ABSTRACT

Clinical Decision Support (CDS) tools are typically designed to assist physicians in clinical decision making at Point Of Care (POC). Existing CDS tools commonly rely on relatively simple rules, deduced from relevant clinical guidelines. However, the increasing pace by which Health Care Organizations (HCOs) adopt Electronic Health Record technologies suggests great potential for CDS tools that directly mine the massive clinical data collected at the HCO. A natural goal for such tools is to exploit Machine Learning (ML) algorithms in order to predict patient’s outcome. However, the technical challenges involved in constructing such a system in practice are quite involved, where in particular treatments outcome are often not available as part of the HCO’s data.

Here, we propose a different strategy in which we use the assigned treatments as the labels in the learning process of the supervised ML algorithms. We present two different use-cases in which our approach could be used. First, in order to highlight the clinical features most associated with the assigned treatments; and second, in order to predict the customary treatment for a patient at POC in a statistically data-driven manner. Altogether, our approach represents a novel strategy that is complementary to the classical paradigm of rule-based clinical guidelines adherence. Experimental results over hypertension clinical data demonstrate the validity of our approach.

Categories and Subject Descriptors

I.2.1 [Artificial Intelligence]: Applications and Expert Systems—Medicine and science

General Terms

Algorithms

Keywords

Clinical Decision Support, Machine Learning, Feature Selection, Naive Bayes, Hypertension

1. INTRODUCTION

Clinical Decision Support (CDS) tools are typically designed to assist physicians in clinical decision making at Point Of Care (POC). Such tools are assumed to hold great potential for improving clinical care and lowering the associated costs [10, 14]. In a recent comprehensive study, Wright et al proposed a taxonomy of currently available CDS front-end tools that includes 53 different types of tools, divided into six categories: medication dosing support; order facilitators; POC alerts/reminders; Relevant information display; Workflow support; and Expert systems [24]. Evidently, existing CDS tools typically rely on relatively simple rules, deduced from relevant clinical guidelines. However, the increasing pace by which Health Care Organizations (HCOs) adopt Electronic Health Record technologies suggests great potential for CDS tools that directly mine the clinical data collected and stored at the HCO as part of its routine care delivery process. A natural goal for such tools would be to exploit Statistical Analysis and Machine Learning (ML) algorithms [6] in mining the HCO’s clinical data, in order to predict the outcome of clinical treatments considered for an individual patient. The underlying intuition is that such tools would be able to identify patients that are “similar” to the examined patient, and further consider the treatments these patients have received along with the associated outcomes; presumably, given these data, the system would be able to accurately predict the outcome of each of the treatments considered for the examined patient.
Obviously, such accurate predictions could be invaluable in improving patients’ care. However, the technical challenges involved in constructing such a system in practice are far from trivial. First, from a statistical perspective, the data collected at an HCO is reminiscent to an observational study and correspondingly is expected to be inherently biased [18]. In addition, the scheme above relies on the assumption that treatments outcomes are recorded and available as part of the HCO data for a relatively large and representative patients population. In practice, however, this assumption is often not valid for various reasons, e.g., in cases where the outcome could be determined only long after the treatment was given. Finally, merely to provide a proper definition of the concept of clinical outcome is often quite challenging, and typically there are different notions of outcome that should be considered, which further complicates the task of the ML algorithms. Thus, it is perhaps not surprising that although many studies report promising results in utilizing ML tools to predict clinical outcomes (e.g., [7, 17, 25]), there seem to be very little report on CDS tools being used in practice within HCO’s that directly analyze the HCO clinical data, aiming to predict patient’s outcome.

In the current work we outline a different approach to harness supervised ML algorithms in mining HCO’s clinical data, in which the ML analysis focuses at the assigned clinical treatment. Specifically, we suggest to consider the assigned treatments as the “class labels” considered by the ML algorithms incorporated within the CDS system. We note that much of the aforementioned difficulties related to considering clinical outcome seem less acute in this proposal. First, it seems reasonable that in many practical scenarios the information regarding which treatments were assigned will be more available in the HCO database, compared to the outcomes associated with these treatments. In addition, the assigned treatments are often selected from a relatively compact and well defined list, and thus are less ambiguous in nature.

