A system for identifying and investigating unexpected response to treatment

Michal Ozery-Flato, PhD 1, Liat Ein-Dor, PhD 1, Hani Neuvirth, MSc 1, Naama Parush, PhD 1, Martin S. Kohn, MS, MD, FACEP, FACPE 2, Jianying Hu, PhD 3, Ranit Aharonov, PhD 1

1 IBM Research - Haifa, Israel; 2 Sentrian, Aliso Viejo CA, US; 3 IBM Research - T.J. Watson, New York, US

Abstract

The availability of electronic health records creates fertile ground for developing computational models for various medical conditions. Using machine learning, we can detect patients with unexpected responses to treatment and provide statistical testing and visualization tools to help further analysis. The new system was developed to help researchers uncover new features associated with reduced response to treatment, and to aid physicians in identifying patients that are not responding to treatment as expected and hence deserve more attention. The solution computes a statistical score for the deviation of a patient’s response from responses observed in individuals with similar characteristics and medication regimens. Statistical tests are then applied to identify clinical features that correlate with cohorts of patients showing deviant responses. The results provide comprehensive visualizations, both at the cohort and the individual patient levels. We demonstrate the utility of this system in a population of diabetic patients.

1. Introduction

Chronic disease represents the leading cause of mortality in the developed world and consumes the majority of healthcare expenditures 1. A chronic illness typically extends over many years and is often associated with progressive deterioration, punctuated with episodes of acute exacerbations. Successfully controlling a chronic condition such as diabetes or hypertension remains difficult for several reasons. Selection of an effective medication regimen varies significantly between patients due to critical differences between their medical and demographic characteristics, as well as the multiple disease progression pathways. While many medications have become available for each chronic condition, it is often impossible to predict which class, dosage, frequency, and combinations of drugs will be required to achieve control for each individual patient. As a result, treatment regimens typically evolve over time through a trial-and-error process.

The rapid increase in the adoption of electronic healthcare records has made large amounts of patient data available. This provides data mining researchers with a unique opportunity to develop novel computational models that can help achieve personalized chronic disease management. Patients with chronic diseases are periodically tested to monitor the condition of their disease and evaluate their response to prescribed medications.

In this paper, we present a system that can evaluate a patient’s clinical outcome in the context of similar patients. Our system compares a patient’s outcome to that of other patients with similar medical conditions who were prescribed a similar medication regimen. The system computes a statistical score for the deviation of the patient’s response from responses observed in similar patients. These scores are used to define cohorts of patients with deviant poorer outcomes; statistical tests are then applied to identify clinical features that correlate with these cohorts. The system provides comprehensive visualizations of the analysis results and the supporting data, both at the cohort level and at the level of individual patients. In developing this system, we made four primary contributions:

1. The formulation of the problem of unexpected response identification as one of “contextual anomaly detection” 2.
2. The use of a patient similarity metric to define the clinical context in which a given patient’s response is evaluated.
3. The use of outlying patients to identify additional clinical features potentially associated with less favorable outcomes.
4. A tool that implements the methods described above alongside a user interface for managing patient cohorts, visualizing patient data, and analyzing results.

2. Background

Anomaly detection methods have been used extensively in many domains, including medicine and healthcare 2. Most prior studies in the medical domain that used anomaly detection methods focused on point anomalies; that is, these studies identified individual data instances that are anomalous with respect to the rest of the data. In this study, we focus on detecting contextual anomalies 7, in which data instances are anomalous in a specific context. In this setting, we defined each data instance using two sets of attributes: (i) contextual attributes, which are used to determine the context of that instance, and (ii) behavioral attributes, which define the non-contextual characteristic of that instance. In our case, patients correspond to data instances; the characteristics that were used to define similarity between patients are the contextual attributes; and the
response to treatment was the single behavioral attribute. In a previous study, Hu and colleagues\textsuperscript{3} applied contextual anomaly detection methods to detect patients with anomalous healthcare utilization patterns.

Various ways of using patient similarity analytics to derive personalized insights have been proposed in recent years. Ebadollahi et al.\textsuperscript{4} developed a supervised learning method to derive patient similarity metrics and used it to perform near-term prognosis for Intensive Care Unit patients. Extensions of this method were later used for disease risk prediction\textsuperscript{5,6} and treatment comparison.\textsuperscript{7} To the best of our knowledge our approach is the first time patient similarity measures have been used for anomaly detection in medication response.

