

# Dissimilarity Measures for Detecting Hepatotoxicity in Clinical Trial Data\*

Matthew Eric Otey   Srinivasan Parthasarathy   Donald C. Trost  
Contact: srini@cse.ohio-state.edu

## Abstract

In clinical trials, pharmaceutical companies test the efficacy and safety of a new drug for the treatment of a disease by comparing the results from a large number of diseased and healthy patients exposed to either the new drug, an existing drug that treats the disease, or a placebo. One primary concern is liver toxicity, which is usually diagnosed by blood analyte tests. Often, such signals of toxicity lead to the discontinuation of drug development or the withdrawal of the drug from the market. Early detection of liver toxicity can save lives and also save such companies billions of dollars. Existing approaches for detecting liver toxicity typically ignore correlations between blood analyte values, but in this work we present novel dissimilarity measures based on principal component analysis which can be used for detecting liver toxicity and identifying subpopulations who may be susceptible by taking into account the correlations structure of the data. Experimental results on real clinical trial data validate our approach.

**Keywords:** Drug efficacy and safety analysis, outlier detection, clustering, dissimilarity measures.

## 1 Introduction

Drug safety issues have received much attention in the past year. Several large pharmaceutical companies have issued warnings or removed their drugs from the market following reports of severe or deadly side-effects. Such events are harmful to the companies' public image, and their financial status. Each company invests large amounts of money in developing and testing new drugs. Any drug under development or clinical trials that does not make it to the market represents a huge loss for the company. Also, any drug that makes it to market but must be withdrawn represents a double loss for the company, as it is unable to recoup development costs, and may be held liable for any harmful effects of that drug. Therefore, pharmaceutical companies have an intense interest in discovering any harmful effects of their drugs as early as possible, so they can cease

development or sales in order to save both lives and money.

The safety and efficacy of new drugs are determined using a set of clinical trials. Clinical trials occur in four phases. The ability to identify harmful drugs and cease development at least one phase or a couple of years earlier than usual can save a pharmaceutical company billions of dollars. In pharmaceutical clinical trials, the efficacy and safety of a drug for treating a particular disease is studied by comparing the results from several groups of subjects. These include groups of healthy subjects and groups of unhealthy subjects who are randomly assigned to the experimental drug, existing therapies for the disease, or a placebo. Safety is studied in many ways; serial clinical laboratory blood tests are used commonly to monitor biochemical changes in the body. A common reason for stopping a drug development project or causing discontinuation in a particular patient or group of patients are abnormal blood test values related to the liver, as it has a major detoxifying function. When liver tests are high, it is assumed that hepatotoxicity, or liver toxicity, is present. However, the rules for determining the presence of drug-induced hepatotoxicity are mostly qualitative and involve considerable clinical judgment. The current state-of-the-art in pharmaceutical research uses ad-hoc univariate rules applied to multiple analytes. For example, "Hy's Rule" [2] requires the crossing of at least two univariate thresholds. The problem of misclassification should be obvious, since hepatotoxicity may not be so much correlated with absolute elevated blood analyte values as it is with how the analytes move together. Our hypothesis is that Hy's rule is not sufficient, and that correlations between analytes are extremely important for understanding the effects of a drug on liver toxicity.

Clinical trial data is usually in the form of a set of multivariate time series, where each variable corresponds to a blood analyte and each series corresponds to a different patient. In this paper we examine the notion of quantifying the dissimilarity between different sets of data with the goal of detecting hepatotoxicity. We propose dissimilarity measures that can be used to quantify the differences between two data sets. Other applications of our measure for clinical trial data involve characterizing the differences between the different subsets of patients and discovering subpopulations

---

\*Matthew Eric Otey and Srinivasan Parthasarathy are in the Department of Computer Science and Engineering at the Ohio State University. Donald C. Trost is an employee of Pfizer, Inc. This work was sponsored by Pfizer as part of a collaborative project in pathodynamics methodology.

that have a greater risk of hepatotoxicity. The measures we propose are based on principal component analysis (PCA). Our measures consists of components that separately take into account differences in the locations and correlations of the data sets being compared. It is also possible to weight the components differently, so one can incorporate domain knowledge into the measure. Finally, our measure is robust towards noise and missing data. We demonstrate the efficacy of the proposed measures using clinical trial data provided by Pfizer that is known to contain subjects exhibiting hepatic changes.