We further outline two concrete use-cases in which the proposed approach could be exploited. The first use case is termed here Retrospective Treatment Analysis (RTA). In this use case, a collection of treatments – assigned in the context of a particular disease stage – is analyzed using a ranking algorithm that considers all clinical features recorded in the HCO database. Single features and/or pairs of features that are statistically identified as most informative on the selected treatment are then highlighted by the system, allowing the user to gain insights regarding the HCO treatment allocation process. These insights could then be compared to existing knowledge as reflected, e.g., by the relevant clinical guidelines. The second use case, termed here Probable Treatment Prediction (PTP), is designed to be used by the physician at POC while considering candidate treatments for an individual patient. Specifically, in this use case the system analyzes the data associated with the examined patient and aims to predict the most probable treatment for this patient, as implied by treatments previously assigned to “similar” patients at the HCO. The system further adds “justification” to the provided prediction, highlighting single clinical features and/or pairs of clinical features that contributed the most to this prediction. In addition, an automatically generated confidence score is added to each prediction, aiming to quantify its reliability. The physician could then consider the provided prediction and accompanied confidence and justification before taking her final decision regarding the assigned treatment. This final decision is further recorded by the system in order to refine future analysis and future predictions.

We have implemented and tested these two use cases in a generic CDS prototype. Our results over hypertension clinical data demonstrate the validity of our approach.

2. METHODS

2.1 Preliminaries

2.1.1 Notations

The proposed system is expected to be used in the context of a particular disease, denoted here disease d. We assume that the HCO database refers to $N_P$ patients that were diagnosed with disease d and treated accordingly, and that the assigned treatment is reported in the database for all these $N_P$ patients. Importantly, the outcomes associated with the assigned treatments are not required by the proposed analysis, and correspondingly are not assumed to be available in the database. We denote by $T$ a random variable with values $\{t_1, \ldots, t_{N_P}\}$, representing all distinct treatments that were assigned to the $N_P$ patients in the context of disease d.

We further assume that the HCO database maintains data about $N_F$ clinical features, denoted by the random variables, $\{F_1, \ldots, F_{N_F}\}$. The data mined by the system can thus be represented by a matrix $M$, where $M(i, j)$ indicates the value of the i-th patient according to the j-th feature, where $i = 1 : N_P$, $j = 1 : N_F$. The treatment variable can be represented via an additional column vector, where $T(i)$ indicates the treatment assigned to the i-th patient. Finally, to denote a patient which is currently examined at POC, we use the index $i^\ast$. The data associated with this patient could be represented via an additional row in $M$, while $T(i^\ast)$ is unknown.

2.1.2 Quantization of Continuous Features

Different clinical features may be described by different types of random variables – categorical, ordinal, or continuous. The following analysis holds for all these types. Nonetheless, for simplicity, in the current study we represent all clinical features as categorical random variables. Thus, clinical features that are continuous in nature – e.g., age, or blood pressure – were quantized and then treated as categorical variables. Quantizing continuous features could be quite involved and the choice of a particular quantization scheme should be guided by the context. Our current implementation supports the following three options:

- **Max-Info Quantization**: A continuous feature is quantized such that the obtained quantized version is most informative about the class label, $T$. This scheme is equivalent to the minimal entropy quantization proposed in [8].

- **Equally Populated Quantization**: A continuous feature is quantized into equally populated bins, as in [20], resulting with a quantization which is relatively robust to outliers.

\footnote{For brevity, here we only consider the most recent treatment assigned to an individual patient in the context of disease d; thus, each row in $M$ refers to a different patient.}
• Knowledge-Based Quantization: A continuous feature is quantized based on existing expert knowledge; for example, it is common to quantize systolic blood pressure into “Normal” (< 120), “Prehypertension” (120 – 139), “Stage 1 Hypertension” (140 – 159), and “Stage 2 Hypertension” (≥ 160) [3].