3. Methods
3.1 Problem formulation
When defining the problem, we used the following definitions: (i) The target population is a group of patients with a certain disease (“the disease”); (ii) the patients are prescribed medications for treating the disease (“the treatment medications”); and (iii) the condition of the disease is monitored by a certain clinical test (“the test”). The clinical test is used to evaluate the efficacy of the treatment medications during the time period preceding the test. The test results can indicate if those effects meet the expected ones, given the patient’s characteristics and the specific medication(s) prescribed. We refer to the time period prior to the test as the “treatment evaluation period” and to the test result as “the patient’s response”. The time period prior to the treatment evaluation period is called “the baseline period”.

![Figure 1: Problem formulation](Image)

In the diabetes case study, we added the following conditions: (i) Patients are tested on the last day before the beginning of the treatment evaluation period; (ii) the treatment evaluation period is limited to three to six months; and (iii) the baseline period is six months. These requirements are aligned with the treatment policy of diabetic patients, which requires HbA\textsubscript{1C} monitoring every three months. Diabetic patients were identified based on abnormal HbA\textsubscript{1C} lab test results. Pregnant women and children under the age of 18, as well as type I diabetic patients (identified by the ICD-9 codes 250.x1, 250.x3, 249.x1, or 249.x3) were excluded from the analysis. Finally, our inclusion criteria required that patients have (i) baseline HbA\textsubscript{1C} equal or higher to 7, and (ii) at least one prescription for anti-diabetic drugs during the treatment evaluation period.

We used the following notations: $Y$ denotes a patient’s response; $X_{\text{drugs}}$ denotes the vector of variables relating to the treatment, such as proportion of days covered and mean dosage (in mg) per drug in the evaluation period, as well as the change (difference) from baseline period; $X_{\text{conf}}$ denotes the vector of confounding variables, in our case age, gender and baseline HbA\textsubscript{1C}. Finally, $X_{\text{other}}$ denotes a set of available clinical features not included in $X_{\text{conf}}$ or $X_{\text{drugs}}$. The aim of this work is to detect patients whose response to treatment significantly deviates from the expected response based on (i) the treatment medications (ii) the additional confounding factors. Thus the feature vector we used to represent a patient, $X=[X_{\text{drugs}}, X_{\text{conf}}]$, was composed of both the variables in $X_{\text{drugs}}$ and the variables in $X_{\text{conf}}$. As discussed in Section 3.5, the variables in $X_{\text{other}}$ are not used to model the patient’s response to treatment, but are tested for association with reduced response.

We filtered out features in $X_{\text{drugs}}$ and $X_{\text{other}}$ that were constant or linearly dependent on other features in their group. As a preprocessing stage, each feature $f$ was normalized to have values between 0 and 1. We assumed the interpretation of the response to be monotonic: a worse response corresponds to either higher values or lower values of the test result. For example, for the HbA\textsubscript{1C} test, the higher the value, the worse the response is. For simplicity, and without loss of generality, we consider below only the former case (i.e., higher values = worse response, while lower values = better response).

3.2 Detecting patients with unexpected responses
To detect patients with anomalous responses to treatment, we first estimated the distribution of each patient’s expected response. We then used this distribution estimate to calculate a statistical score for the deviation of the observed response from the distribution mean. We assumed that the expected response has a Gaussian distribution. To estimate the mean of the distribution, we used the kernel regression method. This is a well-established method for nonlinear regression in which the target value for a test point is estimated using a weighted average of the surrounding samples.\textsuperscript{8} Our rationale for choosing this method was that patients with similar values in the $X_{\text{drugs}}$ and $X_{\text{conf}}$ features are expected to have similar responses to treatment. Following the same line, we used a distance-based weighted average to estimate the distribution variance.

