

# Inlier Analysis using Mahalanobis Distance in Central Monitoring

Abhyuday Chanda  
Biostatistician II

# Agenda

- Background
- Let's see what Google says
- An Initial example of Data
- Approaches
- Illustration
- Methodology
- Results
- Conclusion

# Background

For data verification, the conventional approach is Source Data Verification (SDV), which though commonly used, does have its own limitations in the efficiency in picking up the errors, with a disproportionately high cost.

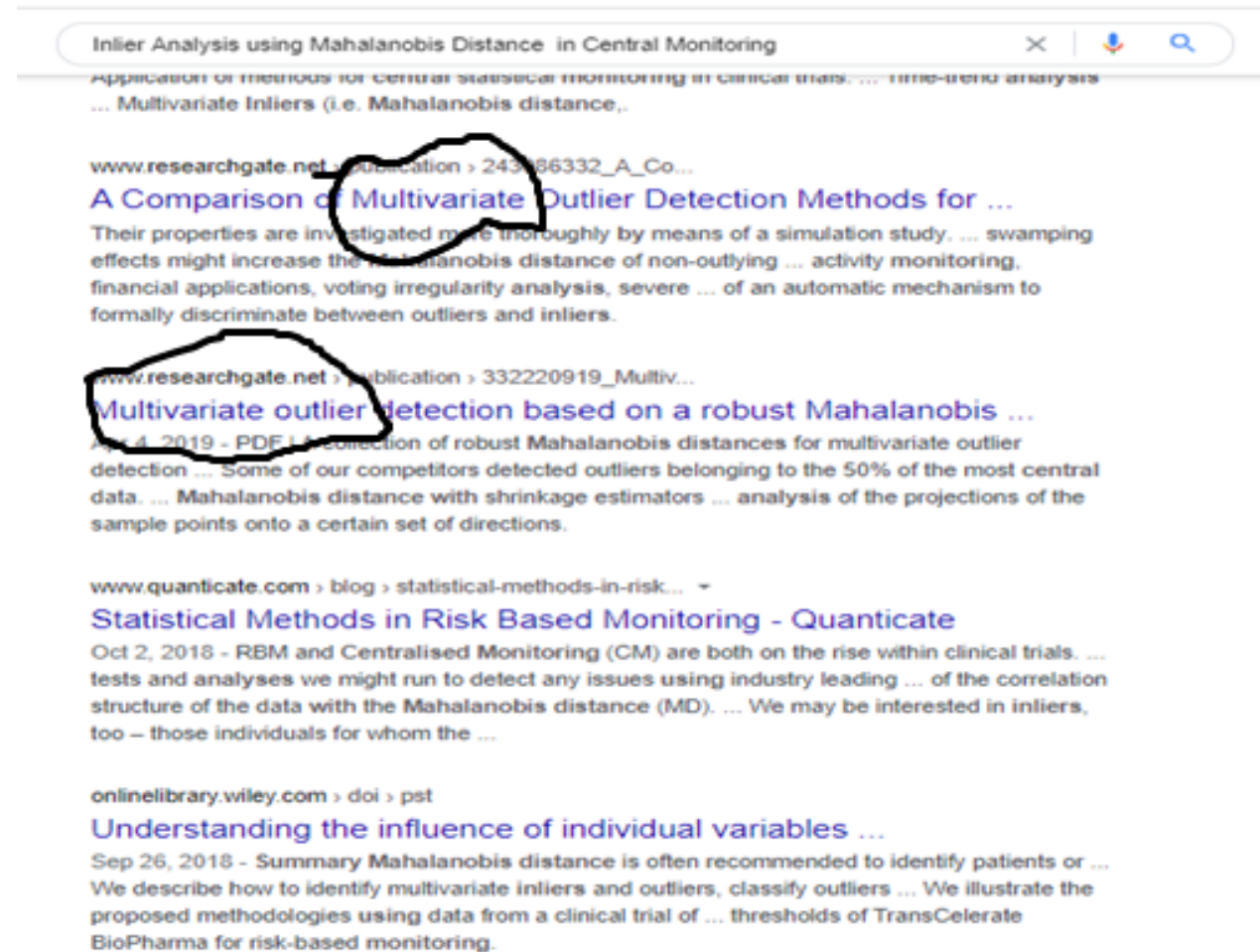
With Statistical Risk based Monitoring

- It is possible to identify possible frauds by giving a broader view of the data,
- increases chances of capturing the fake data and effectively improves the quality of it.
- It gives the sponsor, CRO, Site Management and Monitoring personals essential proofs, backed up by statistically derived methods, whether a site is in risk with respect to the possible fraudulent activities prevailing there.

# Let's see what Google says

Search in Google as 'Inlier Analysis using Mahalanobis Distance in Central Monitoring'

As one can see most of the results are mostly on the outlier approach to identify the fraud!!