In the reported results we used Knowledge-Based Quantization when relevant knowledge was available, and Equally Populated Quantization (into 3 bins) for the remaining continuous features, mainly due to the simplicity and efficiency of this scheme, and since in pilot experiments we saw no improvement while using Max-Info Quantization.

2.1.3 Mutual Information and Synergy

Our analysis relies on information theoretic concepts that we now describe. Let \( n(F_1 = f_1, T = t_i) \) represent the number of patients at the HCO database for which the value of the clinical feature \( F_1 \) is \( f_1 \) and the reported assigned treatment is \( t_i \). Then, the empirical joint distribution of \( F_1 \) and \( T \) is given by

\[
P(F_1 = f_1, T = t_i) = n(F_1 = f_1, T = t_i) / N_P .
\]

The empirical Mutual Information (MI) between \( F_1 \) and \( T \) is then given by [4]

\[
I(F_1; T) = \sum_{f_1, t_i} P(f_1, t_i) \log \frac{P(f_1, t_i)}{P(f_1)P(t_i)} ,
\]

where \( P(f_1) = \sum_{t_i} P(f_1, t_i) \). Similarly, let \( n(F_1 = f_1, F_k = f_k, T = t_i) \) represent the number of patients at the HCO database for which \( F_1 = f_1 \), \( F_k = f_k \), and the reported assigned treatment is \( t_i \). Then, the empirical joint distribution of \( F_1, F_k \) and \( T \) is given by

\[
P(F_1 = f_1, F_k = f_k, T = t_i) = n(F_1 = f_1, F_k = f_k, T = t_i) / N_P .
\]

The empirical MI that \( F_1 \) and \( F_k \) jointly hold over \( T \) is then given by [4]

\[
I(F_1, F_k; T) = \sum_{f_1, f_k, t_i} P(f_1, f_k, t_i) \log \frac{P(f_1, f_k, t_i)}{P(f_1, f_k)P(t_i)} ,
\]

where \( P(f_1, f_k) = \sum_{t_i} P(f_1, f_k, t_i) \). Finally, the synergy of \( F_1 \) and \( F_k \) in terms of the information they convey over the assigned treatment, \( T \), can be defined as [19, 2]

\[
S(F_1, F_k; T) = I(F_1, F_k; T) - (I(F_1; T) + I(F_k; T)) .
\]

Thus, \( S(F_1, F_k; T) \) quantifies the amount of information gained by considering the relation of \( F_1 \) jointly with \( F_k \) over \( T \), as opposed to considering the relation of each of these two features over \( T \), independently.

2.2 Retrospective Treatment Analysis

The Retrospective Treatment Analysis use case aims to highlight clinical features that are most associated with the assigned treatment, \( T \). Considering all clinical features recorded at the database can be overwhelming, especially if one would like to further consider the association between multiple combinations of clinical features and \( T \). Thus, some ranking mechanism is required. In the following we present the relevant workflow for this use case and the ranking algorithm we exploit for this purpose.

2.2.1 Retrospective Treatment Analysis Workflow

The Retrospective Treatment Analysis is targeted for the HCO management and/or senior physicians, allowing the user to inspect off-line the HCO treatment allocation process.

Every fixed period of time, or upon demand, the system performs retrospective analysis of relevant patients’ data currently available in the HCO databases, to incorporate new patients or new information about existing patients that was recorded at the HCO since the last update. This analysis includes

• Pre-processing of the new data, (e.g. cleansing, quantization of continuous features).
• Mutual information and synergy estimation.
• Feature ranking.

At any stage, the Retrospective Treatment View can be invoked to present information about the patients currently processed in the system. This view supports exploration – including visual presentation and drill down capabilities – of all clinical features and pairs of clinical features, ranked according to their relation to the assigned treatments, as described next.

