Quantification of Impact of New Cancer Therapy

Nicola Egües

October 30, 2024

Abstract

This project evaluates the effectiveness of a new cancer drug developed by TheraTech by analysing biomarker data from 2000 samples before and after treatment. Applying Principal Component Analysis and Clustering on the post-treatment data reveals two subgroups, indicating two distinct drug-responses. One cluster can hereby be identified as representing symptom improvement, with 71.55% of patients falling into this category, thus highlighting the effectiveness of this treatment. Further, markers 0, 1 and 2 were found to be especially responsive to the therapy.

1 Introduction

Cancer biomarkers are measurable factors, such as genes or proteins, that can reveal critical information about the progression of cancer in the body. As such, they are essential not only in predicting patient outcomes but also in evaluating the effectiveness of new cancer therapies [1].

The pharmaceutical company TheraTech has recently developed a drug believed to constitute such an effective new cancer therapy and to reflect on a variety of biomarkers.

To evaluate the drug's effectiveness, this project conducts an in-depth analysis of the cancer biomarker data before and after the treatment with this drug, aiming to quantify how many patients presented an improvement in their symptoms as a consequence of it.

In parallel, this analysis aims to provide valuable insights into how these biomarkers interact and reflect patient outcomes as well as determining which markers the drug has the highest impact on.

2 Background Theory

2.1 PCA

2.1.1 Theory

Principal Component Analysis (PCA) is an unsupervised learning method, meaning it analyses data without relying on labeled outcomes. It is most commonly used to obtain a low-dimensional representation of the original data that retains as much information as possible. By projecting the data onto reduced (and thus visualisable), variance-capturing dimensions, PCA can reveal underlying patterns that would have been more difficult to identify in the high-dimensional space [2].

PCA achieves this by creating new dimensions (principal components) that are linear combinations of the original features X (eq. 1). The weights of the first component Z_1, deemed "loadings", are hereby chosen to be such that Z_1 has the largest possible variance (under the condition that the sum of squares of these weights equals 1 so that not arbitrarily large numbers can be chosen for an arbitrarily large variance). Each loading thus represents how much that feature contributes to a given principal component. Additionally, the loadings of the first component can be geometrically interpreted as forming a vector that defines the direction in the feature space along which the variation of the data is maximised. Projecting the data points () leads to the transformed values, the principal component *scores* Z_1 [3].

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p \tag{1}$$

where

$$\sum_{j=1}^{p} \phi_{j1}^2 = 1 \tag{2}$$

The second principal component is then selected to be the linear combination that maximises the remaining variance under the constraint of being *uncorrelated* to the first principal component, and so on for all remaining components [3].

Performing PCA returns as many principal components as there are features in the data, whereby the first few typically explain a large portion of the variance and the subsequent components progressively less. With dimensionality reduction, the aim is to retain as few dimensions as possible without losing too much information. Consequently, one helpful way of figuring out how many components to keep is to set a certain lower bound for the desired percent explained variance and figure out how many components it takes to jointly surpass it [3].

To apply PCA, the features must first be scaled to have a standard deviation of one and a mean of zero. This is because if a few features have very large variance compared to others (e.g. from being measured in different units) - pca will focus on it. Centering the data on the means merely ensures that it is centered on the origin of the principal components [3].

2.2 Clustering

Another powerful unsupervised technique is Clustering, which seeks to group similar data points into distinct clusters. One common method that will be applied here is the Gaussian Mixture Model (GMM), which assumes that the data is composed of several Gaussian distributions (i.e., a weighted sum of Gaussian components) - one representing each cluster. The parameters that characterise each Gaussian component (mean, covariance and weight) are found via the Expectation Maximisation (EM) algorithm. In short, this algorithm works by iteratively updating the parameters of the Gaussian components to maximize the likelihood of observing the data under the present GMM model [4].

In contrast to the K-means algorithm, which is restricted to placing circles (or spheres) in the center of each cluster, GMM can work with clusters of different shapes and sizes and is thus a lot more flexible. Further, it is especially applicable to cases of slight overlap between clusters due to its probabilistic (or "soft") approach in assigning the data points.

