A Semi-Supervised Multi-Task Learning Approach for Predicting Short-Term Kidney Disease Evolution

Michele Bernardini, Luca Romeo, Emanuele Frontoni, Massih-Reza Amini

A Semi-Supervised Multi-Task Learning Approach for Predicting Short-Term Kidney Disease Evolution

Michele Bernardini, Luca Romeo, Emanuele Frontoni, *Senior Member, IEEE,* and Massih-Reza Amini

***Abstract*—Kidney Disease (KD) may hide complex causes and is associated with a tremendous socio-economic impact. Timely identification and management from the first level of medical care represent the most effective strategy to address the growing global burden sustainably. Clinical practice guidelines suggest utilizing estimated Glomerular Filtration Rate (eGFR) for routine evaluation within a screening purpose. Accordingly, the analysis of Electronic Health Records (EHRs) using Machine Learning techniques offers great opportunities to monitor and predict the eGFR trend over time. This paper aims to propose a novel Semi-Supervised Multi-Task Learning (SS-MTL) approach for predicting short-term KD evolution on multiple General Practitioners EHR data. We demonstrated that the SS-MTL approach can (i) capture the eGFR temporal evolution by imposing a temporal relatedness between consecutive time windows and (ii) exploit useful information from unlabeled patients when labeled patients are less numerous with a gain of up to 4.1 % in terms of *Recall*. This situation reflects the real-case scenario, where available labeled samples are limited, but those unlabeled much more abundant. The SS-MTL approach, also given the high level of interpretability, might be the ideal candidate in general practice to get integrated within a decision support system for KD screening purposes.**

***Index Terms*—Machine Learning, Semi-Supervised Learning, Multi-Task Learning, Electronic Health Record, General Practice, Kidney Disease.**

I. Introduction

Kidney Disease (KD), often incautiously underestimated as a comorbidity of diabetes or hypertension, may hide complex causes and is associated with a tremendous socio-economic impact [1], [2]. Worldwide, 19 million disability-adjusted life-years were directly attributable to a reduced Glomerular Filtration Rate (GFR), which measures the health-state of kidney functionality [3]. According to World Health Organization (WHO) recommendations, if KD is early-diagnosed and an effective screening strategy is adopted, the worsening of kidney function can be slowed or averted by inexpensive interventions [4]. Thus, the timely identification and management of chronic KD (CKD) from the first level of medical care (e.g., General Practitioners (GPs)) represent the most effective strategy to address the growing global burden sustainably. Most recent clinical practice guidelines suggest utilizing estimated GFR (eGFR) for routine evaluation within a screening purpose, rather than a GFR measure, needed when an accurate assessment is required [5]. The 6 CKD stages strictly based on eGFR values serve to assess the kidney functionalities [6]. Accordingly, the analysis of Electronic Health Records (EHRs) using Machine Learning (ML) techniques offers a great opportunity to monitor the eGFR trend over time and predict its value in the short-term period. Unfortunately, in a real-case scenario, EHRs collected by GPs include several challenges such as multi-source and non-standardized data, incomplete or missing values, registration errors, data sparsity, privacy-preserving, etc [7].

Patients are followed over some time by GPs, which, at each visit, store a large variety of clinical events (i.e., exam prescriptions, medications, pathologies, lab tests, etc). Thus, eGFR evolution can be modeled using Multi-Task Learning (MTL) approach [8], [9], where the prediction of the eGFR status at a single time point is considered as a task and the predictive models at different time points may be similar because temporally related. Differently from the intensive care unit EHR datasets [10], in GPs scenario the limited availability of i) patients (i.e., spatial-transversal data) and/or ii) patients' medical history (i.e., time-longitudinal data) precludes an adequate labeled sample size (i.e., annotation of eGFR status over time) for exploiting a robust and representative supervised learning strategy. Usually, labeled data are expensive to collect and unlabeled data are abundant. Accordingly, also in the best-case scenario where a large amount of transversal and longitudinal data is available, the label might be sparsely distributed over time. This point is a crucial issue in the clinical-use case, where data labeling is prohibitive (especially for the healthy subjects) and possibly captures only the most important events of pathological subjects, and besides, unlabeled data are abundant.

Starting from these motivations, the work aims to propose a novel Semi-Supervised Multi-task Learning (SS-MTL) approach for predicting short-term KD evolution on multiple GPs EHR data. The SS-MTL approach combines a Semi-Supervised Learning (SSL) strategy with an MTL procedure to i) impose a temporal relatedness between consecutive time windows to predict the eGFR status over time and ii) exploit both labeled and unlabeled samples in the learning procedure for capturing high-discriminative temporal patterns. Thus, two research questions (RQs) are formulated to measure the effectiveness of the proposed approach for state-of-the-art approaches:

- ***RQ1:*** *Is the MTL approach capable to capture the eGFR temporal evolution?*
- ***RQ2:*** *Is the SS-MTL approach capable to capture useful information from unlabeled patients?*

The paper is organized as follows: Section II gives an

M. Bernardini, L. Romeo and E. Frontoni are with Department of Information Engineering, Università Politecnica delle Marche, Ancona, Italy. (e-mails: m.bernardini@pm.univpm.it, l.romeo@univpm.it, e.frontoni@univpm.it)

M.R. Amini is with Grenoble Informatics Laboratory, Université Grenoble Alpes, Saint-Martin-dHères, France. (e-mail: massih-reza.amini@univ-grenoble-alpes.fr)

overview of the state-of-the-art on the MTL and SSL approaches in predictive medicine using EHR data; Section III describes the *m*FIMMG dataset and the preprocessing procedure; Section IV describes the proposed SS-MTL approach and the experimental comparisons; Section V shows the predictive performance and pattern localization results; Section VI answers the two proposed RQs and discusses the clinical significance, limitations and future work; Section VII presents the conclusions of the work.

II. Related Work

Machine Learning techniques have been already adequately proven to be effective in dealing with sequential temporal data in many applicative research areas, including especially healthcare scenarios. In particular, EHR data have been largely exploited to accomplish predictive tasks such as stages of chronic diseases, disease complications, intensive care unit clinical events, etc. These approaches spread from standard ML models such as Logistic Regression (LR) [11], [12], [13], Decision Tree (DT) [14], Random Forest (RF) [11], [13], Gradient Boosting Tree (Boosting) [11], Support Vector Machine (SVM) [14], [15] to more appealing and complex Deep Learning (DL) frameworks, mostly based on feedforward [16], Long-Short Term Memory (LSTM) [11], and Convolutional Neural Network [11] architectures.

The MTL approach is a well-known and consolidated learning paradigm to address health informatics and clinician prediction tasks, capable of extracting useful information from multiple related tasks and improve the overall generalization performance [17]. In [16], [18] authors tried to answer when MTL improved prediction performance for different clinical tasks using EHR data. Multi-task feedforward [16] and multi-task LSTM networks [18] were compared with baseline single task networks and LR models. Most related to our work is the paper [9], where a temporal MTL was adopted to stratify the risk of renal function deterioration. In fact, the different clinical tasks do not differ by their intrinsic nature (i.e., eGFR prediction), but from their temporal evolution (i.e., time windows). Differently from [9], in our work, this problem is modeled as an SSL scenario, where the label is sparse over time. Additionally, the whole raw EHR data is used rather than performing a feature selection for each task, so as to potentially avoid a lack of relevant information to detect hidden patterns.

As mentioned before, in GPs EHR data, labeled data are expensive to collect and unlabeled data are abundant. Moreover, even if originally a huge amount of labeled data is available, during a real-case scenario usually happens that after the preprocessing stage (e.g., inclusion/exclusion criteria) a considerable amount of labeled data is going to be reduced [9], [19]. Thus, the precondition of collecting a huge labeled sample size is necessary, but not easily satisfied especially in the GP scenario where large and publicly available datasets are limited. In [20], [21] the training labeled sample size was augmented using GANs and conditional GANs, respectively, without considering unlabeled data. Given this operational necessity to retrieve labeled information, MTL could be combined with SSL, leading to Semi-Supervised Multi-Task Learning (SS-MTL) paradigm, where a training set of each task consists of both labeled and unlabeled data to exploit useful information contained in the unlabeled data in order to further improve the MTL performance. A similar rationale was proposed in [22], where a multi-task setting based on an SSL technique, named Positive and Unlabeled learning (PU), was implemented for addressing a disease gene prioritization problem. A different SSL technique (i.e., Label Propagation [LP]), which constructs a similarity graph over all input data, was proposed in [23] to generate personalized drug recommendations by leveraging patient similarity and drug similarity analytics. In our proposed approach, the Self-Learning Algorithm (SLA) inspired from [24] is utilized as SSL paradigm, which during the training stage (i.e., negative and positive samples), iteratively assigns pseudo-labels to the set of unlabeled training samples that have their margin above a threshold automatically achieved from this bound.

After evaluating the state-of-the-art, our proposed SS-MTL approach represents the first attempt to combine the SSL paradigm in an MTL scenario where the main goal is to predict the eGFR evolution based on EHR data.

