data, in a principled way. In the future, this explanation capability may underpin epidemiological studies where the precise characterisation of a populations of individuals enables new forms of preventive medicine.

6 CONCLUSIONS AND FUTURE WORK

In this paper we have proposed a novel, data-driven approach towards the definition of *Intrinsic Capacity*, aimed at quantifying and predicting the wellness state of old people who live with HIV. Using a cohort from a multi-centre prospective study as training set, we have shown that a machine learning model that predicts three specific wellness metrics (Falls, SPPB, and Quality of Life), performs equally or better than a manually-defined *Intrinsic Capacity Index*. At the same time, the model is interpretable, making it an ideal complement to expert-based assessment of wellness.

REFERENCES

[1] Ilaria Bellantuono, Rafael DeCabo, Dan Ehninger, et al. 2018. Find drugs that delay many diseases of old age. *Nature* 554 (02 2018), 293–295.
[2] Giulia Belloni and Matteo Cesari. 2019. Frailty and Intrinsic Capacity: Two Distinct but Related Constructs. *Frontiers in Medicine* 6 (06 2019).
[3] Thomas Brothers, Susan Kirkland, Giovanni Guaraldi, et al. 2014. Frailty in People Aging With Human Immunodeficiency Virus (HIV) Infection. *The Journal of infectious diseases* 210 (06 2014).
[4] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proc. of ACM SIGKDD (KDD '16)*. ACM, New York, NY, USA, 785–794.
[5] Nancy Devlin and Richard Brooks. 2017. EQ-5D and the EuroQol Group: Past, Present and Future. *Applied Health Economics and Health Policy* 15 (02 2017).
[6] Iacopo Franconi, Olga Theou, Lindsay Wallace, et al. 2018. Construct validation of a Frailty Index, an HIV Index and a Protective Index from a clinical HIV database. *PLOS ONE* 13 (10 2018), e0201394.

	QoL		SPPB Index		Falls													
	1 - MAPE		1 - MAPE		Acc		P True		P False		R True		R False		F1 True		F1 False	
	KD	DD	KD	DD	KD	DD	KD	DD	KD	DD	KD	DD	KD	DD	KD	DD	KD	DD
Hong Kong																		
w/o FI	93%	93%	91%	92%	84%	**96%**	0%	0%	87%	**100%**	0%	0%	97%	97%	0%	0%	92%	**98%**
w/ FI	94%	**93%**	**94%**	93%	94%	93%	1%	1%	94%	93%	33%	33%	**100%**	**100%**	50%	**50%**	97%	96%
Modena																		
w/o FI	94%	94%	94%	95%	86%	94%	0%	**93%**	86%	94%	0%	41%	**100%**	99%	0%	57%	93%	96%
w/ FI	94%	**94%**	95%	**96%**	93%	**95%**	74%	**93%**	95%	**96%**	53%	**68%**	98%	99%	62%	**79%**	97%	**98%**
Sydney																		
w/o FI	88%	90%	91%	93%	81%	87%	68%	76%	84%	89%	38%	57%	95%	95%	49%	65%	89%	92%
w/ FI	89%	**90%**	93%	**94%**	87%	**95%**	86%	**93%**	88%	**96%**	**69%**	68%	95%	**99%**	77%	**79%**	91%	**98%**

Table 1: Single-clinic models performance. (left: predictive performance for QoL and SPPB, right: classification effectiveness for Falls)

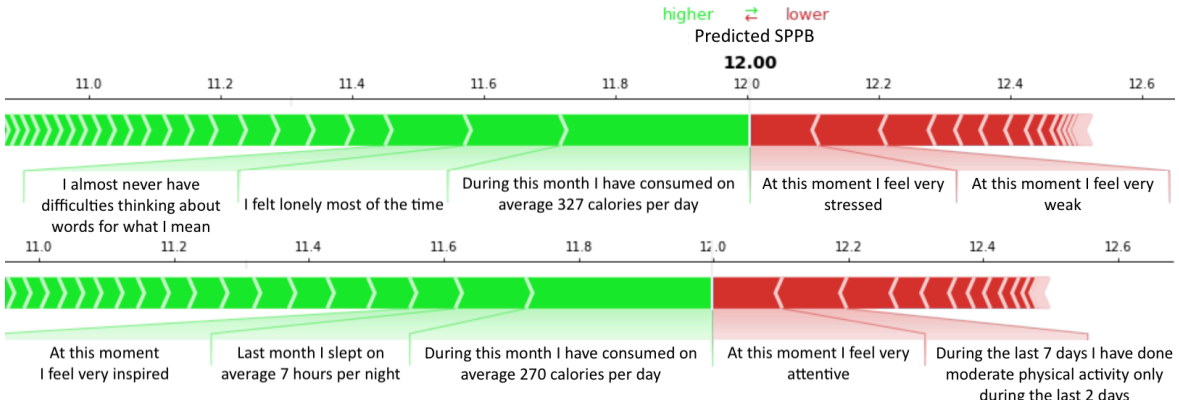


Figure 6: Example of a local interpretation of one patient's SPPB prediction. The 5 most relevant Shapley Values are reported.