2.3 Significance testing

Significance testing is a statistical method that applies hypothesis testing to determine whether there exist meaningful differences between groups.

The Null Hypothesis hereby states that there are *no* significant differences between the groups being compared. The p-value represents the probability of obtaining the observed results if the Null hypothesis is true. A p-value of (typically) 0.05 or lower indicates that one can reject the null hypothesis, meaning that there *is* in fact a statistically significant difference between the groups [5].

The Mann-Whitney U test is a non-parametric test used to compare two groups when the data cannot be assumed to be normally distributed, as is required for the ANOVA test, and in short, works by analysing the rankings of the joined data points and comparing them between the two groups (from which a certain value, the U-statistic, is retrieved) [5].

3 Experimental Methods

The data used in this project includes numerical values from six different markers for 2000 patients, recorded before and after treatment with the drug. Importantly, it is stated that the data has been normalised with appropriate controls, ensuring that the effect of the drug is isolated.

3.1 Required assumptions

The company seeks to quantify the amount of people with improved symptoms after treatment with the drug. Since the post-drug data is unlabeled, this requires the application of an unsupervised technique on the post-drug data that could ideally separate it into groups of ”improved symptoms” and ”no improved symptoms”, based on the knowledge of which cluster a known ”improved symptoms” sample belongs to.

Given the information that the effect of the drug was successfully isolated, any uncovered subgroups in the post-drug data can thus be said to reflect two distinct responses to the administered drug. However, to reliably identify any uncovered subgroups as being those of improved symptoms and not improved symptoms, a key assumption must be made: namely, that there are some patients with improved symptoms after drug treatment and a significant portion without, and the marker values differ significantly between these groups - and more so than any other sources of variability in the data.

This is because if all or almost all patients showed improved symptoms, then any revealed distinct clusters might indirectly reflect other biological differences rather than symptom severity. One example could hereby be males and females responding differently to the drug despite both improving, or the subgroups simply reflecting two distinct levels of improvement due to underlying factors such as genetic predispositions.

These scenarios could also arise in the case where both groups of improved and not improved symptoms are present, but the marker values between these groups are not distinct enough to present any significant patterns. If clusters were to be found in this scenario, then they too would likely reflect a case where the treatment uncovered a different underlying factor.

3.2 Overall procedure

The analysis is conducted in the following steps: First, the marker distributionn for both the pre-drug and post-drug datasets, as well as correlations between the post-drug marker data are visualised.

Though potential outliers can be extracted, they aren’t handled in this case as this may require more domain knowledge and could risk losing valuable information.

Then, PCA is applied on the unlabelled data, aiming to reduce the dimensionality and uncover patterns in the post-drug data, as well as to produce valuable insights about the markers’ responses via the loadings. The Gaussian Mixture Model clustering algorithm is then applied on the clusters found by PCA to assign patients to these uncovered subgroups.

Finally, to quantify which of the markers the drug has the most significant impact on, significance testing was done on the ”improved symptoms” versus not improved symptoms groups for each marker within the post-drug data.

4 Analysis and Discussion

4.1 Data visualisations

Figure 1 illustrates the distribution of marker values before and after treatment with the drug. While all pre-drug marker values consistently follow a normal distribution, the post-drug marker distributions are a lot more varied.

Specifically, the marker values for 0, 1 and 2 clearly deviate from a Gaussian distribution as marker 0 is skewed to the right, and markers 1 and 2 exhibit a bimodal distribution with a lower right peak. Post-drug values for markers 3, 4 and 5 do approximate a normal distribution, but present differing medians and variability than their pre-drug values. This is more pronounced for markers 3 and 4, while the post-drug distribution for marker 5 appears relatively similar to its pre-drug equivalent. The boxplots present in the accompanying dashboard provide an additional visual comparison of the medians and spreads of pre- versus post-drug marker data.