2.2.2 Univariate and Pairwise Feature Ranking by Mutual Information and Synergy

In general, the MI is the unique measure of relatedness between a pair of variables that obeys several simple and desirable requirements independent of assumptions about the form of the underlying joint distributions [4]. In particular, it is independent of invertible transformations on the individual variables. As a result, the units by which \( F_1 \) is reported at the database will have no impact on \( I(F_1; T) \), as long as it is being reported consistently. The absolute scale of the MI also has a clear meaning. For example, if there are only two candidate treatments, then \( I(F_1; T) = 1 \text{ bits} \) implies that \( F_1 \) solely determines the assigned treatment. In addition, \( I(F_1; T) \) will reflect any type of dependence between \( F_1 \) and \( T \), whereas ordinary correlation measures typically ignore nonlinear dependences [20]. For these reasons, the first step of our feature ranking algorithm is simply to sort all features, \( F_j, j = 1 : N_F \), according to \( I(F_j; T) \).

In some cases, high \( I(F_j; T) \) value may represent a non-direct relation. For example, if gender affects treatment assignment, we expect to observe high MI between the gender feature and \( T \), but also between the height feature and \( T \), as gender is highly informative about height. Ideally, we would like to highlight only features with a presumably direct relation to \( T \). Typically, we can expect that these feature will be characterized by higher MI values with \( T \) than features with an indirect relation to \( T \). Thus, we define \( F_k \) as having an “indirect relation” with \( T \) if and only if there exists another feature \( F_j \) such that \( I(F_j; T) > I(F_k; T) \) (i.e., \( F_j \) is ranked higher than \( F_k \)) and \( I(F_j; T) > 0.3 \text{ bits} \) (i.e., \( F_k \) demonstrates relatively strong dependency with \( F_j \)). Otherwise, we will say that \( F_k \) has a “direct” relation with \( T \). Various ways could be used to communicate this distinction to the user. For example, if \( F_k \) is considered to have “indirect relation” with \( T \) due to its relation to \( F_j \), then \( F_k \) can be moved to the end of the ranked list, and/or include a special “indirect relation” mark at the user-interface that
further indicates its high association with $F_j$. These issues are out of the scope of this manuscript and will be described in detail elsewhere. In Section 3 we focus on features that were identified to have a "direct" relation with $T$.

It is often useful to examine the joint relation of a combination of clinical features to the assigned treatment. For example, normal waist circumference differs between men and women. Thus, while the relation of this single feature to $T$ could be somewhat vague, the relation of this feature – jointly with the gender feature – over $T$, could reveal meaningful insights. To address this issue, we propose to accompany the univariate feature ranking described above with a corresponding pairwise feature ranking mechanism. Specifically, we start by sorting all feature pairs according to their $I(F_j, F_k; T)$ value. However, to avoid redundancy, we focus our attention on pairs with a synergistic relation over $T$, i.e., pairs that satisfy $S(F_j, F_k, T) > 0$. Furthermore, we use the same mechanism described above to distinguish pairs with a "direct" relation to $T$ vs. pairs with an "indirect" relation to $T$, and discuss only pairs with a "direct" relation to $T$ in Section 3.

In all cases, the MI terms were estimated using the method described in [20] in order to correct for finite sample effects. We note that the proposed scheme can be extended to consider higher order relations of clinical features to the assigned treatment, as long as the examined population of patients is large enough.

2.3 Probable Treatment Prediction

In the Probable Treatment Prediction (PTP) use case we aim to identify the most probable treatment for a new patient examined at POC, as implied by the statistics associated with treatments previously assigned to patients at the HCO. To that end, here we use the Naive Bayes (NB) classifier which is a relatively simple and well established classification scheme [6]. In addition, we describe the details of the "justification" and "confidence" that accompany each prediction provided by the system.

2.3.1 Probable Treatment Prediction Workflow

The Probable Treatment Prediction use case is targeted for the physician at POC, that is currently considering treatment allocation for a specific patient. The proposed workflow is as follows:

- Clinical data for the examined patient is retrieved by the system.
- The system predicts the most probable treatment, denoted $t^*$, but does not present it to the physician.
- The physician makes an initial decision regarding the treatment that should be administered, denoted $t_p$, and inputs her decision to the system.
- The system presents the prediction, $t^*$, with the associated justification and confidence.
- The physician makes the final treatment decision, after considering the system outputs.
- This final decision is further recorded by the system and can be exploited to refine future analysis and future predictions.