## 2 Related Work

There have been many metrics proposed that find the distance or similarity between the records (e.g. the Euclidean distance) or between the attributes of a data set (e.g. correlation). However, these metrics are not suitable for comparing two different data sets. Similarity metrics for comparing two data sets have been used in image recognition [6], and hierarchical clustering [7]. However, these metrics do not explicitly take into account the correlations between attributes in the data sets. Parthasarathy and Ogihara [12] propose a similarity metric for clustering data sets based on frequent itemsets. This metric takes into account correlations between the attributes, but it is only applicable for data sets with categorical or discrete attributes. There has also been work on defining distance metrics that take into account the correlations present in continuous data. The most popular metric is the Mahalanobis distance [13], which accounts for the covariances of the attributes of the data. However this can only be used to calculate the distance between two points in the same data set.

## 3 Algorithms

As we discussed in Section 1, clinical trial data are presented in the form of a multivariate time series for each subject in the trial. At each time point, the values of various blood analytes are recorded. While there are many techniques for analyzing (multiple) times series data [1, 3, 5], clinical trial time series data is quite challenging. Clinical trial time series data sets suffer from irregular sampling, missing data, and varying lengths. This may be due to a variety of reasons, including missed appointments, unexplained absences, and drop outs. Furthermore, there are also several potential sources of noise. Measurement errors, laboratory bias <sup>1</sup>, and circadian effects on analyte values (depending on when the blood sample was drawn) can

<sup>1</sup>Different laboratories, where these tests are often analyzed, often have different protocols resulting in a significant variation in analyte values for the same subject.

be contributing factors to noise.

The basis of Hy’s rule, and the typical signal physicians look for when evaluating liver toxicity, is usually a significant and consistent departure from the normal levels of one or more liver analytes. Moreover, it is usually the case that not all the analytes are affected simultaneously. A conclusion one can draw from these two statements is that the correlation among analytes should be capable of identifying such significant departures from the norm.

This key intuition leads us to use correlation or covariance matrices to represent patient data. We subsequently use principal component based methods for computing dissimilarity measures for such datasets. We note that correlation and covariance matrices can easily be imputed in the presence of missing data by using the Expectation-Maximization algorithm [4] to find the maximum-likelihood values of the covariance or correlation matrices. Moreover, principal components based techniques have been shown in the literature to be noise-tolerant [11]. We note that such measures are general-purpose, and can be used to compare any two data sets, so long as they have the same dimensionality [9].

**3.1 Dissimilarity Measures** Our goal is to quantify the dissimilarity of two  $k$ -dimensional data sets  $\bar{\mathbf{X}}$  and  $\bar{\mathbf{Y}}$ . Our measures take into account the correlations between the attributes of the two data sets. In general, the dissimilarity of two data sets  $\bar{\mathbf{X}}$  and  $\bar{\mathbf{Y}}$  is denoted as  $D(\bar{\mathbf{X}}, \bar{\mathbf{Y}})$ . We define the function  $D$  in terms of two dissimilarity functions that take into account the differences the magnitude and direction of the variance in the data sets. These components are combined by means of a weighted sum, which allows one to weight the components differently, so as to incorporate domain knowledge.