Therefore, the applicative theoretical novelty of this work actively contributes to the biomedical informatics when a large number of unlabeled samples and a temporal relatedness between consecutive tasks are involved. In this work, the SS-MTL approach, capable to predict and explain the short-term KD evolution, contributes to improve the KD management especially at an early stage. Thus, in general practice, the SS-MTL approach may be integrated in a decision support system for screening purposes.

III. Data

The publicly available *m*FIMMG dataset[1], which is extracted from the standardized FIMMG Netmedica Cloud computing infrastructure [25], [26], stores a 10-year ($2010-2019$) activity collected by 6 GPs and consists of 14175 patients and 6 main fields. The demographic field is composed of age and gender. The monitoring field (i.e., diastolic and systolic blood pressure, height, weight, and waist) contains only continuous predictors, as well as the lab tests field where all the laboratory outcomes (e.g., eGFR) are stored. The remaining fields (i.e., pathologies, drugs, exam prescriptions [exams]) are all categorical.

A. Preprocessing

Figure 1 shows all the preprocessing procedure: i) *eGFR*, ii) *Labeled samples*, and iii) *Temporal data*.

eGFR: The eGFR index was calculated by the authors using a unique CKD-EPI formula [27], [28], as a combination of 4 factors:

$$eGFR = f(creatinin, age, gender, race) \tag{1}$$

This rationale, that is adopted also in [9], mitigated the inter-laboratory variability. Among all patients, let call labeled

[1] https://vrai.dii.univpm.it/content/mfimmg-dataset

samples (#5812) the subset of patients whose at least a single eGFR index is known, unlabeled samples (#8363) the remaining.

***Labeled samples**:* Let t_g the time-stamp of the last eGFR observation, the previous 1-year time-stamp is defined as:

$$t_t = t_g - 12 \text{ months} \tag{2}$$

Among all labeled samples only those which satisfy the following criteria were selected:

- At least a single observation of all fields (#5494);
- At least 2-year eGFR medical history before t_t, that must include 2 or more eGFR observations (#2176).

Table I shows the eGFR distribution at t_g time-stamp of the *selected* samples (i.e., from now on mentioned as labeled samples) in according with the CKD stages [6]. The remaining samples named *discarded samples* (#3636) from now on were merged with unlabeled samples and named as such (#11999).

Additionally, for each field, only features whose appearances are less than 5% of the total of labeled samples were excluded. Regarding the monitoring field, only the blood pressure feature is over cut, but it was then grouped with the lab tests field because of the same continuous nature. From now on, all the included features were named predictors.

TABLE I: Distribution of eGFR for the labeled samples (#2176) in according with the CKD stages.

CKD stage	**eGFR** [ml/min/1.73m^2]	**%**
I	≥ 90: normal	19.35
II	6089: mild reduction	53.31
IIIa	4559: mild-moderate reduction	16.59
IIIb	3044: moderate-severe	7.49
IV	1529: sever reduction	2.85
V	< 15: kidney failure	0.41

***Temporal data**:* Following [9], 6-month granularity was chosen to define a time window. For each labeled sample, only the first five consecutive non-overlapping time windows (i.e., 2.5-year medical history) before t_t were chosen. If few time windows were chosen the eGFR temporal evolution could not be caught by the predictive model; on the other hand, the model would risk overfitting because the more observations the patient would have, the more the patient would tend to have chronic kidney complications (i.e., low eGFR values) [29]. Thus, for each field, all the patients that did not contain observations in any of the selected five-time windows were deleted (#2136). The information about eGFR before t_t was deleted only from the lab test field (i.e., eGFR continuous values) because already indirectly present through the predictors used in CKD-EPI formula (see Eq. 1), while from exams field was left (i.e., times of eGFR examination prescription). Finally, a supplementary field named 'Overall'- which consists of the aggregation of drugs, exams, lab test predictors only if they were fully shared by the same patient - was provided (#1833 samples and 494 predictors). Additionally, Overall* field included also demographic predictors (i.e., gender and age).

On the contrary, for each field of unlabeled samples, five random consecutive time windows were chosen if patients shared at least a single observation of the same predictors extracted from the labeled samples, by obtaining the final Overall and Overall* fields (#4996 samples).

Both categorical and continuous features were appropriately standardized during the preprocessing stage. The one-hot encoding was used on categorical features (i.e., pathologies, exams, drugs), while the z-score was used on continuous features (i.e., lab tests) by removing the mean and scaling to unit variance. Thus, categorical fields reflect the presence or the absence of a given pathology, drug, or exam without displaying any missing values. On the other hand, the continuous field (lab tests) may present missing values or outliers. For that reason, an outlier detection strategy based on scaled median absolute deviation and an extra-values imputation of missing values was performed for both labeled and unlabeled samples of the lab tests field. Table II shows the final configuration of the mFIMMG dataset after the preprocessing stage.

TABLE II: Final configuration of the mFIMMG dataset after the preprocessing stage.

	Pathologies	Drugs	Exams	Lab tests	**Overall**	**Overall***
Predictors	38	309	135	50	494	496
Total samples	5660	9533	9530	7479	6829	6829
Labeled samples	707	1853	1887	1877	1833	1833
Unlabeled samples	4953	7680	7643	5602	4996	4996

IV. Method

The binary classification task consists of predicting the short-term (1-year) eGFR evolution. Given the longitudinal information of each patient, according to Table I we suppose to predict CKD stage I (e.g., negative or normal samples, y^-) from the others (e.g., positive or risky samples, y^+).

Firstly, in Sec. IV-A a summary of the mathematical notations used from now on is provided. Then, in Sec. IV-B starting from baseline approaches such as no-temporal and stacked-temporal (see Figure 2a and Figure 2b), the proposed SS-MTL approach is thoroughly described in Sec. IV-C.

A. Notations

The main mathematical notations used in the following Sec. IV were summarized in Table III.

B. Baseline approaches

***No-temporal**:* In this approach, the continuous predictors were averaged across all time windows, while the categorical ones were aggregated. Even if the temporal information has vanished, this approach handles the challenge of irregular sampling and missing values.

***Stacked-temporal**:* In this approach temporal information was preserved by concatenating longitudinally all the time windows. This approach can capture temporal information across time windows, but it may suffer from overfitting, considering the increasing number of predictors which is directly proportional to the number of time windows.

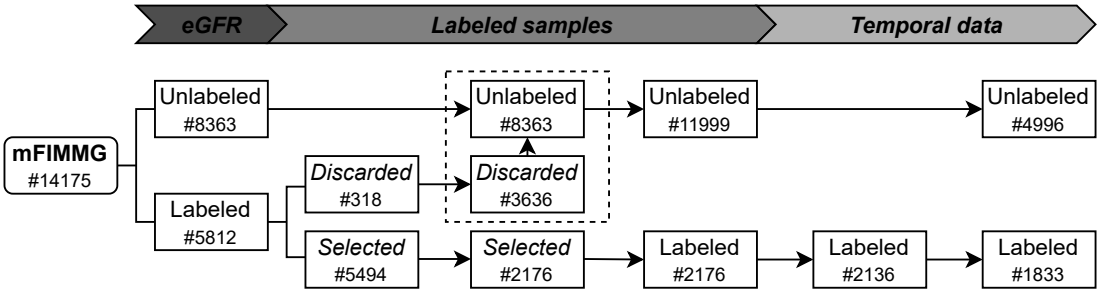

Fig. 1: mFIMMG dataset preprocessing: labeled and unlabeled samples.

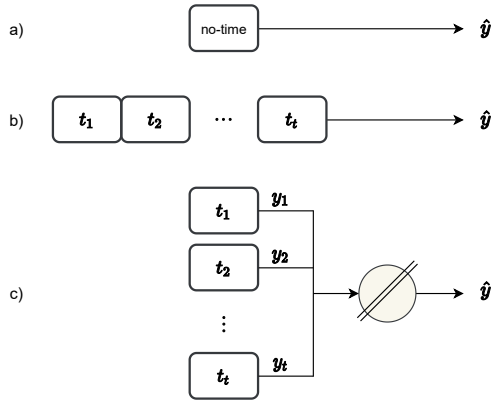

Fig. 2: Three different approaches. a) No-temporal: the temporal information is averaged across all time-windows; b) Stacked-temporal: the temporal information is preserved by concatenating longitudinally all the time windows; and c) Multitask-temporal: each time-window is treated as a separate task.

TABLE III: Notations.