The bimodal distributions of markers 1 and 2 show how two subgroups are present within their post-drug data, reflecting two distinct responses to the administered drug -

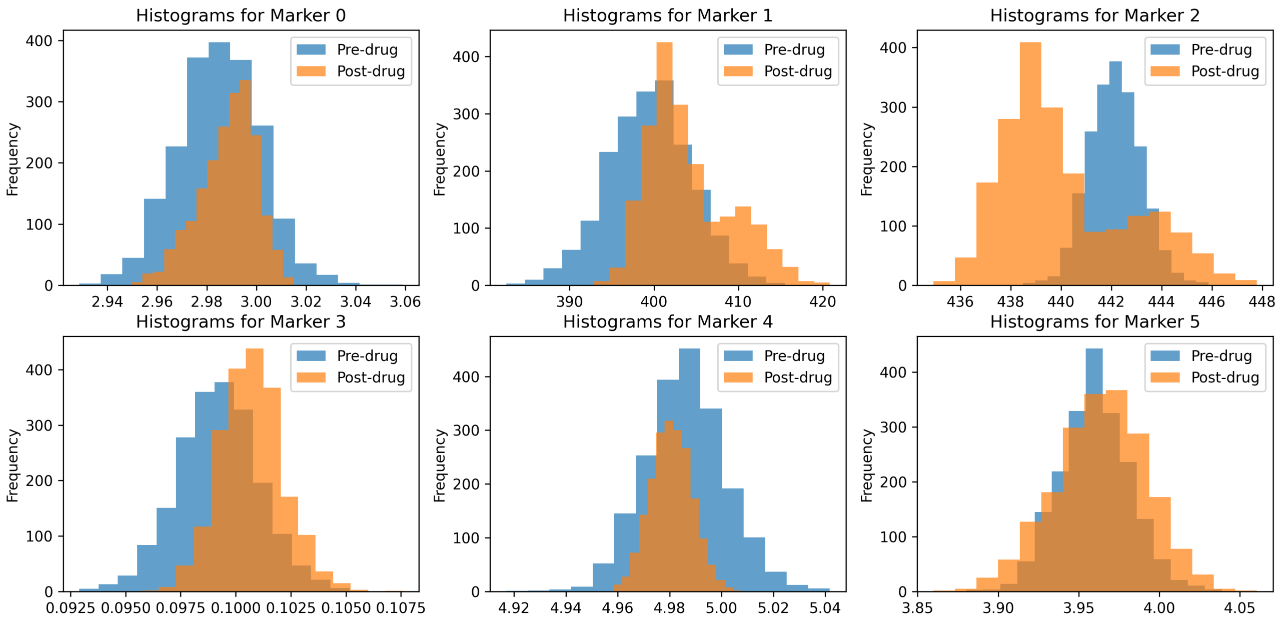

Figure 1: Histograms of pre-drug and post-drug data for each marker.

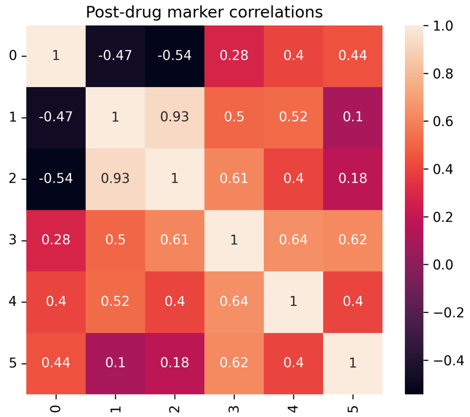

Figure 2: Correlations between the post-drug marker data.

which, under the assumption that "improved symptom" marker values are significantly different from "not improved symptoms values", most likely represent these two groups.

One could thus take skewness and/or present bimodality to be a good indicator of how well this marker will separate the two subgroups. From the histograms alone, however, one would not be able to extract the groups with good accuracy - a more advanced technique such as PCA is necessary.

4.1.1 Correlations analysis

Plotting the correlations of the markers from the post-drug data (fig. 2) shows how marker 1 and 2 are very strongly correlated to each other. This suggests that they may be linked to each other via the same underlying biological mechanism affected by the drug. It also suggests that one of these markers may be redundant and could potentially be dropped by the company.

Markers 1 and 2 are also moderately positively correlated to markers 3 and 4, indicating that they probably share some overlap in their response to the drug, and are moderately anticorrelated to marker 0, suggesting that they have an opposing biological response to the drug. Finally, the low correlation with marker 5 most likely points to them being involved in a process that is independent of the process underlying marker 5.