The first step in using our dissimilarity measures is to the find the principal components of the data sets being compared. The principal components of a data set are the set of orthogonal vectors such that the first vector points in the direction of greatest variance in the data, the second points in the orthogonal direction of the second greatest variance in the data, and so on [14]. We consider  $\bar{\mathbf{X}}$  and  $\bar{\mathbf{Y}}$  to be most similar to each other when their principal components, paired according to their ranks, are aligned and have the same magnitude, and most dissimilar when all of the components of  $\bar{\mathbf{X}}$  are orthogonal to those of  $\bar{\mathbf{Y}}$ .

Using the principal components, we can represent each data set  $\bar{\mathbf{X}}$  as a single feature vector  $F_{\bar{\mathbf{X}}}$ :

$$(3.1) \quad F_{\bar{\mathbf{X}}} = \sqrt{\lambda_1} \times X_1$$

where  $X_1$  is the first principal component of the data set, and  $\lambda_1$  is its corresponding eigenvalue. That is to say each data set is represented by the scaled primary

principal component *vector* pointing in the direction of greatest variance.

Having such a feature vector, we can then apply any standard distance metric. For example, applying the Euclidean distance metric:

$$(3.2) \quad D_e(F_{\underline{X}}, F_{\underline{Y}}) = |F_{\underline{X}} - F_{\underline{Y}}|_2$$

on the first principal component derived from the covariance matrix of the data would result in a value that simultaneously measures differences in direction and magnitude of the vector.

This measure can be extended to account for the differences in the mean of the data sets. First we define the dissimilarity of the means of the data sets as follows:

$$(3.3) \quad D_\mu(\underline{X}, \underline{Y}) = |\mu_{\underline{X}} - \mu_{\underline{Y}}|_2.$$

that is to say, the Euclidean distance between the centroids of the two data sets. We can then define the extended  $D_e$  measure as follows:

$$(3.4) \quad D_e(\underline{X}, \underline{Y}) = \beta_0 + \beta_1 \times D_\mu + \beta_2 \times D_e.$$

This formulation allows us to weight differences in the means and correlations according to domain information. For example, in clinical trial data, differences in the means of the observations of two different subjects may be caused more by differences in demographic characteristics (e.g. sex, age, weight) than by any effect of the drug, and so one would want to weight the differences in correlations higher.

Finally, we note that we can generalize these measures to account for all the principal components as follows. Let  $F_{\underline{X}}^i$  be the feature vector for the  $i$ th component:

$$(3.5) \quad F_{\underline{X}}^i = \sqrt{\lambda_i} \times X_i.$$

Then the  $D_e$  measure can be generalized as:

$$(3.6) \quad D'_e(\underline{X}, \underline{Y}) = \sum_{i=1}^k D_e(F_{\underline{X}}^i, F_{\underline{Y}}^i).$$

## 3.2 Applications

**3.2.1 Anomaly Detection** Detection of anomalies or outliers in clinical trial data is very important. Subjects' analyte values may be anomalous for many reasons related to sample processing including subject ingestion of interfering substances, sampling handling conditions, analyzer error, and transcription error. If these data points can be identified and the cause attributed to a non-treatment-related event, then the data point may need to be removed from a particular analysis. Subjects' values may be anomalous because they are having abnormal reactions to the drug. If this is the case, the drug maker may want to study more subjects

similar to the anomalous ones to see if they are true anomalies or indicative a small sub-populations that may have toxic reactions to the drug. Using our dissimilarity measures, it is straightforward to implement basic outlier detection algorithms such as those described in [8]. These are nested-loop approaches that calculate the dissimilarity between each pair of data points (or in our case, each pair of subjects). Having calculated these values one can rank the data points (subjects) according to the sum of the dissimilarities from the  $k$  most similar subjects.

**3.2.2 Clustering** The dissimilarity measures we present above allow us to easily perform clustering of the subjects. Finding clusters of subjects in clinical trial data is helpful in that it allows us to identify sub-populations who may have a greater risk of hepatotoxicity, sub-populations on whom the drug may have little or no effect, sub-populations that may have a higher risk of severe side-effects, et cetera. This allows the drug makers to determine the efficacy of the drug, to determine dosage levels for different patients, and to determine if the side-effects are too severe or widespread to continue development of the drug. It is straightforward to perform agglomerative hierarchical clustering of data sets using our dissimilarity measures. If one has  $n$  data sets, one can construct an  $n$  by  $n$  table containing the pairwise dissimilarities of the data sets. Once this table has been constructed, one can use any distance metric (e.g. single-link) to perform the hierarchical clustering.