The screenshot shows a Google search interface with the query "Inlier Analysis using Mahalanobis Distance in Central Monitoring" entered in the search bar. Below the search bar, the first result is from ResearchGate, titled "A Comparison of Multivariate Outlier Detection Methods for ...". The second result is also from ResearchGate, titled "Multivariate outlier detection based on a robust Mahalanobis ...". The third result is from Quanticate.com, titled "Statistical Methods in Risk Based Monitoring - Quanticate". The fourth result is from Wiley Online Library, titled "Understanding the influence of individual variables ...".

Inlier Analysis using Mahalanobis Distance in Central Monitoring

Application of methods for central statistical monitoring in clinical trials: ... time-series analysis ... Multivariate Inliers (i.e. Mahalanobis distance, ...

www.researchgate.net/publication/243186332\_A\_Comparison\_of\_Multivariate\_Outlier\_Detection\_Methods\_for\_...

**A Comparison of Multivariate Outlier Detection Methods for ...**

Their properties are investigated more thoroughly by means of a simulation study. ... swamping effects might increase the Mahalanobis distance of non-outlying ... activity monitoring, financial applications, voting irregularity analysis, severe ... of an automatic mechanism to formally discriminate between outliers and inliers.

www.researchgate.net/publication/332220919\_Multiv...

**Multivariate outlier detection based on a robust Mahalanobis ...**

Apr 4, 2019 - PDF | Correction of robust Mahalanobis distances for multivariate outlier detection ... Some of our competitors detected outliers belonging to the 50% of the most central data. ... Mahalanobis distance with shrinkage estimators ... analysis of the projections of the sample points onto a certain set of directions.

www.quanticate.com/blog/statistical-methods-in-risk...

**Statistical Methods in Risk Based Monitoring - Quanticate**

Oct 2, 2018 - RBM and Centralised Monitoring (CM) are both on the rise within clinical trials. ... tests and analyses we might run to detect any issues using industry leading ... of the correlation structure of the data with the Mahalanobis distance (MD). ... We may be interested in inliers, too – those individuals for whom the ...

onlinelibrary.wiley.com/doi/10.1002/pst

**Understanding the influence of individual variables ...**

Sep 26, 2018 - Summary Mahalanobis distance is often recommended to identify patients or ... We describe how to identify multivariate inliers and outliers, classify outliers ... We illustrate the proposed methodologies using data from a clinical trial of ... thresholds of TransCelerate BioPharma for risk-based monitoring.

# An Initial example of Data

| Site | Subject ID | Visit   | SBP | DBP | PR | RR | TEMP |
|------|------------|---------|-----|-----|----|----|------|
| 114  | 022        | Visit 2 | 120 | 74  | 70 | 20 | 98.4 |
| 120  | 043        | Visit 3 | 120 | 78  | 82 | 18 | 97.3 |
| 174  | 028        | Visit 4 | 105 | 86  | 81 | 16 | 98.6 |
| 174  | 025        | Visit 4 | 100 | 85  | 64 | 16 | 98.6 |
| 174  | 031        | Visit 4 | 105 | 86  | 80 | 16 | 98.6 |
| 174  | 036        | Visit 4 | 100 | 86  | 64 | 16 | 98.6 |
| 120  | 021        | Visit 3 | 120 | 78  | 82 | 18 | 97.3 |
| 104  | 052        | Visit 2 | 124 | 76  | 78 | 16 | 97   |
| 114  | 023        | Visit 2 | 120 | 74  | 70 | 20 | 98.4 |
| 104  | 052        | Visit 3 | 124 | 76  | 78 | 16 | 97.1 |

A data with an innocent display

A little bit of Sorting and a clearer picture emerges

| Site | Subject ID | Visit   | SBP | DBP | PR | RR | TEMP |
|------|------------|---------|-----|-----|----|----|------|
| 104  | 052        | Visit 2 | 124 | 76  | 78 | 16 | 97   |
| 104  | 052        | Visit 3 | 124 | 76  | 78 | 16 | 97.1 |
| 114  | 022        | Visit 2 | 120 | 74  | 70 | 20 | 98.4 |
| 114  | 023        | Visit 2 | 120 | 74  | 70 | 20 | 98.4 |
| 120  | 043        | Visit 3 | 120 | 78  | 82 | 18 | 97.3 |
| 120  | 021        | Visit 3 | 120 | 78  | 82 | 18 | 97.3 |
| 174  | 025        | Visit 4 | 100 | 85  | 64 | 16 | 98.6 |
| 174  | 036        | Visit 4 | 100 | 86  | 64 | 16 | 98.6 |
| 174  | 028        | Visit 4 | 105 | 86  | 81 | 16 | 98.6 |
| 174  | 031        | Visit 4 | 105 | 86  | 80 | 16 | 98.6 |