Symbol	Description
n	# of samples
d	# of predictors
t	# of tasks (time-windows)
$X \in \mathbb{R}^{n \times d}$	Observations
- $x \in Z_l$	- labeled
- $x^{'} \in V_u$	- unlabeled
- $\widetilde{x}^{'} \in \widetilde{Z}_u$	- pseudolabeled
- $\widetilde{x} \in \widetilde{Z}$	- labeled and pseudolabeled
$W \in \mathbb{R}^{d \times t}$	Weights
$Y \in \mathbb{R}^{n \times t}$	Targets
- $y \in Z_l$	- labeled
- $y^{'} \in V_u$	- unlabeled
- $\widetilde{y}^{'} \in \widetilde{Z}_u$	- pseudolabeled
- $\widetilde{y} \in \widetilde{Z}$	- labeled and pseudolabeled
$\hat{y} \in \mathbb{R}^{n \times 1}$	Target predictions

C. Semi-Supervised Multi-Task Learning (SS-MTL)

In the following subsection the SS-MTL approach is introduced by providing: i) multi-task temporal Lasso formulation (see Sec. IV-C1), ii) Self-Learning Algorithm formulation (see Sec. IV-C2), and iii) SS-MTL approach implementation (see Sec. IV-C3).

1) Multi-Task Learning (MTL): multi-task temporal Lasso:

Multitask-temporal: In this approach (see Figure 2c) the temporal information was handled as a MTL problem. Each time-window was treated as a separate task and then, the resulting intermediate outputs $(y_1, y_2, ..., y_t)$ were combined to obtain the final prediction $\hat{y}$.

Considering the following MTL problem with t tasks (time windows), n samples, and d predictors, the model encodes the temporal information using regularization terms. Let $\{x_1, ..., x_n\}$ be the input data and $\{y_1, ..., y_n\}$ be the targets, where each $x_i \in \mathbb{R}^d$ represents a sample, and $y_i \in \mathbb{R}^t$ is the corresponding target at different time-windows. We denote $X = [x_1, ..., x_n]^T \in \mathbb{R}^{n \times d}$ as the data matrix, $Y = [y_1, ..., y_n]^T \in \mathbb{R}^{n \times t}$ as the target matrix, and $W = [w_1, ..., w_t] \in \mathbb{R}^{d \times t}$ as the weight matrix. The whole formulation of multitask temporal Lasso is given by [8], [30]:

$$\min_W \quad L(W) + \rho_1 \|W\|_F^2 + \rho_2 \sum_{i=1}^{t-1} \|W_i - W_{1+1}\|_F^2 + \rho_3 \|W\|_{2,1} \tag{3}$$

where $L(W)$ is the loss function and ρ_1, ρ_2, ρ_3, represent the regularization penalties: the first penalty controls the complexity of the model; the second penalty couples the neighbor tasks, encouraging every two neighbor tasks to be similar (temporal smoothness); and the third penalty induces the grouped sparsity, which performs the joint feature selection on the tasks at different time windows (longitudinal feature selection). The temporal information is modeled as a type of graph regularization A, where neighbor tasks are coupled via edges. A is the structure matrix which encodes the task relatedness. In the temporal group Lasso formulation, A is defined as an $(t-1) \times t$ sparse matrix, in which $A_{i,i} = 1$ and $A_{i,i+1} = -1$; and thus, the formulation can be written in a simpler form:

$$\min_W \quad L(W) + \rho_1 \|W\|_F^2 + \rho_2 \|WA\|_F^2 + \rho_3 \|W\|_{2,1} \tag{4}$$

However, this formulation assumes that for each sample a predictor is simultaneously selected or not at all time windows. The convex fused sparse group Lasso (CFG) formulation overcomes this issue [30]:

$$\min_W \quad L(W) + \rho_1 \|W\|_1 + \rho_2 \|AW^T\|_1 + \rho_3 \|W\|_{2,1} \tag{5}$$

Accordingly, the CFG with Logistic loss model solves the CFG regularized multi-task Logistic regression problem:

$$\min_{W,c} \quad \sum_{i=1}^{t}\sum_{j=1}^{n_i} \log\Big\{1+exp\Big[-Y_{i,j}\Big(W_j^T X_{i,j}+c_i\Big)\Big]\Big\}+ \\ +\rho_1\|W\|_1+\rho_2\|AW^T\|_1+\rho_3\|W\|_{2,1} \tag{6}$$

where ρ_3 controls group sparsity for joint feature selection, while ρ_1, which controls element-wise sparsity and ρ_2 which controls the fused regularization represent the parameters for the fused Lasso.

2) Semi-Supervised Learning (SSL): Self-Learning Algorithm (SLA): SSL, also referred as learning with partially labeled data, concerns the case where a prediction function is learned on both labeled and unlabeled training samples. Unlabeled training samples may contain valuable information on the prediction problem at hand which exploitation may lead to a performant prediction function. For a binary classification scenario, we define a set of labeled training samples $Z_l = \{(x_i, y_i) \mid i = 1,...,l\}$ and a set of unlabeled training samples $V_u = \{x_i^{'} \mid i = l+1,...,l+u\}$.

Considering learning algorithms that work in a fixed hypothesis space H of binary classifiers and given the whole training set $S = Z_l \cup V_u$, the task of the learner $h \in H$ is to choose a posterior distribution Q over H such that the Q-weighted majority vote classifier B_Q (i.e., Bayes classifier) will have the smallest possible risk on samples of V_u. Defining the Bayes classifier:

$$B_Q(x) = sign[E_{h\sim Q}h(x)] \tag{7}$$

We can define its empirical error over the unlabeled set V_u, called the transductive risk, as:

$$R_u(B_Q) = \frac{1}{u}\sum_{x^{'}\in V_u}[B_Q(x^{'}) \neq y^{'}] \tag{8}$$

The corresponding Gibbs classifier, G_Q, is randomly chosen from the hypothesis space H according to the posterior distribution Q and its transductive risk over the unlabeled training set is defined by:

$$R_u(G_Q) = \frac{1}{u}\sum_{x^{'}\in V_u} E_{h\sim Q}[h(x^{'}) \neq y^{'}] \tag{9}$$

Note that these risks cannot be estimated as the labels of unlabeled examples are unknown.

In [31, Ch. 3] the margin of a Bayes classifier was shown to be an indicator of confidence respecting the cluster assumption in SSL which stipulates that the decision boundary passes through low-density regions. Supposing that we have a tight upper bound $R_u^{\delta}(G_Q)$ over the risk of the Gibbs classifier G_Q which holds with probability $1-\delta$, [24] showed that it is possible to bound the transductive risk of the Bayes classifier with high probability.

This result follows from a bound on the joint Bayes risk depending on a threshold θ :

$$R_{u\wedge\theta}(B_Q) = \frac{1}{u}\sum_{x^{'}\in V_u}[B_Q(x^{'}) \neq y^{'} \wedge m_Q(x^{'}) > \theta] \tag{10}$$

where $m_Q(\cdot) = |E_{h\sim Q}h(\cdot)|$ is the absolute value output of the Bayes classifier, denoted as the unsigned margin function.

This bound over the joint Bayes risk can be estimated by considering the distribution of unsigned margins regarding the threshold θ and it constitutes the working hypothesis of the margin-based Self-Learning Algorithms (SLA). This algorithm first trains a classifier on the labeled training set. The output of the learner can then be used to assign pseudolabels to unlabeled examples (denoted by the set $\widetilde{Z}_u$ in what follows) having a margin above a certain threshold θ and the supervised method is repeatedly retrained upon the set of the initial labeled and unlabeled examples that have been classified in the previous steps. The threshold θ is iteratively estimated at each step of the algorithm as the one which minimizes the conditional Bayes error defined as:

$$R_{u|\theta}(B_Q) = P_u(B_Q(x^{'}) \neq y^{'} \mid m_Q(x^{'} > \theta) = \frac{R_{u\wedge\theta}(B_Q)}{P_u(m_Q(x^{'} > \theta)} \tag{11}$$

In practice, the upper bound $R_Q^{\delta}(G)$ of the risk of the Gibbs classifier which is involved in the computation of θ in equation (see Eq. 8) is fixed to its worst value 0.5.

Algorithm SLA

Input: Labeled and Unlabeled training sets: Z_l, V_u
Initialize
Train a classifier H on Z_l
Set $\widetilde{Z}_u \leftarrow \varnothing$
repeat
  Compute the margin threshold θ from (see Eq. 8)
  S $\leftarrow \{ (x^{'}, y^{'}) \mid x^{'} \in V_u; m_Q(x^{'} \geq \theta \wedge y^{'} = sign(H(x^{'})) \}$
  $\widetilde{Z}_u \leftarrow \widetilde{Z}_u \cup S$, $V_u = V_u \backslash S$
  Learn a classifier H by optimizing a global loss function on Z_l and $\widetilde{Z}_u$
until V_u is empty or no adds to $\widetilde{Z}_u$;
Output: The final $\widetilde{Z} = Z_l \cup \widetilde{Z}_u$

3) Implementation of Semi-Supervised Multi-Task Learning (SS-MTL): The training experimental procedure adopted by our proposed method is shown in Figure 3.

At the beginning of the outer 10-fold cross-validation (10-CV) procedure, negative labeled samples y^{-} were around four times more numerous than those positive y^{+}, thus SMOTE [32] was utilized to balance the labeled samples ($y^{-} = y^{+}$). In our experiments, within the SLA algorithm, we considered two different Bayes classifiers, such as Decision Tree (DT) and SVM for Overall and Overall* fields, respectively. On the contrary for the other single fields, only the DT model was used. This rationale is justified by the fact that after having tested all the possible combinations of classifiers (i.e., LR, DT, RF, Boosting, SVM) within the SLA algorithm, in terms of predictive performance, the SVM resulted the best classifier for Overall* field, while the DT classifier for all the others.