## 4 Experimental Results

**4.1 Setup** The first dataset we use, henceforth referred to as  $D1$ , consists of a set of subjects suffering from diabetes, who, in addition to their regular diabetes therapy, were receiving either a placebo or the study drug (drug A) for a diabetic complication. Since we are primarily concerned with hepatotoxicity, following suggestions from our domain experts, we only consider eight serum analytes (often referred to in the literature as the liver panel): ALT (alanine aminotransferase), AST (aspartate aminotransferase), GGT ( $\gamma$ -glutamyltransferase), LD (lactate dehydrogenase), ALP (alkaline phosphatase), total bilirubin, total protein, and albumin. Using advice from a domain expert, we use the logarithm transformation of the first six analytes' values (total protein and albumin are excepted). There is strong evidence that the data after this transform is multi-variate Gaussian [10]. This dataset consists of 446 patients on placebo and 680 patients on drug. Development on this drug was discontinued in Phase III for various reasons including possible hepatotoxicity.

The second dataset we use, henceforth referred to as  $D2$ , consists of a set of post-menopausal women, who

again were given either a placebo or one of two drugs ( $B, C$ ) (both are different from drug  $A$ ). Again, we limit our focus to the liver panel. This dataset consists of 201 patients on placebo, 41 patients on drug  $B$ , and 126 patients on drug  $C$ . Both drugs  $B$  and  $C$  are on the market, and are expected to have little or no hepatotoxicity.

Both datasets suffer from the problems we mentioned earlier. They contain missing data, unequally spaced time series data for different patients, some patients had many readings over a period of time, others had much fewer etc. Since the differences in the mean are not significant in these data sets, we use the basic forms of the  $D_e$  measure defined in equation 3.2 in these experiments. All of our implementations are done using Octave, an open-source version of Matlab.

**4.2 Anomaly Detection** In our first experiment, we want to see how our dissimilarity measures perform on the clinical trial data set of diabetic patients. As noted earlier we have two groups of patients: one on placebo, and another on the drug under study. The experiment we conduct is to flag outliers from the dataset using the dissimilarity measures discussed in the previous section. Note that previous to drug intake *the distributions of the two groups are nearly identical*, according to standard Q-Q plots. If the people on drug tend to be flagged as outliers with a greater probability than expected, then a reasonable conclusion is that there may be a hepatotoxic effect resulting from drug intake.

We rank the subjects according to the approach presented in section 3.2.1. Once we have these outlier rankings for all the subjects in a given study, we can use them to determine not only which subjects are the most anomalous, but also to determine if the drug being studied has any appreciable effect. For example, if we examine a ranking of the subjects, we would expect the hepatotoxic patients to be highest-ranked, followed by the remaining subjects who were on the drug, and finally the patients who were given a placebo. However, a drug that has little or no effect on the liver tests is less likely to cause hepatotoxicity, and subjects on such a drug should not be very dissimilar from those on placebo, meaning that the ranking would be random. To examine the effects of the drug being studied, we use graphs such as those in Figure 1, where we plot the cumulative number of subjects on drug and on placebo given the outlier ranking using thick lines. The thin lines express the expected cumulative number of subjects on drug or placebo for a given ranking assuming the ranking is random. For more details see [10].

In Figure 1 (A) we plot the outlier ranking arising from both the  $D_e$  measure for the top 10% (113) of the outliers in  $D1$ . We observe that the expected number of

outliers from the drug group is exceeded by the actual number indicating a clear signal that the drug under question is causing a change in analyte behavior in the patients being flagged as outliers. We would like to note that Phase III continued for approximately two more years after these cases were completed. Had this signal been detected at that time, Pfizer might have been able to save on the resources it expended to continue Phase III.