The systems predictions are based on patient records gathered in the HCO databases. As in the Retrospective Treatment Analysis workflow these records are processed every fixed period of time to incorporate new information gathered in the HCO databases.

2.3.2 The Naive Bayes Classifier

The NB classifier relies on the assumption that the features used by the classifier are independent given the class label. In our context, this implies that clinical features are independent given the assigned treatment, which is certainly void. Nonetheless, there is much evidence that NB classifiers tend to perform well even when their underlying assumptions are not valid (e.g., [5]).

Specifically, given the data associated with a new patient – denoted here as $f(i^*)$ – our classifier estimates the probability of each candidate treatment, $t_i$, via

$$P(t_i | f(i^*)) \propto P(t_i) \cdot (\prod_{j=1}^{N} P(f_j(i^*)|t_i)) \cdot (\prod_{j,k} P(f_j(i^*), f_k(i^*)|t_i)).$$

(6)

where $f_j(i^*)$ is the value associated with the feature $F_j$ for the $i^*$-th patient, the probabilities at the right hand side are estimated as described in Section 2.1.3, and the second product is taken over all pairs of features for which $S(F_j, F_k, S) > 0$. $^2$

Given the estimates of $P(t_i | f(i^*))$, the most probable treatment, $t^*_p = \arg\max_{t_p} P(t_p | f(i^*))$, is provided as the Probable Treatment Prediction for patient $i$.

2.3.3 Justification Module

In cases in which the systems prediction differs from the physician's initial decision, it is important to supply the physician with a possible explanation to this potential disagreement. Such an explanation may enable the physician to re-examine her decision in light of the system’s input, and determine if indeed her decision represents a deviation from the HCO guidelines and/or common practice. Thus, in the Justification Module of our system we aim to highlight clinical features that were most dominant in leading the system to predict $t^*$ rather than the physicians initial choice, $t_p$. To address that, for each feature $F_j$ we calculate the ratio of its conditional probability given the systems prediction and the physicians choice:

$$x_j = \log \frac{P(f_j(i^*)|t^*_p)}{P(f_j(i^*)|t_p)}$$

(7)

In the NB classifier, the higher this ratio is, the bigger the effect of the feature on the classification. Thus, we rank all features according to this ratio. We then present to the physician the patients values for the top ranking features, on top of $P(F_j | T)$, i.e., on top of the features conditional distribution given the treatment (see Section 3).

2.3.4 Confidence Estimation

Finally, there are various straightforward techniques to exploit the posterior distribution, $P(t_i | f(i^*))$, generated by the NB classifier, in order to approximate the confidence associated with each prediction. The common concept behind these methods is that when the posterior distribution tends to be more uniform the prediction of the NB classifier is presumably less reliable. Next, we describe three such measures.

$^2$Technically, the usage of synergistic pairs of features suggest that the classification scheme used here can be viewed as a semi-NB classifier [26].
for quantifying prediction confidence that are supported by our system:

- MAP – The posterior probability of the most probable treatment.
- Gap – The difference between the MAP value and the second most probable treatment [13, 21].

In our experiments all three methods provided similar results, hence in Section 3 we only report the results of the MAP measure. These results demonstrate that the reliability of the provided prediction can be properly estimated, allowing the user to ignore relatively unreliable predictions so to increase the precision of the system at the cost of reducing recall.

3. RESULTS

3.1 Data

We demonstrate our methodology over clinical data collected for hypertensive patients as part of the Hypergenes project [1]. These data were collected in several cohorts, from different countries and hospitals. To eliminate biases in data collection and treatment assignment between the different cohorts, and to try and simulate data which is routinely collected at an HCO database, we focused our analysis on only one cohort, EPOGHLuven, for which we had the most data. For hypertensive patients, the physician can consider one of three treatment options - Initiating life style changes (which we term “No Drug”), treating with a single drug (“Monotherapy”) or treating with a combination of drugs (“Combination therapy”). If choosing pharmacological therapy, the physician next faces the question of which agent to use. However, as the data about drug choice in our data is very limited, we do not handle this resolution here.