More formally, let $\hat{Y}_i$ be a random variable denoting the expected response of patient $i$. We estimate the mean, $\mu(\hat{Y}_i)$, and the standard deviation, $\sigma(\hat{Y}_i)$, of the expected response $\hat{Y}_i$ in the following way:
\[
\mu(\bar{Y}_i) = \frac{\sum_{j \neq i} w_{ij} Y_j}{\sum_{j \neq i} w_{ij}}, \quad \sigma(\bar{Y}_i) = \sqrt{\frac{\sum_{j \neq i} w_{ij} (\bar{Y}_j - \mu(\bar{Y}_i))^2}{\sum_{j \neq i} w_{ij}}},
\]

where \(w_{ij}\) is a weight corresponding to the distance between patients \(i\) and \(j\). Along with the assumption that samples in proximity to one another exhibit similar responses, one could expect the weights to decrease in value with increased distance from the target patient. Here we used the Gaussian kernel to determine the weights such that \(w_{ij} = e^{-cd_{ij}^2}\) where \(d_{ij}\) is the distance between patients \(i\) and \(j\), and \(c\) is a constant determining the Gaussian width. We absorbed \(c\) in \(d_{ij}\), and fixed \(c = 1\). The distance metric we selected for calculating \(d_{ij}\) is described in Section 3.3.

Using \(\mu(\bar{Y}_i)\) and \(\sigma(\bar{Y}_i)\), we compute a Z-score, \(Z(Y_i) = \frac{Y_i - \mu(\bar{Y}_i)}{\sigma(\bar{Y}_i)}\), for each patient \(i\). This score reflects the anomaly level of the response given the responses of other patients. Clearly, the higher the Z-score is, the lower the expected response value. We considered patients as outliers by the following procedure. We used the Gaussian distribution to compute the right tail p-value for each Z-score. We then flagged a patient as an outlier if his or her associated P-value was significant at a false discovery rate (FDR) of 10%. Thus, outliers are patients whose responses to treatment are significantly worse than expected.

### 3.3 Distance metric learning

The distance function that is used to calculate \(d_{ij}\) has a crucial influence on the resultant \(\mu\) and \(\sigma\) values. The most trivial selection of a distance function is the Euclidean distance. However, this function may not be appropriate in our case, since different features may have different relevance to the ultimate response, while the Euclidean distance gives the same influence to all features. To find a distance function that takes into account the relevance of the features to the final response, we learned a linear transformation of the data points that minimizes the mean squared error (MSE) of the Gaussian kernel regression of the response \(Y\) \(^8\). Under this transformation, denoted by \(A\), the distance between two vectors \(x_i, x_j\) is defined as

\[d_{ij} = \|A(x_i - x_j)\|^2.\]

We limited \(A\) to be diagonal for two main reasons. First, the obtained metric is easier to interpret in terms of the features’ influence on the response. Second, in this way the number of parameters to learn was significantly reduced, simplifying the learning process and lessening the risk of overfitting. We further limited the number of parameters by selecting only features that significantly correlated with the outcome at FDR-10%. Finally, when estimating \(\mu(\bar{Y})\) and \(\sigma(\bar{Y})\), we considered only the 100 closest patients. Not only did this speed up the computation, it also resulted in better prediction accuracy (lower MSE).

### 3.4 Singleton patients exclusion

The confidence of model results for patients whose closest neighbors are relatively far is much lower. Due to this expected reduced confidence, we seek to identify those relatively distant patients who correspond to point anomalies in the feature space. There are many algorithms for point anomaly detection \(^2\). In this study, we tested two common algorithms: One-class Support Vector Machines (OCSVM) \(^3\)\(^^5\), and distance to K-nearest neighbor (KNN). We applied the two algorithms with the learned distance metric to find the top 5% of the patients who are most anomalous in the feature space. We compared the two algorithms by the improvement in the prediction accuracy (MSE) after exclusion of the singleton patients they identified, and then selected the method leading to better accuracy.

### 3.5 Identifying features associated with reduced response

Given a cohort of outlying (i.e., reduced response) patients and a second cohort of non-outlying patients, we looked for features whose distribution differs between the two cohorts. In this study we used T-test and Chi-square tests. Additionally, we applied “context-normalization” to each feature and repeated the statistical tests. This special normalization uses the model’s metric to assign Z-scores to the feature values, similar to the way response Z-scores are computed for the response variable \(Y\) (see Section 3.2). In the case of binary features, we assumed that their values are drawn from a Bernoulli distribution. For given feature \(f\) and patient \(i\) we estimated the probability \(P(f_i = 1) = \frac{\sum_{j \neq i} w_{ij} f_j}{\sum_{j \neq i} w_{ij}}\). Using context normalization can uncover new associations involving features whose values for patients with unexpected response is high (equivalently, low) compared to similar patients.