During each SLA iteration, every candidate pseudo-label is chosen only if selected (i.e., above threshold θ) for all time

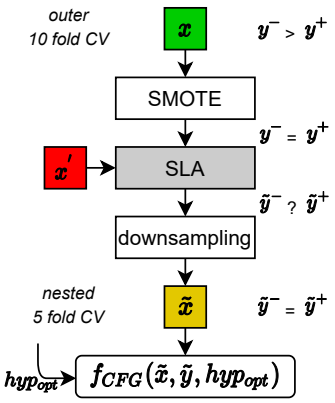

Fig. 3: SS-MTL: training experimental procedure.

windows. After that, the final prediction associated with the pseudo-label was selected by testing 3 different strategies (i.e., majority voting (majvot), unanimous, and Gibbs).

From the final SLA output $\in \widetilde{Z}$, the imbalance ratio between $\widetilde{y}^-$ and $\widetilde{y}^+$ is unknown ($\widetilde{y}^-$? $\widetilde{y}^+$), thus random downsampling over the pseudolabel majority class was performed in order to achieve again a balanced condition. The hyperparameters tuning was performed by implementing a grid-search and maximizing the *Macro-Recall* within a nested 5-fold cross-validation (5-CV) procedure. The rationale behind the optimization of the *Macro-recall* in the validation set is justified by the fact of achieving an objective that is more clinical relevant for a screening purpose. Thus, the authors, following this rationale, preferred to minimize the false negatives and achieve a trade-off between sensitivity and specificity. This choice has been also performed according to the most recent state-of-the-art approaches in predictive medicine scenario [15], [19], [33]. The optimal hyperparameters (*hyp opt*), $\widetilde{x}$, and $\widetilde{y}$ were fed to the MTL model (i.e., CFG) for the training stage. The final prediction of the SS-MTL was computed by averaging the margin outputs of each single t task and then taking the decision based on the *sign* function:

$$\hat{y}_i = sign\left(\frac{\sum_{i=1}^{t} \widetilde{x}^T w_i + c_i}{t}\right) \tag{12}$$

The code[2] to replicate the SS-MTL approach is publicly released by the authors.

D. Experimental Comparisons

Our proposed SS-MTL approach was compared with baseline approaches (i.e., no-temporal, stacked-temporal) and with the MTL approach. Moreover, to better contextualize the proposed SS-MTL in the Semi-Supervised Learning (SSL) literature, the Self-Learning Algorithm (SLA) procedure was also compared with other existing SSL techniques, such as Positive and Unlabeled learning and Label Propagation. These baseline approaches adopted as ML models those employed in the state-of-the-art closer to our setting (see Sec. II), such as LR [11], [12], [13] with Lasso regularizer; DT [14]; RF [11], [13]; Boosting [11]; and SVM [14], [15] with Lasso regularizer. Experimental results were provided both for single (i.e., pathologies, drugs, exams, lab tests) and Overall/Overall* fields, by utilizing or not (i.e., noSLA) the SLA procedure. The same ML model adopted externally for the 10-CV was utilised also within the SLA procedure.

[2]https://github.com/michelebernardini/SS-MTL

TABLE IV: Range of hyperparameters (hyp) for each model: Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), Support Vector Machine (SVM) with Lasso regularizer, and Convex Fused Group Lasso (CFG) with LR model.

Model	**Hyp**	**Range**
LR [11], [12], [13]	Lambda	$\{10^{-5}, 10^{-4}10^{-3}, 10^{-2}, 0.1, 1\}$
DT [14]	max # of splits	$\{100, 200, 300, 400, 500\}$
RF [11], [13]	# of DT # of predictors	$\{25, 50, 75, 100, 125, 150\}$ $\{\frac{all}{4}, \frac{all}{3}, \frac{all}{2}, all\}$
Boosting [11]	max # of splits learning rate	$\{50, 100, 150, 200\}$ $\{10^{-2}, 0.1, 1\}$
SVM [14], [15]	Lambda	$\{10^{-5}, 10^{-4}10^{-3}, 10^{-2}, 0.1, 1\}$
CFG [8], [30]	ρ_1 ρ_2 ρ_3	$\{10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1\}$ $\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1\}$ $\{10^{-3}, 10^{-2}, 0.1\}$

1) Measures: The predictive performance was evaluated according to the following standard metrics for classification task: *Accuracy*, *Macro-precision*, *Macro-recall*, *Macro-F1* and *Area Under Receiver Operating Characteristic curve (AUC)*. From now on we refer to the *Macro-precision*, *Macro-recall* and *Macro-F1* as *Precision*, *Recall* and *F1* respectively. Table IV summarizes the range of the hyperparameters optimized in all the experiments.

V. Experimental Results

The experimental results of the SS-MTL approach are shown as predictive performance comparison with baseline approaches (i.e., no-temporal [Sec. V-B], stacked-temporal [Sec. V-C]) and with the MTL approach (Sec. V-D). For the baseline approach (i.e., no-temporal), the SLA procedure (i.e., the SSL technique from which our proposed approach is originated) is firstly compared with other SSL techniques, such as PU and LP (Sec. V-A).

In particular, Section V-D shows the trend of the predictive performance in relation to different portions of labeled training samples. This rationale is due to the intention of measuring the reliability of SS-MTL on dealing with a higher portion of unlabeled samples as excepted in a real-case scenario. Finally, the experimental results of the SS-MTL approach are shown in terms of pattern localization (Sec. V-E) to measure the importance of the predictors.

A. State-of-the-art comparison: Semi-Supervised Learning (SSL)

Table V shows the comparison of the experimental of the SLA procedure with other SSL techniques (i.e., PU, LP). The comparison was performed only for the Overall* field of the baseline (i.e., no-temporal) approach. The predictive performance of all ML models that used the SLA procedure

is clearly superior to the other SSL techniques (i.e., PU, LP), thus the SLA procedure was selected as the SSL paradigm for the proposed SS-MTL approach.

TABLE V: Experimental results comparison of the Self-Learning Algorithm (SLA) procedure with other Semi-Supervised Learning (SSL) techniques (i.e., Positive and Unlabeled learning [PU], Label Propagation [LP]). The comparison was performed only for the Overall* field of the baseline (i.e., no-temporal) approach. The best result in terms of *Recall* was highlighted in bold.

SLA	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.744	0.629	0.620	**0.660**	0.741
DT	0.792	0.677	0.670	**0.697**	0.693
RF	0.838	0.730	0.731	**0.734**	0.827
Boosting	0.849	0.687	0.760	**0.660**	0.847
SVM	0.716	0.627	0.623	**0.685**	0.749
LP	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.651	0.575	0.582	0.632	0.724
DT	0.698	0.583	0.592	0.618	0.616
RF	0.788	0.687	0.644	0.692	0.811
Boosting	0.813	0.646	0.707	0.640	0.829
SVM	0.598	0.559	0.576	0.655	0.710
PU	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.759	0.610	0.611	0.607	0.680
DT	0.795	0.553	0.594	0.538	0.532
RF	0.816	0.598	0.638	0.646	0.721
Boosting	0.811	0.516	0.662	0.508	0.692
SVM	0.705	0.601	0.609	0.640	0.729

B. State-of-the-art comparison: No-temporal

Table VI shows the comparison results for the no-temporal approach. Considering the SS-MTL an evolution of standard LR model, the comparison of the SS-MTL approach with the LR model would represent the most fair and straight comparison. The SS-MTL approach performance ($Recall = 0.737 \pm 0.054$) for Overall* field was greater than no-temporal (LR: $Recall = 0.660 \pm 0.048$) and stacked-temporal (LR: $Recall = 0.657 \pm 0.042$) in SLA configuration. Again for Overall field, the SS-MTL approach performance ($Recall = 0.668 \pm 0.053$) was greater than no-temporal (LR: $Recall = 0.616 \pm 0.062$) and stacked-temporal (LR: $Recall = 0.588 \pm 0.034$) in SLA configuration.

Nevertheless, if a global overview is considered, the best performance ($Recall = 0.734 \pm 0.051$) for no-temporal approach was obtained by the RF model for Overall* field in SLA configuration, but still lower than the best ones obtained by MTL ($Recall = 0.742 \pm 0.060$) approach and SS-MTL ($Recall = 0.737 \pm 0.054$) approach. Instead for Overall field, the best performance ($Recall = 0.665 \pm 0.062$) was obtained by the SVM in noSLA configuration. This result is comparable with the one extracted for the SS-MTL approach ($Recall = 0.668 \pm 0.053$).