Figure 7: Global distribution of one of the PRO's SVs based on the value of the possible answers.

[7] Leilani H. Gilpin, David Bau, Ben Z. Yuan, et al. 2018. Explaining explanations: An overview of interpretability of machine learning. In *Proc. DSAA 2018*. 80–89.

[8] Giovanni Guaraldi and Jovana Milic. 2019. The Interplay Between Frailty and Intrinsic Capacity in Aging and HIV Infection. *AIDS Research and Human Retroviruses* 35 (08 2019).

[9] Giovanni Guaraldi, Mirko Orsini, Agnese Caselgrandi, et al. 2019. Fitness tracking wearable devices and a dedicated smart phone app (MySAwH App) to predict quality of life in PLWH: a multi-centre prospective study.. In *17th European AIDS Conference (EACS)* (2019-08-05).

[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, et al. 2018. A survey of methods for explaining black box models. *Comput. Surveys* (2018).

[11] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.

[12] Scott M Lundberg, Bala Nair, Monica S Vavilala, et al. 2018. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *Nature Biomedical Engineering* 2, 10 (2018), 749–760.

[13] Llew Mason, Jonathan Baxter, Peter L Bartlett, et al. 2000. Boosting algorithms as gradient descent. In *Advances in neural information processing systems*. 512–518.

[14] Ankur Teredesai Muhammad Aurangzeb Ahmad, Carly Eckert et al. 2018. Interpretable Machine Learning in Healthcare. *IEEE Intelligent Informatics Bulletin* 19, 1 (2018), 1–7.

[15] Harsha Nori, Samuel Jenkins, Paul Koch, et al. 2019. InterpretML: A Unified Framework for Machine Learning Interpretability. *arXiv preprint arXiv:1909.09223* (2019).

[16] World Health Organization. 2015. World report on aging and health. https://www.who.int/ageing/events/world-report-2015-launch/en/

[17] World Health Organization. 2018. Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

[18] Mirko Orsini, Marco Pacchioni, Andrea Malagoli, et al. 2017. My smart age with HIV: An innovative mobile and IoMT framework for patient's empowerment. In *2017 IEEE 3rd International Forum on Research and Technologies for Society and Industry (RTSI)*. 1–6.

[19] Nathan D Price, Andrew T Magis, John C Earls, et al. 2017. A wellness study of 108 individuals using personal, dense, dynamic data clouds. *Nature Biotechnology* 35 (jul 2017), 747.

[20] Mashfiqui Rabbi, Shahid Ali, Tanzeem Choudhury, et al. 2011. Passive and in-situ assessment of mental and physical well-being using mobile sensors. In *Proc. of UbiComp'11*. 385–394.

[21] Martin Ritt, Karl Gassmann, and Cornel Sieber. 2016. Significance of frailty for predicting adverse clinical outcomes in different patient groups with specific medical conditions. *Zeitschrift fur Gerontologie und Geriatrie* 49 (09 2016).

[22] Samuel Searle, Arnold Mitnitski, Evelyne Gahbauer, et al. 2008. A standard procedure for creating a frailty index. *BMC geriatrics* 8 (10 2008), 24.

[23] Lloyd Stowell Shapley. 1953. *A Value for n-Person Games*. Contributions to the Theory of Games, Vol. 2. Princeton University Press, Chapter 17.

[24] Catherine Tong, Matthew Craner, Matthieu Vegreville, et al. 2019. Tracking Fatigue and Health State in Multiple Sclerosis Patients Using Connnected Wellness Devices. *Proc. of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (sep 2019), 1–19.

[25] Berk Ustun and Cynthia Rudin. 2015. Supersparse Linear Integer Models for Optimized Medical Scoring Systems. *Machine Learning* 102 (02 2015).

[26] Petar Veličković, Laurynas Karazija, Nicholas D. Lane, et al. 2018. Cross-modal Recurrent Models for Weight Objective Prediction from Multimodal Time-series Data. In *Proc. of PervasiveHealth '18 (PervasiveHealth '18)*. 178–186.

[27] Danding Wang, Qian Yang, Ashraf Abdul, et al. 2019. Designing theory-driven user-centric explainable AI. In *Proc. of Conference on Human Factors in Computing Systems*.

[28] Rui Wang, Emily A. Scherer, Megan Walsh, et al. 2018. Predicting Symptom Trajectories of Schizophrenia Using Mobile Sensing. *GetMobile: Mobile Computing and Communications* (2018).

[29] Rui Wang, Weichen Wang, Alex DaSilva, et al. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. *Proc. of ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.