We limit our analysis to patients which suffer from hypertension (Diastolic blood pressure >90 or Systolic blood pressure >140), and for which the chosen treatment is known 3. Missing values are highly prevalent in clinical data, complicating the classification task, as most classifiers are not able to handle them. We therefore discarded features with more than 50% missing values and imputed missing values in the remaining features using KNN imputation [22]. This pre-process resulted in data for 1149 distinct patients and 15 clinical features.

3.2 Retrospective Treatment Analysis

The features MI score and ranking are summarized in Table 1. This Table 2 summarizes the features that were defined as having indirect interactions. Examining the features indicated as “indirect” reveals two probable causes for this indication. The first feature, height, has high mutual information with gender, and thus its relation to treatment assignment is most likely through gender (see below). The rest of the “indirect” features are all highly informative with BMI. For these features, it is possible that their relation to treatment assignment is “direct” but the information they contain is highly overlapping with BMI. Figure 1 presents the distribution of the four top ranking features amongst the different treatments. According to the European Society of Cardiology guidelines for the management of arterial hypertension [16] the decision which antihypertensive treatment to initiate is based on initial blood pressure levels and risk factors. In accordance we find blood pressure, and risk factors (Cardiovascular disease, patient age) amongst the top ranking features. Additionally we also find that gender holds much information about treatment decision, as 44% of the women receive drug therapy, compared to 24% of the men. Although gender is not a consideration according to the guidelines, it has already been observed that women with hypertension are more likely to be treated pharmacologically [15]. This demonstrates the utility of our system in recapturing known guidelines together with customary behaviours of the organization which are not depicted in guidelines.

We rank all 43 synergistic pairs \( (S(F_k, F_i, T) > 0) \) according to their MI with the target. Figure 2 presents the distribution of the three top ranking feature pairs. Examining the bottom left panel, we notice that high waist circumference (bins 2 and 3, 88–97 cm and > 97 cm respectively) is uncommon in women that do not receive drug treatment, while it is common in men that do not receive drug treatment. One possible explanation is that waist circumference above 88 in women is associated with an increased risk for chronic disease, and is considered a risk factor taken into account when deciding on treatment for hypertension [16]. Indeed, examining the bottom right panel, we observe that high waist circumference is relatively more common in women receiving drug combination therapy. However, for men the threshold is higher (102 cm), thus most men with a waist circumference falling in bins 2 and 3 do not suffer from this risk factor.

Table 1: Mutual information of features with “direct” relation to \( T \).

<table>
<thead>
<tr>
<th>Feature name</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cardiovascular diseases</td>
<td>0.1</td>
</tr>
<tr>
<td>Patient age</td>
<td>0.092</td>
</tr>
<tr>
<td>Systolic blood pressure</td>
<td>0.031</td>
</tr>
<tr>
<td>Gender</td>
<td>0.031</td>
</tr>
<tr>
<td>Myocardial infarction</td>
<td>0.026</td>
</tr>
<tr>
<td>BMI</td>
<td>0.025</td>
</tr>
<tr>
<td>Lipid lowering drugs given</td>
<td>0.025</td>
</tr>
<tr>
<td>Heart rate standing</td>
<td>0.011</td>
</tr>
<tr>
<td>Diastolic blood pressure</td>
<td>0.009</td>
</tr>
<tr>
<td>Current alcohol consumption</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 2: Mutual information of features with “indirect” relation to \( T \). For each “indirect” feature, \( F_k \), we report its MI with the assigned treatment and the MI with \( F_i \), the feature with the maximum MI with \( F_k \) amongst all features for which \( I(F_k; T) < I(F_i; T) \).