Markers 3 and 4 have a moderate positive correlation to all markers, while marker 5 has a moderate correlation to all markers except 1 and 2, where it is much lower.

4.2 PCA

Plotting the percent explained variance against the components as well as the cumulative sum for the *post-drug* marker data in figure 4, shows how the first two components would be needed to explain at least 75% of the variance in the data. In fact, by jointly capturing 84% of the variance, we can expect them to provide a reliable summary of the underlying patterns in the data without too much information being lost. Performing PCA on the *pre-drug* data finds that the first two components explain about 86% of the variance and are thus also very reliable, whereas this number decreases to 73% when PCA is done on both pre-and post-drug data combined. While this is still a relatively high number, it points to a more careful interpretation of any combined-data results.

Figure 3 shows the projections of the original data points onto the first two principal components found from doing PCA on the pre-drug data, the post-drug data and both

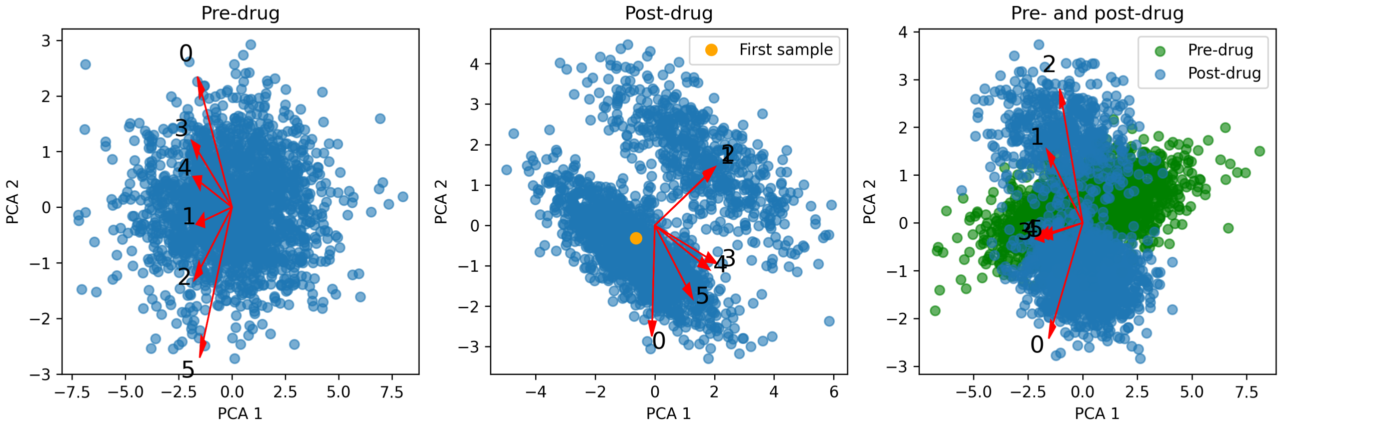

Figure 3: PCA biplot showing the projections of the data onto the first principal component (x-axis) and the second principal component (y-axis) in blue/green, as well as the loadings for each marker as red vectors, for the pre-drug marker data (left), the post-drug data (middle), and the pre-and post-drug data combined (right).

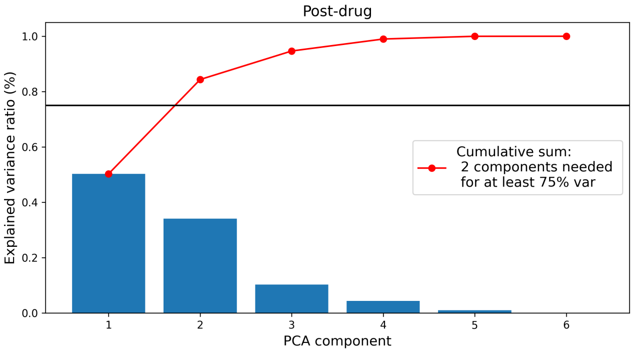

Figure 4: Explained variance ratios from the post-drug PCA plotted against the principal components, with their cumulative sum shown in red.