C. State-of-the-art comparison: Stacked-temporal

Table VII shows the comparison results for the stacked-temporal approach. Focusing on the comparison of the LR model, the SS-MTL approach performance ($Recall = 0.737 \pm 0.054$) for Overall* field was greater than stacked-temporal (LR: $Recall = 0.657 \pm 0.042$) in SLA configuration. Again for Overall field, the SS-MTL approach performance ($Recall = 0.668 \pm 0.053$) was greater than stacked-temporal (LR: $Recall = 0.588 \pm 0.034$) in SLA configuration.

Nevertheless, if a global overview is considered, the best performance ($Recall = 0.709 \pm 0.057$) was obtained by the SVM model for Overall* field in SLA configuration. Accordingly for Overall field, the best performance ($Recall = 0.659 \pm 0.047$) was still obtained by the SVM model in SLA configuration. These results were lower than those extracted by the SS-MTL approach for Overall* ($Recall = 0.737 \pm 0.054$) and Overall field ($Recall = 0.668 \pm 0.053$).

D. Multitask-temporal comparison

Figure 4 compares the performance trend (i.e., *Recall*) over the fraction of labeled training samples $x, y \in Z_l$ for MTL and SS-MTL approaches considering both Overall and Overall* fields. Starting from a total of 4996 unlabeled samples $x' \in V_u$, figure 5 shows the trend of the pseudolabels samples $\widetilde{x}', \widetilde{y}' \in \widetilde{Z}_u$ selected by the SS-MTL approach (after random downsampling) over the fraction of labeled training samples. Table VIII shows more in detail the predictive performance for MTL and SS-MTL approaches. In particular, two configurations were highlighted, where both the full amount (f=100%) and a specific portion (f=30%) of labeled samples was utilised in the training stage. Comparable performance were obtained by the MTL ($Recall = 0.742 \pm 0.060$) approach and SS-MTL ($Recall = 0.737 \pm 0.054$) for Overall* field with f=100%. On the contrary, if f=30% the best performance ($Recall = 0.731 \pm 0.049$) was obtained by the SS-MTL approach with an important gain of 4.1% with respect to MTL ($Recall = 0.692 \pm 0.035$). The rationale to emphasize this result was due to the fact that the performance of SS-MTL remained stable from f=100% to f=30% while the performance of MTL decreased (see Figure 4).

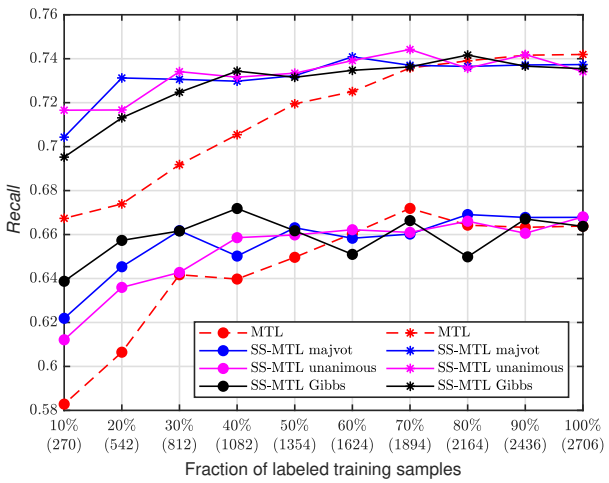

Fig. 4: MTL and SS-MTL approaches: Recall trend over the fraction of labeled training samples $x, y \in Z_l$. In the legend, stars indicate that gender and age were included as predictors (Overall*), filled circles were not (Overall).

TABLE VI: **No-temporal:** Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), and Support Vector Machine (SVM) with Lasso regularizer. In the SLA procedure the same classifier adopted externally in 10-CV was used. Overall* indicates that also gender and age were included as predictors. Best result in terms of *Recall* was highlighted in bold for each field.

	noSLA					SLA				
Pathologies	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.519	0.456	0.529	0.556	0.557	0.528	0.452	0.516	0.530	0.568
DT	0.614	0.499	0.531	0.553	0.576	0.652	0.511	0.531	0.556	0.549
RF	0.628	0.506	0.530	0.554	0.608	0.642	0.517	0.537	0.562	0.579
Boosting	0.652	0.523	0.544	**0.572**	0.585	0.651	0.518	0.539	0.563	0.580
SVM	0.488	0.437	0.524	0.546	0.563	0.501	0.445	0.532	0.559	0.568
Drugs	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.638	0.557	0.573	0.618	0.643	0.631	0.552	0.570	0.612	0.650
DT	0.694	0.540	0.540	0.549	0.559	0.628	0.543	0.561	0.598	0.602
RF	0.759	0.561	0.568	0.559	0.618	0.724	0.542	0.542	0.543	0.538
Boosting	0.781	0.557	0.580	0.554	0.608	0.767	0.538	0.552	0.537	0.596
SVM	0.594	0.536	0.574	0.625	0.645	0.597	0.541	0.579	**0.633**	0.659
Exams	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.607	0.534	0.559	0.594	0.647	0.610	0.537	0.562	0.600	0.643
DT	0.707	0.552	0.551	0.559	0.557	0.670	0.550	0.553	0.574	0.604
RF	0.778	0.548	0.576	0.546	0.663	0.774	0.543	0.569	0.541	0.602
Boosting	0.798	0.526	0.600	0.531	0.662	0.797	0.528	0.593	0.533	0.639
SVM	0.593	0.536	0.571	0.617	0.667	0.606	0.547	0.579	**0.629**	0.670
Lab tests	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.645	0.559	0.572	0.611	0.661	0.656	0.559	0.566	0.599	0.656
DT	0.743	0.574	0.574	0.576	0.576	0.710	0.574	0.572	0.588	0.588
RF	0.806	0.630	0.661	**0.619**	0.761	0.789	0.605	0.622	0.598	0.731
Boosting	0.815	0.498	0.565	0.522	0.759	0.815	0.514	0.649	0.529	0.743
SVM	0.633	0.550	0.565	0.602	0.657	0.668	0.567	0.571	0.603	0.653
Overall	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.703	0.582	0.579	0.607	0.676	0.706	0.587	0.584	0.616	0.683
DT	0.741	0.576	0.574	0.581	0.569	0.711	0.597	0.593	0.629	0.598
RF	0.803	0.640	0.654	0.632	0.762	0.777	0.615	0.620	0.612	0.695
Boosting	0.821	0.532	0.673	0.541	0.770	0.816	0.542	0.635	0.546	0.783
SVM	0.651	0.583	0.601	**0.665**	0.709	0.677	0.592	0.599	0.654	0.706
Overall*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.739	0.622	0.615	0.650	0.727	0.744	0.629	0.620	0.660	0.741
DT	0.796	0.658	0.660	0.662	0.666	0.792	0.677	0.670	0.697	0.693
RF	0.830	0.717	0.716	0.722	0.854	0.838	0.730	0.731	**0.734**	0.827
Boosting	0.847	0.678	0.761	0.650	0.875	0.849	0.687	0.760	0.660	0.847
SVM	0.693	0.613	0.617	0.683	0.747	0.716	0.627	0.623	0.685	0.749

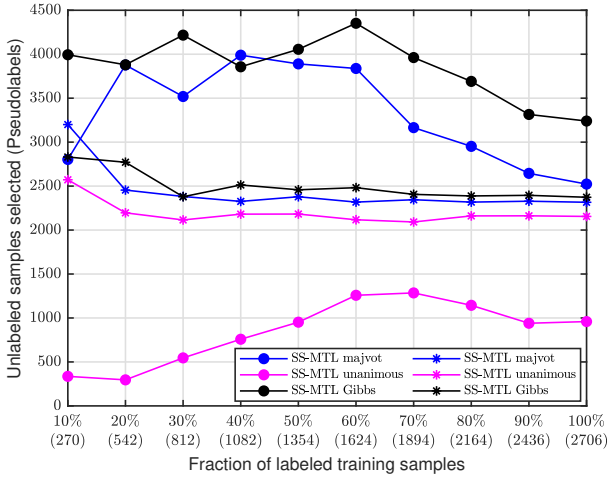

Fig. 5: Pseudolabel samples $\tilde{x}^{l}, \tilde{y}^{l} \in \widetilde{Z_u}$ selected from SLA procedure (after random downsampling) over fraction of labeled training samples $x \in Z_l$. In the legend, stars indicate that gender and age were included as predictors (Overall*), filled circles were not (Overall).

E. Pattern localization

Table IX explains which predictors in SS-MTL majvot (f=30%) configuration were more decisive to predict the next 1-year eGFR state. The final percentage weight of each predictor showed in Table IX was calculated by averaging the weights of the model over 10 folds, and then, over the t tasks.

VI. Discussions

This work has mainly contributed to the biomedical informatics field for the following points:

- Introduction of the SS-MTL paradigm for predicting short-term KD evolution. The proposed high-interpretable approach seeks to learn from labeled and unlabeled samples while imposing a temporal relatedness between consecutive tasks (i.e., time windows).
- Measurement and demonstration of the effectiveness of the SS-MTL approach with respect to the state-of-the-art in real-use case scenario (i.e., GP EHR dataset). The benefits in terms of predictive performance are particularly pronounced the more numerous the unlabeled samples are than those labeled. This condition reflects the real clinical use case where the observations of each patient lack annotation or are only partially labeled.