In our second experiment we evaluate the performance of our method on the second dataset composed of healthy post-menopausal women. In Figure 1 (B) and (C), we plot the top 10% of the outliers for both drugs using the  $D_e$  measure. As expected, since these are healthy women taking either a placebo or drugs with no known hepatotoxic effects, the mixture of subjects on drug and placebo marked as outliers are near the expected levels.

These experiments demonstrate an advantage of our approach over Hy's rule. They show that we are capable on not only finding important differences in magnitude, but also in direction (correlation) that may be missed by Hy's rule.

**4.3 Data Set Clustering** In our final experiment we demonstrate the utility of using our dissimilarity measures to perform clustering. In this case we use a subset of the subjects corresponding to all males in dataset  $D1$  with diabetes who were taking drug  $A$ , for a total of 450 subjects. We use the Euclidean dissimilarity measure  $D_e$  and the covariance matrices and performing single-link hierarchical clustering. We find that clustering results in an intuitive grouping of the subjects.

One branch of the resulting cluster dendrogram corresponds to a cluster of subjects with relatively small spikes in analyte values, as can be seen in Figure 2 (A). Another branch contained subjects with spikes an order of magnitude larger, such as the subject seen in Figure 2 (B). These spikes may not be large enough to be considered a sign of hepatotoxicity according to Hy's rule. Still another branch contained subjects with spikes two orders of magnitude larger than those in the first branch. low spikes in analyte values, such as the one in Figure 2 (C). Although we only plot the ALT levels here, we note these spikes extend to the other blood analytes as well and affect the overall covariances. Many of the other branches contain subjects with only small random fluctuations in analyte values. Such fluctuations are not indicative of hepatotoxicity, but the subjects may cluster together due to demographic or other health attributes, which may aid in determining dosage levels.

## 5 Conclusion

Efficient and precise analysis of clinical trial data is very important to pharmaceutical companies, as it allows

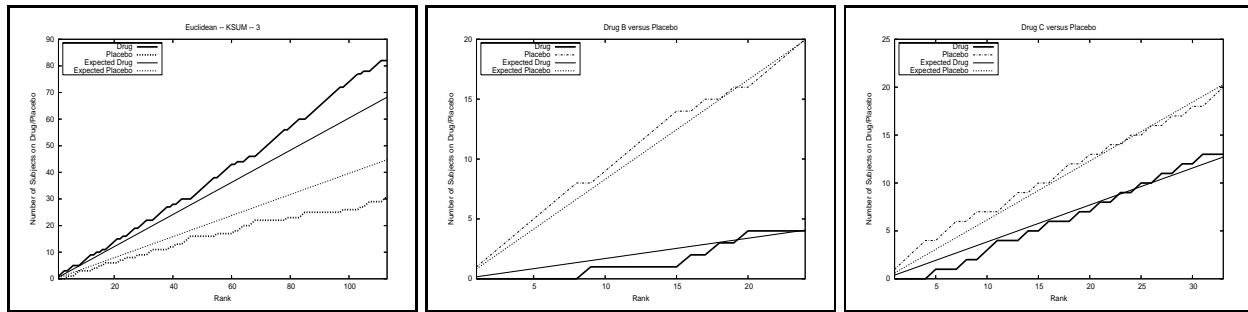


Figure 1: Outlier rankings using (A)  $D_e$  on Drug A in  $D_1$ ; (B)  $D_e$  on Drug B in  $D_2$ ; (C)  $D_e$  on Drug C in  $D_2$ .