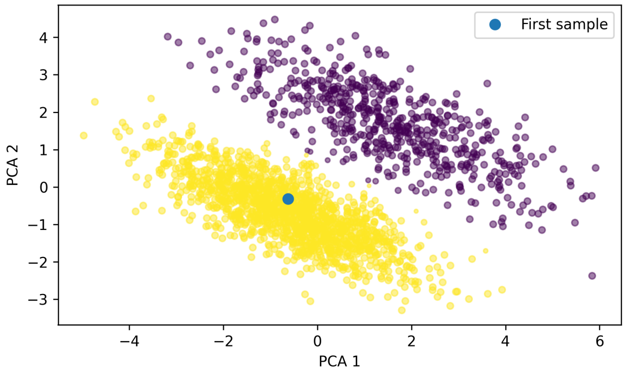

Figure 5: The resulting cluster assignments after the application of the Gaussian Mixture Model.

datasets combined. As expected, the pre-drug data only forms one big cluster, indicating no distinguishing factors in the marker data. This changes with the administration of the drug, however, which leads to the formation of two clearly distinct clusters.

Under the assumption laid out in section 3.1 that marker values between patient groups of improved symptoms and not improved symptoms are significantly different, one can conclude that these are the groups that the clusters most likely represent. Further, given the information that the first sample is known to present a case of improved symptoms, it can be inferred that the lower cluster (which the first sample falls into), represents the subgroup of improved symptoms.

Interestingly, the projection plot of both datasets combined shows the formation of three clusters, representing the pre-drug data as well as the two clusters found within the post-drug data. Under the current assumptions, this result implies that the drug does not simply divide patients into ”improved” and unchanged” symptoms (as is normally the case). If the symptoms were unchanged, then one would have expected the marker value distributions for ”unchanged symptoms” to be the same as the pre-drug equivalents, and thus the PCA on the combined dataset to find only two clusters. This would suggest that these patients that did not improve instead underwent some other changes as a consequence of the drug.

Alternatively, this finding may also present a reason to question the current assumptions of what the clusters represent. Thus, to further increase the reliability of interpreting any clusters as indicating symptomatic responses to the drug, we suggest the following recommendations:

1) Further investigation into patient outcomes should be done such that instead of only having one sample known to have exhibited improved symptoms, there is a small amount of labeled data for *each* group - so, for example, five known ”improved symptoms” and five ”not improved symptoms” samples. If the samples from each group consistently fall into different clusters, a more confident conclusion can be drawn.

2) If possible, the final clusters found via the statistical methods applied in this paper should be cross-checked with domain knowledge to check whether, for example, the median biomarker values in the clusters (present in the dashboard) align with what is expected for groups that present symptom improvement (and that don’t).

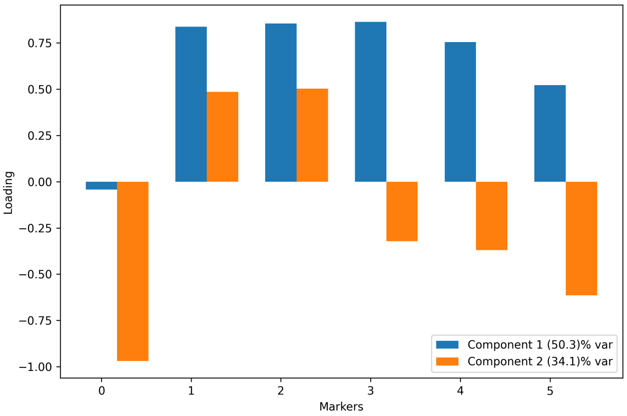

Figure 6: The loadings of the first two principal components of the post-drug data for each marker.

Continuing with the assumption that the lower cluster represents the group with improved symptoms, the Gaussian Mixture Model clustering algorithm can be applied to find out which samples are more likely to form a part of it, as seen in figure 5. This leads to the estimate that symptoms improved and thus the drug worked for about 71.55% of patients in the study.