### 4. The system

We developed a tool that implements the methods described above and offers a user interface for applying these methods, as well as visualizing the data and analysis results. It allows users to define and manage different patient cohorts, and analyze the differences between them. Cohorts are represented in a tree-like structure allowing a series of nested definitions of cohorts. For each cohort, users can view summary statistics, review patients, and flag and add comments to specific individuals. At the level of specific patients, the tool offers a comprehensive and configurable visualization of the longitudinal clinical data. See Figure 2 for an example of the system’s patient view.
5. Results

5.1 Data statistics
We used a database of outpatient encounters for approximately 200,000 patients, cared for by a network of primary care physicians over a three-year period. See \textsuperscript{10} for additional information on this database. We derived a cohort of 2209 type II diabetic patients and computed $Y$ (final HbA$_{1C}$), $X_{\text{conf}}$, $X_{\text{drugs}}$ (135 features), and $X_{\text{other}}$ (1391 features). $Y$ was strongly correlated with baseline HbA$_{1C}$ ($r=0.6$) and age ($r=-0.17$), which are both contained in $X_{\text{conf}}$.

5.2 Metric features
Out of the 138 features in $X=[X_{\text{drugs}}, X_{\text{conf}}]$, 20 features were selected for the full model (see Section 3.3). Baseline HbA$_{1C}$ was the dominating feature in the metric, as its weight was larger than the sum of all other features’ weights. Interestingly, in this metric, features with information concerning the change in the consumption of anti-diabetic drugs had larger weights than corresponding features with data on the actual consumption.

5.3 Model evaluation
We defined the baseline model as the one returning the average final HbA$_{1C}$, excluding the target patient. Our model had a better accuracy (MSE) than the baseline model (1.03 vs. 1.51). Removing $X_{\text{drugs}}$ reduced our model’s accuracy (MSE =1.15), indicating the importance of treatment features. The exclusion of singleton patients improved the accuracy, with KNN outperforming the OCSVM method (MSE 0.9 vs. 0.96, see Section 3.4). When using the baseline model to detect unexpected responses, there were 50 outliers ($Y \geq 11.1$). Our model detected 29 patients with outlying response, of which only 6 belonged to the baseline outliers. This demonstrates that when considering the medication regimen and confounding factors, the group of patients with unexpected response may differ substantially from the group of patients having extreme outcome values.

5.4 Associated features
We used our system to systematically explore associations with cohorts of patients among a large set of available features, (see Sections 3.5 and 4). Being a distinct cohort, it was interesting to characterize the patients identified as singletons. Baseline HbA$_{1C}$ showed the strongest correlation, and indeed 21 of the singleton patients belonged to the group of baseline model outliers. Age, use of anti-diabetic drugs, and specifically the use of oral anti-hyperglycemic drugs were negatively associated with this cohort. The frequency of visits to the doctor’s office and pharmacy was significantly lower, and the standard deviation of the time between visits to these facilities was significantly larger. This suggests that singleton patients are characterized by fewer regular meetings with physicians, leading to poor diabetes management.

We next analyzed which features were associated with the cohort of patients identified by our model as having an unexpected response to treatment. The set of features correlating with this cohort included many sparse features. Among the less sparse features were: general use of anti-depressant drugs; the use of the antibiotic levofloxacin during the evaluation period; “difl Mg pioglitazone hydrochloride” (Actos) showed a negative correlation. We further analyzed this cohort, searching for correlations with context-normalized features (see Section 3.5). This analysis yielded additional significant correlations with the consumption of certain drugs during the evaluation period. Some of these drugs are known to be associated with elevated blood sugar, such as Cymbalta (duloxetine hydrochloride), an anti-depressant drug, and simvastatin, a cholesterol-lowering drug. Another associated drug was the antibiotic amoxicillin. The use of antibiotic drugs such as levofloxacin and amoxicillin might indicate infections that may cause metabolic stress, which is known to be associated with poor diabetes control.
5.5 Identifying additional correlating feature by manual inspection