The impact of the predictive performance and pattern localization experimental results will be thoroughly discussed in

TABLE VII: **Stacked-temporal:** Logistic Regression (LR) with Lasso regularizer, Decision Tree (DT), Random Forest (RF), Gradient Boosting Trees (Boosting), and Support Vector Machine (SVM) with Lasso regularizer. In the SLA procedure the same classifier adopted externally in 10-CV was used. Overall* indicates that also gender and age were included as predictors. Best result in terms of *Recall* was highlighted in bold for each field.

	noSLA					**SLA**				
Pathologies	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.658	0.506	0.521	0.537	0.561	0.682	0.534	0.543	0.566	0.562
DT	0.607	0.500	0.533	0.563	0.578	0.703	0.539	0.546	**0.568**	0.563
RF	0.550	0.468	0.524	0.549	0.571	0.545	0.467	0.529	0.558	0.554
Boosting	0.680	0.526	0.536	0.557	0.561	0.685	0.512	0.522	0.533	0.541
SVM	0.646	0.511	0.529	0.550	0.570	0.726	0.539	0.544	0.555	0.557
Drugs	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.679	0.533	0.534	0.544	0.623	0.653	0.553	0.561	0.591	0.641
DT	0.693	0.540	0.540	0.550	0.548	0.613	0.532	0.554	0.588	0.567
RF	0.750	0.549	0.554	0.547	0.610	0.743	0.561	0.562	0.562	0.529
Boosting	0.758	0.555	0.565	0.554	0.620	0.701	0.538	0.537	0.545	0.617
SVM	0.587	0.535	0.579	**0.633**	0.665	0.582	0.530	0.574	0.624	0.666
Exams	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.665	0.540	0.543	0.560	0.603	0.648	0.539	0.546	0.567	0.616
DT	0.686	0.528	0.528	0.535	0.530	0.648	0.530	0.537	0.554	0.544
RF	0.764	0.551	0.565	0.548	0.611	0.748	0.531	0.540	0.531	0.561
Boosting	0.791	0.480	0.515	0.504	0.629	0.787	0.487	0.512	0.507	0.611
SVM	0.624	0.547	0.566	**0.605**	0.648	0.615	0.540	0.562	0.599	0.646
Lab tests	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.651	0.552	0.560	0.590	0.645	0.657	0.555	0.562	0.591	0.642
DT	0.723	0.554	0.553	0.556	0.554	0.675	0.541	0.542	0.557	0.537
RF	0.785	0.573	0.602	0.566	0.711	0.777	0.578	0.593	0.573	0.670
Boosting	0.818	0.496	0.699	0.521	0.731	0.811	0.490	0.585	0.516	0.713
SVM	0.611	0.545	0.572	0.617	0.663	0.638	0.563	0.580	**0.628**	0.671
Overall	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.727	0.571	0.567	0.579	0.660	0.721	0.576	0.572	0.588	0.669
DT	0.727	0.570	0.567	0.577	0.567	0.690	0.573	0.573	0.601	0.586
RF	0.795	0.625	0.638	0.617	0.745	0.775	0.602	0.611	0.598	0.660
Boosting	0.817	0.509	0.646	0.526	0.751	0.816	0.503	0.633	0.522	0.738
SVM	0.661	0.583	0.595	0.652	0.713	0.668	0.590	0.600	**0.659**	0.713
Overall*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
LR	0.755	0.611	0.606	0.621	0.706	0.762	0.638	0.630	0.657	0.742
DT	0.772	0.638	0.633	0.648	0.622	0.769	0.648	0.639	0.668	0.619
RF	0.816	0.696	0.691	0.704	0.834	0.805	0.678	0.675	0.684	0.777
Boosting	0.840	0.632	0.745	0.610	0.857	0.835	0.622	0.733	0.602	0.835
SVM	0.745	0.651	0.641	0.700	0.782	0.751	0.659	0.647	**0.709**	0.787

TABLE VIII: **Multitask-temporal:** Decision Tree (DT) classifier was used to select pseudo-labels in SLA procedure, except for Overall* where Support Vector Machine (SVM) with Lasso regularizer was used. Overall* indicates that also gender and age were included as predictors. Fraction (f) represents the number of labeled samples used in the training stage. The table depicts the SS-MTL majvot configuration. Best result in terms of *Recall* was highlighted in bold for each field.

f = 100%	**MTL**					**SS-MTL**				
	Accuracy	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Pathologies	0.766	0.510	0.542	0.510	0.517	0.782	0.523	0.541	**0.522**	0.511
Drugs	0.568	0.526	0.584	**0.640**	0.680	0.581	0.533	0.582	0.638	0.690
Exams	0.580	0.532	0.579	**0.631**	0.677	0.573	0.522	0.568	0.611	0.673
Lab tests	0.622	0.561	0.587	**0.643**	0.687	0.621	0.560	0.587	0.642	0.687
Overall	0.625	0.568	0.598	0.664	0.720	0.636	0.575	0.601	**0.668**	0.713
Overall*	0.765	0.681	0.668	**0.742**	0.816	0.750	0.670	0.661	0.737	0.820
f = 30%	**MTL**					**SS-MTL**				
	Accuracy	*F1*	*Precision*	*Recall*	*AUC*	*Accuracy*	*F1*	*Precision*	*Recall*	*AUC*
Pathologies	0.730	0.512	0.516	0.515	0.542	0.600	0.457	0.542	**0.549**	0.564
Drugs	0.584	0.531	0.575	0.626	0.671	0.640	0.566	0.584	**0.635**	0.680
Exams	0.625	0.550	0.569	**0.610**	0.651	0.626	0.539	0.556	0.587	0.630
Lab tests	0.662	0.570	0.576	0.616	0.656	0.660	0.575	0.584	**0.629**	0.658
Overall	0.682	0.590	0.593	0.642	0.686	0.707	0.612	0.610	**0.662**	0.700
Overall*	0.758	0.655	0.644	0.692	0.784	0.746	0.665	0.657	**0.731**	0.811

Section VI-A and Section VI-B. Then, limitations and future work will be argued in Section VI-C and Section VI-D.

A. Predictive performance

In the following section the two RQs formulated in Sec.I will be discussed.

TABLE IX: Top-10 predictors for SS-MTL majvot approach with f =30 %. Overall* indicates that also gender and age were included as predictors. D=Drugs; E=Exam; M=Monitoring.

	Overall			Overall*		
Rank	Field	Predictors	W [%]	Field	Predictors	W [%]
1)	D	Valsartan and diuretics	3.78	M	Age	44.85
2)	D	Colecalciferol (vitamin D3)	3.59	D	Furosemide	3.26
3)	D	Levothyroxine	3.17	D	Metformin	2.42
4)	D	Alfuzosin	3.16	D	Amlodipine	1.40
5)	D	Lansoprazole	3.08	D	Ramipril and amlodipine	1.38
6)	D	Furosemide	2.89	D	Valsartan and diuretics	1.32
7)	D	Acetylsalicylic acid	2.78	D	Pravastatin	1.28
8)	D	Pantoprazole	2.67	D	Atorvastatin	1.27
9)	E	Interview and evaluation	2.51	D	Bisoprolol	1.19
10)	D	Nebivolol	2.28	D	Omeprazole	1.16
		Others	70.09		*Others*	40.47

1) ***RQ1:*** *Is the MTL approach capable to capture the eGFR temporal evolution?:* The MTL approach as showed in Table VIII was capable to capture the eGFR temporal evolution, because for Overall* configuration in terms of $Recall = 0.742 \pm 0.060$ was superior than the best competitors for no-temporal (Table VI) and stacked-temporal (Table VII) approaches (RF: $Recall = 0.722 \pm 0.036$; RF: $Recall = 0.704 \pm 0.067$, respectively). Instead, if age and gender were not considered (i.e., Overall), the MTL performance (*Recall* = 0.664 ± 0.048) was close to SVM (*Recall* = 0.665 ± 0.062) for no-temporal approach but superior than SVM (*Recall* = 0.652 ± 0.053) for stacked-temporal approach. However, the performance of the MTL approach remained greater than the baseline LR model for both Overall* and Overall configurations. These outcomes highlighted the importance to include the temporal evolution of the predictors in the ML model. Moreover, experimental results suggested how demographic information was highly discriminative in terms of predictive performance.

The single fields of the MTL approach that mostly affected the predictive performance were drugs and lab tests, while exams seemed to impact less. For instance, the single lab tests field in MTL reached a *Recall* until 0.643 ± 0.039, much more superior than the other competitors. On the contrary, the pathologies field obtained very poor results and for this reason, was excluded from the Overall and Overall* fields. Results evidenced how the predictive performance of MTL and no-temporal approaches were globally superior to one of the stacked-temporal approaches, which encapsulated the temporal information by aggregating longitudinally the time windows, and this aspect may suffer much the high temporal data sparsity. However, the MTL approach was capable of modeling and interpreting through the regularization strategy the progression of the temporal information, otherwise lost in the no-temporal approach.