<table>
<thead>
<tr>
<th>Feature ( F_k )</th>
<th>Feature ( F_i )</th>
<th>( I(F_k; T) )</th>
<th>( I(F_i; T) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>Gender</td>
<td>0.031</td>
<td>0.44</td>
</tr>
<tr>
<td>Hip circ.</td>
<td>BMI</td>
<td>0.011</td>
<td>0.465</td>
</tr>
<tr>
<td>Waist circ.</td>
<td>BMI</td>
<td>0.01</td>
<td>0.346</td>
</tr>
<tr>
<td>Body weight</td>
<td>Waist circ.</td>
<td>0.001</td>
<td>0.491</td>
</tr>
<tr>
<td>Waist to hip ratio</td>
<td>Waist circ.</td>
<td>0.001</td>
<td>0.495</td>
</tr>
</tbody>
</table>

3For simplicity, if a patient received more than one treatment course we perform our analysis on the first treatment course she received.
3.3 Probable Treatment Prediction

3.3.1 Run details

To test the performance of the prediction algorithm we perform 10 repeats of stratified 3-fold cross validation. In the training phase the following steps are performed:

- kNN imputation of missing values in the training data.
- Quantization of continuous features into 3 equally populated bins.
- Mutual information and synergy estimation for all features and feature pairs.
- Feature pair selection according to synergy score, requiring $S > 0$.
- Training of the NB classifier.

The test phase consists of the following steps:

- kNN imputation of missing values in the test set, using only data of patients from the training set.
- Quantization of continuous features using the bins determined in the training phase.
- Treatment prediction for each patient in the test set.
- Prediction confidence estimate for each patient.

For each fold we calculate the precision and recall of our predictions, and average them over all folds and all repeats.
Figure 3: Mean precision and recall per treatment option. The error precision bars indicate one standard deviation above and below the mean.

Figure 4: Precision and recall for different confidence score thresholds. Macro average recall (horizontal axis) and precision [23] (vertical axis). All predictions were sorted by confidence score, and at each point the recall and precision score were calculated by taking into account only the top $p$ predictions.

### 3.3.2 Recall and Precision

The accuracy of the treatment prediction in terms of precision and recall are presented in Figure 3. These results demonstrate that our ability to distinguish between the decision to treat pharmacologically or not, is much higher than our ability to distinguish between mono and combination therapy. One possible method to improve prediction accuracy is to split the prediction into 2 successive tasks, in accordance with the European Society of Cardiology [16] guidelines approach; that is, first predicting whether the customary treatment is drug treatment, and then for patients for which drug treatment is determined as customary, aim to identify if customary treatment is mono or combination therapy. We leave this additional investigation for future work.

For each prediction, the system also generates a confidence score. As depicted in Figure 4, predictions for which the estimated confidence is higher indeed tend to be more correct. Specifically, the macro average precision is improved when considering only predictions with a high confidence score for different thresholds.

### 3.3.3 Justification

To examine the justification module we studied in detail cases in which the system's prediction is inconsistent with the physician's decision. In each case we rank the features by their contribution to the selection of the predicted treatment (Section 2.3.3). Figure 5 shows the top three ranking features, for a patient for which the physician's decision was no drug treatment, while the system predicted monotherapy as the most probable treatment. This view is similar to the retrospective view, with two differences: First, the features are ranked with respect to the specific patient; Second, the patient information is presented on top of the features distribution. This allows the physician to reconsider her choice in light of these features. For example, from the top selected feature, one can observe that this patient is a female suffering from a heart disease. As we can see from the top left panel, this combination of features is very uncommon among patients that did not receive drug therapy, implying that this treatment choice is indeed less common for such patients.

### 4. DISCUSSION

Mining the clinical data collected at HCOs has long been considered to hold great potential for improving patients care. Here, we proposed to apply – within a CDS system – ML techniques that focus on examining the treatments assigned to patients as part of the routine care delivery process at the HCO. We outlined two concrete use-cases that arise from this paradigm – Retrospective Treatment Analysis (RTA) and Probable Treatment Prediction (PTP), and demonstrated the applicability of both use cases over real world data of hypertension patients. Other use cases under the same paradigm merits further investigation.