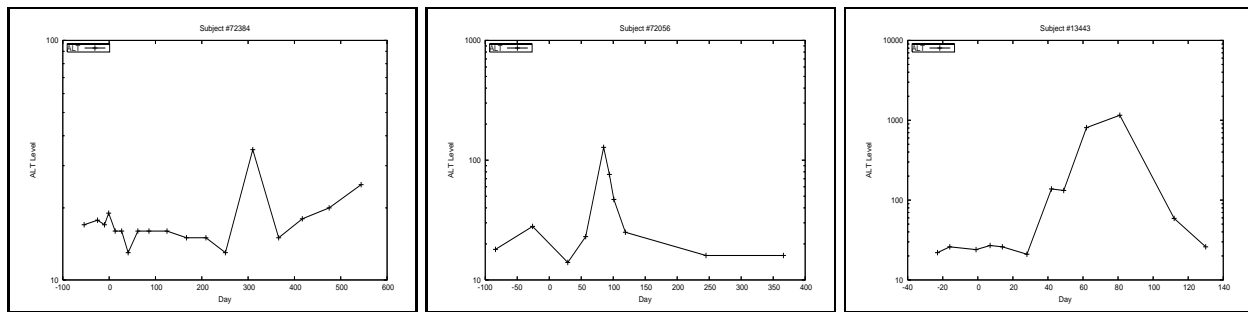


Figure 2: Subject with (A) a small spike, (B) a medium spike, and (C) a large spike in analyte values.

them to determine the efficacy and safety of a drug. Current approaches for detecting hepatotoxicity in clinical trial data sets have limited effectiveness, since they typically ignore correlations between blood analytes. In this paper we presented several dissimilarity measures for data sets that takes into account the means and covariance structures of the data sets. Our results on real clinical trial data show that our measures can be very helpful in detecting true hepatotoxicity and finding subpopulations of subjects who may have different reactions to the drug under study.

## References

- [1] Francesco Audrino and Peter Buhlmann. Synchronizing multivariate financial time series. Technical Report Research Report 97, Seminar fur Statistik, May 2001.
- [2] Einar Bjornsson and Rolf Olsson. Outcome and prognostic markers in severe drug-induced liver disease. *Hepatology*, 42(2):481–489, August 2005.
- [3] Tak chung Fu, Fu lai Chung, Robert Luk, and Chak man Ng. Financial time series indexing based on low resolution clustering. In *Proceedings of the IEEE International Conference on Data Mining Workshop on Temporal Data Mining: Algorithms, Theory and Applications*, November 2004.
- [4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.
- [5] Emre Erdogan, Sheng Ma, Alina Beygelzimer, and Irina Rish. Statistical models for unequally spaced time series. In *SIAM*, 2004.
- [6] D. P. Huttenlocher, G. A. Klanderman, and W. A. Rucklidge. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 15(9):850–863, 1993.
- [7] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.
- [8] E. Knorr and R. Ng. A unified notion of outliers: Properties and computation. In *ACM SIGKDD*, 1997.
- [9] Matthew Eric Otey and Srinivasan Parthasarathy. A dissimilarity measure for comparing subsets of data: Application to multivariate time series. In *TDM*, 2005.
- [10] Matthew Eric Otey, Srinivasan Parthasarathy, and Donald C. Trost. Dissimilarity measures for detecting hepatotoxicity in clinical trial data. Technical report, Department of Computer Science and Engineering, The Ohio State University, 2006.
- [11] Srinivasan Parthasarathy and C. C. Aggarwal. On the use of conceptual reconstruction for mining massively incomplete data sets. *IEEE Transactions on Knowledge and Data Engineering*.
- [12] Srinivasan Parthasarathy and Mitsunori Ogihara. Clustering distributed homogeneous datasets. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 566–574, London, UK, 2000. Springer-Verlag.
- [13] Jose C. Principe, Neil R. Euliano, and W. Curt Lefebvre. *Neural and Adaptive Systems: Fundamentals through Simulations*. John Wiley and Sons, 2000.
- [14] Richard Reymont and K. G. Joreskog. *Applied Factor Analysis in the Natural Sciences*. Cambridge University Press, 1996.