2) ***RQ2:*** *Is the SS-MTL approach capable to capture useful information from unlabeled patients?:* The SS-MTL approach was mostly capable to gain useful information from unlabeled patients, in terms of predictive performance concerning to MTL, when labeled patients were less numerous than those unlabeled. This situation commonly reflects the real-case general practice scenario, where available labeled samples size is limited, while unlabeled samples are much more abundant.

Specifically (see Figure 4), the SS-MTL approach did not add an important gain compared to MTL in predictive performance both for Overall and Overall* fields when the full fraction (f=100%) of labeled training sample size was considered. But, if f was progressively decreased (i.e., both for MTL and SS-MTL), the predictive performance kept on being still similar until f=70% for Overall* and until f=60% for Overall. After these cut points, the more f decreases, the more the spread between SS-MTL and MTL increased due to an MTL predictive performance worsening. This finding suggested that our proposed SS-MTL approach was convenient since at least unlabeled samples (# 4996) were almost 2.5 times more numerous than labeled samples (# 1894 at f=70%). Additionally, the SS-MTL predictive performance until f=30% remained almost constant if compared to f=100%, while the MTL performance decreased much earlier as seen before. This further finding proved how the SS-MTL approach was reliable in dealing with unlabeled information.

Basically, for the Overall* field, the *Recall* trend across SLA majvot, unanimous, and Gibbs seemed to be more stable. On the contrary, for the Overall field, the *Recall* trend was more fluctuating and it appeared that SLA unanimous was less performing than the others. These considerations may be fully explained in Figure 5, from which it has emerged that the number of pseudo-labels selected by SLA directly interfered with the SS-MTL working stability. Indeed, the most stable performance of the SS-MTL for the Overall* field was influenced by almost constant pseudo-labels selected by SLA. However, even if for the Overall field at f=30% more pseudo-labels were selected by SS-MTL Gibbs and SS-MTL majvot than Overall*, the gap between the predictive performance remained fairly constant across different f thresholds (see figure 4). These findings suggested how an increase of almost $2K$ pseudo-labels between Overall and Overall* fields was not related to an increase in predictive performance. Indeed, the pseudo-labels may not be necessarily informative enough to improve the generalization performance of the ML model.

We demonstrated that all the models used for SSL techniques obtained the best predictive performance with the SLA procedure. A central hypothesis in SSL, based on which discriminant models are developed, is the *low density separation* assumption (**H**) [34] which stipulates that the decision boundary should pass through low-density regions. In this sense, contrary to PU [22], the negative class has a central role in finding the decision boundary. In this sense, SLA follows assumption **H** which is also shown to be effective in our experiments. Instead, graphical models, as LP [23], are based on manifold assumption and construct a graph where the nodes represent training examples and the edges reflect similarities between them. The class label of each labeled node is then propagated to its neighbors using label spreading techniques. The similarity between the two observations is based on their Euclidean distance in the feature space, and due to the curse of dimensionality when the dimension of the space is high - as in our case - the Euclidean distance does not reflect well the proximity between examples.

B. Clinical significance

The proposed SS-MTL approach was high-interpretable and this aspect assumes an important relevance in the general

practice scenario. In fact, obtaining only satisfactory predictive results might be useless if then the results cannot be interpreted by GPs, which need to understand and explain which factors have mostly determined a prediction. From experimental results, it has turned out how gender and age may play a key role compared to other predictors for forecasting the next 1-year eGFR state. In fact, the predictive performance of the SS-MTL approach for the Overall* field (f=30%) was much greater than the one for the Overall field (see Table VIII). This finding was fully clarified in Table IX, from which it emerged that age was the leading predictor with importance of 44.85%, while gender did not appear as a discriminant factor. Although age has already been adequately demonstrated to be one of the major factors in kidney functionality, a prediction merely based on age provided inferior predictive performance, as proven also in [13]. The remaining predictors belonged to the drugs field and this aspect suggested how highly discriminative the past patient's pharmacological treatment might be. In particular, the best contenders such as furosemide and metformin are strictly correlated to variations of eGFR value. Furosemide treatment reduces kidney functionalities for patients with cardiovascular pathologies [35], [36], while metformin administration in patients suffering from moderate CKD is associated with clinical outcome improvements [37]. The creatinin, even if it has been used in Eq. 1 for the calculation of the CKD-EPI formula, did not appear as one of the best top-10 predictors. Since the demographic predictors (i.e., gender and age) are included in the eGFR formula (see Eq. 1) the performance of the predictive model improved in the Overall* experiment. On the other hand, if demographic information (i.e., gender and age) were not considered (i.e., Overall and single modality experiments) to discover further discriminative predictors besides demographic information, there was no predominant predictor over others, but the pharmacological pathway remained still decisive with respect to the other fields.

C. Limitations

In this work, the Overall/Overall* fields did not account for the pathologies' information, which caused a predictive performance worsening. In fact, the pathologies field contains much more static information than the others, and it may have found difficult to offer discriminative temporal information to the predictive model. Perhaps, the exclusion of pathologies among the predictors may limit the global contextualization of the clinical problem. To better combine and make coexist heterogeneous feature sets consisting of various EHR fields (e.g., pathologies, exams, drugs, lab tests) of different data types (e.g., categorical, continuous), multi-view learning approaches [38] may be explored as an intriguing future direction.

We used linear models, which assume linear relationships between the variables, and the outcome and we did not take into account the non-linear combination of different predictors that could potentially affect the outcome. In this context, we may explore non-linear models with different features map in order to discover new hidden high-discriminative temporal patterns.

D. Future work

Future work may be addressed to explore interesting directions by including different experimental procedures, task definitions, and data processing.

It would be interesting to apply the SS-MTL approach considering only patients enclosed within a specific range of CKD stages and/or, unlike our strategy, predict CKD stages I and II from the others. Alternatively, binary classification could be applied to the prediction of the variation in time of the eGFR value above a certain experimental threshold [9]. Other very promising and attractive solutions could be to extend the current SS-MTL binary classification problem to a multiclass classification problem [39] or to learning to rank approach (i.e., learning the risk prediction using an ordinal structure of all CKD stages).

For what concernc the data processing, the strong class imbalance may be addressed using more advanced data imputation strategies rather than SMOTE, median/mean imputation [40] and KNN [41]. For instance, the missing values of the EHR field may be imputed by using conditional GAN [42] across different temporal windows and different spatial views (i.e., EHR fields).

VII. Conclusions

We proposed the SS-MTL approach for predicting short-term KD evolution on multiple GPs' EHR data. We demonstrated that the SS-MTL approach was capable to predict and discriminate the CKD stage I (normal samples) from the other more severe CKD stages (risky samples), by modeling the temporal evolution of EHR data (e.g., imposing temporal relatedness between consecutive time windows). The SS-MTL approach was mostly capable to gain useful information from unlabeled patients when labeled patients are less numerous than those unlabeled. This situation reflects commonly the real-case general practice scenario, where available labeled samples are limited, but those unlabeled are much more abundant. The SS-MTL approach, exhibiting also a high level of interpretability (i.e., age and pharmacological pathway were the most important predictors), might be the ideal candidate in general practice to get integrated within a decision support system for CKD screening purposes.

Acknowledgment

The authors would thank Federazione Italiana Medici di Medicina Generale (FIMMG), NetMedica Italia (NMI), and Alessandro Ferri for their support in data collection and organization; Arianna Vecchi and Anastasiya Tymoshenko for their effort in proofreading the English language of the manuscript; and Vasilii Feofanov for his contribution in SLA code migration.

References

[1] V. A. Luyckx, M. Tonelli, and J. W. Stanifer, "The global burden of kidney disease and the sustainable development goals," *Bulletin of the World Health Organization*, vol. 96, no. 6, p. 414, 2018.

[2] A. Levin, M. Tonelli, J. Bonventre, J. Coresh, J.-A. Donner, A. B. Fogo, C. S. Fox, R. T. Gansevoort, H. J. Heerspink, M. Jardine *et al.*, "Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy," *The Lancet*, vol. 390, no. 10105, pp. 1888–1917, 2017.

[3] N. J. Kassebaum, M. Arora, R. M. Barber, Z. A. Bhutta, J. Brown, A. Carter, D. C. Casey, F. J. Charlson, M. M. Coates, M. Coggeshall *et al.*, "Global, regional, and national disability-adjusted life-years (dalys) for 315 diseases and injuries and healthy life expectancy (hale), 1990–2015: a systematic analysis for the global burden of disease study 2015," *The Lancet*, vol. 388, no. 10053, pp. 1603–1658, 2016.

[4] World Health Organization *et al.*, "Tackling NCDs:'best buys' and other recommended interventions for the prevention and control of noncommunicable diseases," World Health Organization, Tech. Rep., 2017.

[5] A. S. Levey, J. Coresh, H. Tighiouart, T. Greene, and L. A. Inker, "Measured and estimated glomerular filtration rate: current status and future directions," *Nature Reviews Nephrology*, pp. 1–14, 2019.