The physician on our team used the tool described in Section 4 to review the clinical data of patients detected as having deviant responses by our model. For many of these outlying patients, a thorough inspection of their medical data indicated possible explanations for their unexpectedly high blood sugar. Some of these explanations were connected to features correlated with the cohort of unexpected response, such as parallel use of drugs that increase blood sugar. Some other explanations were connected to high stress events, such as hospitalizations due to congestive heart failure. Several patients stopped using Actos (pioglitazone hydrochloride) and Avandia (rosiglitazone). Actos and Avandia are associated with elevated cardiovascular risk, which may explain why physicians consider switching them. We tested for each anti-diabetic drug and whether its discontinuation was associated with the cohort of unexpected response. Indeed, for both Actos and Avandia this feature was found to be significantly correlated at FDR-10% with the cohort of unexpected response. An additional drug that arose in this analysis was Byetta (exenatide), which is generally recommended as a supplementary drug.

6. Discussion

In this study we presented a system for detecting and studying unexpectedly reduced responses to treatment in the context of similar patients. Based on this retrospective analysis, patients detected as anomalous may be considered to be at higher risk for continuing treatment failures. Therefore, investigating these patients is of interest for both clinicians and researchers seeking to improve the given treatment per patient. Our system is designed to meet the needs of such investigations. Our system can be configured to use alternative techniques for metric learning (Section 3.3), singleton identification (Section 3.4), and association tests (Section 3.5). The techniques we used here were selected to exemplify our methodology.

Our tool provides different visualizations of the data, both at the cohort level and for individual selected patients. Data visualizations are vital for validating and interpreting analysis results and sifting important correlating features from ones that are less interesting. From a physician’s perspective, viewing the entire medical profile of a patient in a succinct but comprehensive manner is critical. This functionality provides physicians with a better and quicker understanding of a patient’s clinical status. In particular, the patient view in our tool allowed our team’s physician to quickly review the clinical status of patients detected as having unexpectedly bad responses.

We note that some of the known confounding factors for diabetes control, such as obesity measurements, were not available in our dataset. Clearly, the implementation of our methods depends strongly upon the availability of the medication and known confounding features in the medical database being used. This study demonstrates that even in the presence of incomplete data, it is possible to identify novel factors that associate with unexpected response and merit further research. As more medical information becomes available, the more accurate will be the findings our tool reports.

Acknowledgement

We thank Gerald Tesauro for sharing the code for metric learning for kernel regression.

References

1. Bodenheimer T, Chen E, Bennett HD. Confronting the growing burden of chronic disease: can the U.S. health care workforce do the job? Health Aff Proj Hope. 2009;28(1):64–74.
2. Chandola V, Banerjee A, Kumar V. Anomaly detection: A survey. ACM Comput Surv. 2009;41(3):15:1–15:58.
3. Hu J, Wang F, Sun J, Sorrentino R, Ebadollahi S. A healthcare utilization analysis framework for hot spotting and contextual anomaly detection. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2012;360–9.
4. Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C. Predicting Patient’s Trajectory of Physiological Data using Temporal Trends in Similar Patients: A System for Near-Term Prognostics. AMIA Annu Symp Proc. 2010;192–6.
5. Wang F, Sun J, Hu J, Ebadollahi S. iMet: Interactive Metric Learning in Healthcare Applications. In: Liu B, Liu H, Clifton C, Washio T, Kamath C, editors. Proceedings of the 2011 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2011. p. 944–55.
6. Wang F, Sun J, Ebadollahi S. Integrating Distance Metrics Learned from Multiple Experts and its Application in Patient Similarity Assessment. In: Liu B, Liu H, Clifton C, Washio T, Kamath C, editors. Proceedings of the 2011 SIAM International Conference on Data Mining. Philadelphia, PA: Society for Industrial and Applied Mathematics; 2011. p. 59–70.
7. Markatou M, Don PK, Hu J, Wang F, Sun J, Sorrentino R, et al. Case-based reasoning in comparative effectiveness research. IBM J Res Dev. 2012;56(5):4:1–4:12.
8. Weinberger KQ, Tesauro G. Metric learning for kernel regression. International Conference on Artificial Intelligence and Statistics. 2007. p. 612–9.
9. Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. ACM Trans Intell Syst Technol. 2011;2(3):27:1–27:27.
10. Neuvirth H, Ozery-Flato M, Hu J, Laserson J, Kohn MS, Ebadollahi S, et al. Toward Personalized Care Management of Patients at Risk: The Diabetes Case Study. Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2011. p. 395–403.