4.2.1 Marker Loadings

The loadings reveal how much each marker contributes to each principal component. Figure 6, shows how in the post-drug data, markers 1, 2, 3 and 4 have strong positive loadings on the first principal component and marker 5 a moderately positive loading, indicating that high values for these markers will lead to high principal component 1 scores (i.e., high x-axis values in our scatter plots), and low values for these markers to a low PC1 score. Marker 0, with a weak negative loading, contributes very little to this component.

These conclusions can be drawn more rapidly from the red vectors depicted in the PCA scatter plots of figure 3, which represent these loadings of each marker. Their orientation describes how a given marker influences where a sample ends up in this PCA space. One can see, in the post-drug data, how higher values for marker 1 and 2, which by the proximity of the vectors are clearly correlated, will push a sample more into the cluster of no improved symptoms (with high PC1 and PC2 scores). Meanwhile the opposite is true for high values of marker 0, which will push a sample more into the improved symptom cluster (with strongly low PC2 scores and low PC1 scores).

To put the above abstract observations into a more concrete example, one could imagine a (very hypothetical) scenario where the first component represented inflammation, and the second component represented the presence of some beneficial protein. High inflammation and low protein would thus push a sample into the ”no improvements” cluster, while low inflammation and high good protein would do the opposite.

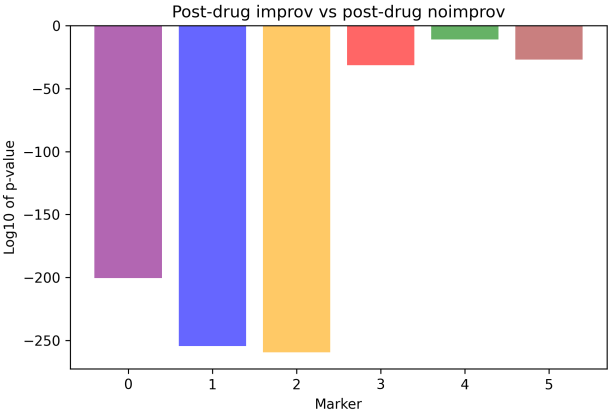

Figure 7: The log10 p-values for each marker after performing the Mann-Whitney U Test on the groups of improved and not-improved samples within the post-drug data.

4.2.2 Significance testing

To perform significance testing, the Mann-Whitney U test was applied, given that the post-drug data do not all follow normal distributions, as seen in figure 1. Performing this on the groups showing improved versus ”not improved” symptoms returns p-values significantly smaller than 0.05 for all markers, indicating that all six effectively distinguish between the two symptomatic groups. Markers 2, 1 and 0 (named in ascending p-value order) can hereby separate the two groups significantly better than markers 3, 4 and 5 (fig. 7, suggesting that the former three are more responsive to treatment with the drug.

5 Conclusions

In conclusion, principal component analysis was used to uncover two subgroups within the post-drug marker data, which reflect two distinct responses to the drug, suggesting a grouping of patients into those for which the symptoms improved and those for which it didn’t. This inference, however, relies on the assumption that marker values differ significantly between these two groups and more so than other potential sources of variability. Through Gaussian Mixture Model clustering and with the known label of the first sample, it was found that 71.55% of patients presented reduced cancer symptoms after being treated. Further, the analysis shows that markers 0, 1 and 2 proved particularly effective in distinguishing between the two subgroups, as shown by their PCA loading vectors and statistically significant differences identified through the Mann-Whitney U test.

References

[1] S. Qi et al. “High-resolution metabolomic biomarkers for lung cancer diagnosis and prognosis”. In: *Scientific Reports* 11 (2021), p. 11805. DOI: `10.1038/s41598-021-91276-2`.

[2] I. T. Jolliffe and J. Cadima. “Principal component analysis: a review and recent developments”. In: *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 374.2065 (2016), p. 20150202. DOI: `10.1098/rsta.2015.0202`.

[3] Gareth James et al. *Introduction to Statistical Learning with Applications in Python*. Springer, 2013.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. New York: Springer, 2006.

[5] S. Parab and S. Bhalerao. “Choosing Statistical Test”. In: *International Journal of Ayurveda Research* 1.3 (2010), pp. 187–191. DOI: `10.4103/0974-7788.72494`.