[6] P. E. Stevens and A. Levin, "Evaluation and management of chronic kidney disease: synopsis of the kidney disease: improving global outcomes 2012 clinical practice guideline," *Annals of Internal Medicine*, vol. 158, no. 11, pp. 825–830, 2013.

[7] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, and Z. He, "Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review," *Journal of the American Medical Informatics Association*, 2020.

[8] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2011, pp. 814–822.

[9] A. Singh, G. Nadkarni, O. Gottesman, S. B. Ellis, E. P. Bottinger, and J. V. Guttag, "Incorporating temporal EHR data in predictive models for risk stratification of renal function deterioration," *Journal of Biomedical Informatics*, vol. 53, pp. 220 – 228, 2015.

[10] S. Bailly, G. Meyfroidt, and J.-F. Timsit, "Whats new in ICU in 2050: big data and machine learning," *Intensive Care Medicine*, vol. 44, no. 9, pp. 1524–1527, 2018.

[11] J. Zhao, Q. Feng, P. Wu, R. A. Lupu, R. A. Wilke, Q. S. Wells, J. C. Denny, and W.-Q. Wei, "Learning from longitudinal data in electronic health record and genetic data to improve cardiovascular event prediction," *Scientific Reports*, vol. 9, no. 1, pp. 1–10, 2019.

[12] P. Fraccaro, S. van der Veer, B. Brown, M. Prosperi, D. ODonoghue, G. S. Collins, I. Buchan, and N. Peek, "An external validation of models to predict the onset of chronic kidney disease using population-based electronic health records from Salford, UK," *BMC Medicine*, vol. 14, no. 1, p. 104, 2016.

[13] S. Ravizza, T. Huschto, A. Adamov, L. Böhm, A. Büsser, F. F. Flöther, R. Hinzmann, H. König, S. M. McAhren, D. H. Robertson *et al.*, "Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data," *Nature Medicine*, vol. 25, no. 1, pp. 57–59, 2019.

[14] N. A. Almansour, H. F. Syed, N. R. Khayat, R. K. Altheeb, R. E. Juri, J. Alhiyafi, S. Alrashed, and S. O. Olatunji, "Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study," *Computers in Biology and Medicine*, vol. 109, pp. 101–111, 2019.

[15] M. Bernardini, L. Romeo, P. Misericordia, and E. Frontoni, "Discovering the type 2 diabetes in electronic health records using the sparse balanced support vector machine," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 1, pp. 235–246, 2020.

[16] D. Y. Ding, C. Simpson, S. Pfohl, D. C. Kale, K. Jung, and N. H. Shah, "The effectiveness of Multitask Learning for Phenotyping with Electronic Health Records Data." in *Pacific Symposium on Biocomputing*. World Scientific, 2019, pp. 18–29.

[17] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[18] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan, "Multitask learning and benchmarking with clinical time series data," *Scientific Data*, vol. 6, no. 1, pp. 1–18, 2019.

[19] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and L. Burattini, "Early temporal prediction of type 2 diabetes risk condition from a general practitioner electronic health record: A multiple instance boosting approach," *Artificial Intelligence in Medicine*, p. 101847, 2020.

[20] Z. Che, Y. Cheng, S. Zhai, Z. Sun, and Y. Liu, "Boosting deep learning risk prediction with generative adversarial networks for electronic health records," in *IEEE International Conference on Data Mining*, 2017, pp. 787–792.

[21] C. Esteban, S. L. Hyland, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional gans," *arXiv preprint arXiv:1706.02633*, 2017.

[22] F. Mordelet and J.-P. Vert, "ProDiGe: Prioritization of Disease Genes with multitask machine learning from positive and unlabeled examples," *BMC Bioinformatics*, vol. 12, no. 1, p. 389, 2011.

[23] P. Zhang, F. Wang, J. Hu, and R. Sorrentino, "Towards personalized medicine: leveraging patient similarity and drug similarity analytics," *Summits on Translational Science Proceedings*, vol. 2014, p. 132, 2014.

[24] M. R. Amini, N. Usunier, and F. Laviolette, "A transductive bound for the voted classifier with an application to semi-supervised learning," in *Advances in Neural Information Processing Systems*, 2009, pp. 65–72.

[25] E. Frontoni, A. Mancini, M. Baldi, M. Paolanti, S. Moccia, P. Zingaretti, V. Landro, and P. Misericordia, "Sharing health data among general practitioners: The nu. sa. project," *International journal of medical informatics*, vol. 129, pp. 267–274, 2019.

[26] E. Frontoni, L. Romeo, M. Bernardini, S. Moccia, L. Migliorelli, M. Paolanti, A. Ferri, P. Misericordia, A. Mancini, and P. Zingaretti, "A decision support system for diabetes chronic care models based on general practitioner engagement and ehr data sharing," *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 8, pp. 1–12, 2020.

[27] A. S. Levey, L. A. Stevens, C. H. Schmid, Y. L. Zhang, A. F. Castro, H. I. Feldman, J. W. Kusek, P. Eggers, F. Van Lente, T. Greene *et al.*, "A new equation to estimate glomerular filtration rate," *Annals of Internal Medicine*, vol. 150, no. 9, pp. 604–612, 2009.

[28] L. A. Inker, C. H. Schmid, H. Tighiouart, J. H. Eckfeldt, H. I. Feldman, T. Greene, J. W. Kusek, J. Manzi, F. Van Lente, Y. L. Zhang *et al.*, "Estimating glomerular filtration rate from serum creatinine and cystatin C," *New England Journal of Medicine*, vol. 367, no. 1, pp. 20–29, 2012.

[29] A. K. Bello, M. Alrukhaimi, G. E. Ashuntantang, S. Basnet, R. C. Rotter, W. G. Douthat, R. Kazancioglu, A. Köttgen, M. Nangaku, N. R. Powe *et al.*, "Complications of chronic kidney disease: current state, knowledge gaps, and strategy for action," *Kidney International Supplements*, vol. 7, no. 2, pp. 122–129, 2017.

[30] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2012, pp. 1095–1103.

[31] M.-R. Amini and N. Usunier, *Learning with Partially Labeled and Interdependent Data*. Springer, 2015.

[32] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[33] L. Romeo, G. Armentano, A. Nicolucci, M. Vespasiani, G. Vespasiani, and E. Frontoni, "A novel spatio-temporal multi-task approach for the prediction of diabetes-related complication: a cardiopathy case of study," in *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, 7 2020, pp. 4299–4305.

[34] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-supervised learning," *IEEE Transactions on Neural Networks*, vol. 20, no. 3, pp. 542–542, 2009.

[35] D. Singh, K. Shrestha, J. M. Testani, F. H. Verbrugge, M. Dupont, W. Mullens, and W. W. Tang, "Insufficient natriuretic response to continuous intravenous furosemide is associated with poor long-term outcomes in acute decompensated heart failure," *Journal of Cardiac Failure*, vol. 20, no. 6, pp. 392–399, 2014.

[36] Y. Okuhara, S. Hirotani, Y. Naito, A. Nakabo, T. Iwasaku, A. Eguchi, D. Morisawa, T. Ando, H. Sawada, E. Manabe *et al.*, "Intravenous salt supplementation with low-dose furosemide for treatment of acute decompensated heart failure," *Journal of Cardiac Failure*, vol. 20, no. 5, pp. 295–301, 2014.

[37] M. J. Crowley, C. J. Diamantidis, J. R. McDuffie, C. B. Cameron, J. W. Stanifer, C. K. Mock, X. Wang, S. Tang, A. Nagi, A. S. Kosinski *et al.*, "Clinical outcomes of metformin use in populations with chronic kidney disease, congestive heart failure, or chronic liver disease: a systematic review," *Annals of Internal Medicine*, vol. 166, no. 3, pp. 191–200, 2017.

[38] M. Kandemir, A. Vetek, M. Goenen, A. Klami, and S. Kaski, "Multi-task and multi-view learning of user state," *Neurocomputing*, vol. 139, pp. 97–106, 2014.

[39] V. Feofanov, E. Devijver, and M.-R. Amini, "Transductive bounds for the multi-class majority vote classifier," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3566–3573.

[40] M. Bernardini, M. Morettini, L. Romeo, E. Frontoni, and L. Burattini, "TyG-er: An ensemble Regression Forest approach for identification of clinical factors related to insulin resistance condition using Electronic

Health Records," *Computers in Biology and Medicine*, vol. 112, p. 103358, 2019.

[41] C. Ye, T. Fu, S. Hao, Y. Zhang, O. Wang, B. Jin, M. Xia, M. Liu, X. Zhou, Q. Wu *et al.*, "Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning," *Journal of Medical Internet Research*, vol. 20, no. 1, p. e22, 2018.

[42] A. Doinychko and M.-R. Amini, "Biconditional generative adversarial networks for multiview learning with missing views," in *European Conference on Information Retrieval*. Springer, 2020, pp. 807–820.