The proposed approach is somewhat complementary to ML techniques aiming to predict the clinical outcome of treatments considered for an individual patient [7, 17, 25]. However, it does not require the availability of treatments outcome, i.e., it is less demanding in terms of the required patient data, which naturally increases its potential to positively impact healthcare providers’ performance in practice [12]. Furthermore, in some cases, identifying the most probable treatment – as addressed in the proposed PTP use-case – is an important pre-requisite for outcome prediction [18].

It is important to note that the most probable treatment identified by our system is not necessarily the treatment leading to the most favourable outcome for the examined patient. Rather, it merely represents the customary treatment at the HCO for patients that are found somewhat similar to the examined patient. In particular, if a non optimal treatment is often given at the HCO to a group of patients with similar statistical characteristics, this trend might be increased if the proposed PTP use-case is applied without caution. Two measures should be taken to avoid such undesirable trends. First, the Retrospective Treatment Analysis supported by our system should be executed regularly in order to highlight the clinical features associated with treatment assignment, aiming to improve the internal understanding at the HCO regarding treatment allocation, and further invoke internal discussions and perhaps initiate research, as appropriate. Second, it is crucial that the proposed paradigm will be used in the context of existing clinical knowledge. Thus, for example, in the mature sys-
Another potential difficulty is that a statistical relation between clinical features and assigned treatments could arise due to various reasons, each having different interpretations. For example, if a large scale clinical trial is conducted at the HCO it will obviously bias the statistical analysis performed by the proposed system. Thus, ideally, treatments assigned as part of clinical trials should be marked at the HCO database and correspondingly be treated separately.

In addition, a statistically based prediction algorithm is naturally prune to errors. As a result, there could be different reasons underlying cases where the most probable treatment – as identified under the PTP use case – is different from the treatment considered by the physician as most appropriate for the examined patient. One possibility is that the system properly identified the most probable treatment, i.e., the physician initial recommendation is not typical, compared to the data collected at the HCO database. However, it might be that the system prediction is wrong, e.g., due to over simplistic statistical modelling assumptions used by the prediction algorithm. In this case, the disagreement between the physician and the provided prediction represents an error of the prediction algorithm. Obviously, each of these two scenarios have different implications. Thus, being able to properly distinguish between these two cases is an important direction for future work.

We note that the final decision taken by the physician – after considering the provided prediction, could be exploited in this context. Specifically, it might be that cases where the physician eventually assigns a treatment which is different from the one predicted by the system are more often associated with wrong system predictions, compared to cases where the physician has changed her initial recommendation to match the system prediction. Thus, focused analysis of these physician/system disagreement cases may highlight problems associated with the statistical analysis performed by the system; updating the prediction algorithm accordingly could gradually reduce the rate of wrong system predictions.

The proposed approach can be viewed as a data-driven statistically based “soft” form of guidelines adherence. In parallel, evidence-based approaches for healthcare systems are advancing. Recently IBM has introduced Watson [9], a computer system that can deeply analyse large quantities of unstructured text to answer questions in different domains. In the context of treatment allocation such a system can analyse research papers and clinical guidelines to provide answers about best-practice practice treatment for a specific patient. We believe that this approach can be considered complementary to the one presented here, providing different perspective to the same question.

The statistical insights emerging from using our proposed system are expected to reflect the statistics of the routine care delivery process at the HCO where the system is installed. These insights may echo with existing general guidelines, but may also add novel refinements to these guidelines, and at the extreme may even represent contradictions with these guidelines. It is therefore important that the mature system will be able to smoothly integrate relevant clinical knowledge, and will further clearly highlight the relation of...
existing clinical evidence to the results obtained by the statistical analysis.

5. ACKNOWLEDGMENTS

We thank the Hypergenes project (http://www.hypergenes.eu/) for making their data available, Costanza Conti for helpful insights about the data, Ariel Farkash and Daphna Landau for assisting in the data preprocessing, and Amnon Shabo and Ehud Aharoni for insightful discussions.

6. REFERENCES