# The Approaches

Instead of the outlier approach, we concentrated on the Inlier approach in the following 4 ways:

1. Same Site , Same Visit, Different Subjects

Identifying the pairs of subjects in the same visit, belonging to same site having same/almost same values for a set of parameters

2. Same Site , Same date, Different Subjects

Identifying the pairs of subjects in the same date, belonging to same site having same/almost same values for a set of parameters

3. Same Subject, Different Visits

Identifying subjects having same/ almost same results at different visits

4. Site specific central tendency of parameters

The distribution of parameter values for too many subjects hovering around the central( mean ) value of it

The formula of Mahalanobis Distance is given below:

$$D^2 = (\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m})$$

where:

$D^2$  = Mahalanobis distance

$\mathbf{x}$  = Vector of data

$\mathbf{m}$  = Vector of mean values of independent variables

$\mathbf{C}^{-1}$  = Inverse Covariance matrix of independent variables

$\mathbf{T}$  = Indicates vector should be transposed

# Illustration

Same Site , Same  
Visit, Different  
Subjects

| site | subjectid | visit | Date/Time          | MD       | SBP | DBP | HR |
|------|-----------|-------|--------------------|----------|-----|-----|----|
| 163  | 004       | 2_1_6 | 11/1/2019<br>11:40 | 0        | 134 | 84  | 79 |
| 163  | 011       | 2_1_6 | 12/1/2019<br>14:54 | 0        | 134 | 84  | 79 |
| 508  | 003       | 5_2_3 | 12/1/2019<br>11:20 | 0.014077 | 118 | 68  | 60 |
| 508  | 005       | 5_2_3 | 1/1/2020<br>11:40  | 0.014077 | 119 | 68  | 60 |
| 305  | 003       | 2_2_3 | 8/1/2019<br>12:06  | 0.023308 | 120 | 80  | 65 |
| 305  | 014       | 2_2_3 | 5/1/2020<br>11:25  | 0.023308 | 120 | 80  | 66 |

Same Site ,  
Same Date,  
Different  
Subjects

| Site | subjectid | Visit                    | visitdate | vstm  | MD       | SBP | DBP | HR |
|------|-----------|--------------------------|-----------|-------|----------|-----|-----|----|
| 297  | 297187    | Visit 3 Day 29           | 17-Feb-20 | 07:21 | 0        | 120 | 80  | 72 |
| 297  | 297224    | Screening Visit 1b       | 17-Feb-20 | 09:18 | 0        | 120 | 80  | 72 |
| 297  | 297193    | Screening Visit 1b       | 14-Jan-20 | 08:28 | 0.028802 | 120 | 80  | 69 |
| 297  | 297198    | Screening Visit 1b       | 14-Jan-20 | 09:20 | 0.028802 | 120 | 80  | 68 |
| 234  | 234115    | Visit 3 Day 29           | 5-Nov-19  | 09:07 | 0.04349  | 128 | 76  | 83 |
| 234  | 234128    | Visit 2<br>Randomization | 5-Nov-19  | 08:11 | 0.04349  | 128 | 76  | 82 |

Same  
Subject ,  
Different  
Visit

| site | subjectid | visit | Date/Time      | md       | SBP | DBP | HR |
|------|-----------|-------|----------------|----------|-----|-----|----|
| 100  | 010       | 3_1_2 | 11/1/2019 9:41 | 0        | 124 | 71  | 78 |
| 100  | 010       | 3_2_1 | 11/1/2019 9:39 | 0        | 124 | 71  | 78 |
| 103  | 006       | 1_2_1 | 4/1/2019 13:12 | 0.004814 | 122 | 84  | 68 |
| 103  | 006       | 8_1_2 | 8/1/2019 12:11 | 0.004814 | 123 | 84  | 68 |
| 305  | 011       | 2_1_4 | 2/1/2020 12:39 | 0.023308 | 120 | 80  | 70 |
| 305  | 011       | 5_2_1 | 5/1/2020 13:35 | 0.023308 | 120 | 80  | 71 |

# Site Specific Central Tendency of Parameters

| site | subjectid | visit | Date/Time          | sbp | dbp | hr | md_mean  | m_sbp  | m_dbp | m_hr  |
|------|-----------|-------|--------------------|-----|-----|----|----------|--------|-------|-------|
| 103  | 008       | 3_2_1 | 11/1/2019<br>11:50 | 128 | 78  | 74 | 0.209219 | 131.37 | 78.18 | 72.88 |
| 111  | 004       | 8_2_1 | 2/1/2020<br>10:27  | 124 | 74  | 78 | 0.136778 | 124.33 | 73.58 | 76.88 |
| 121  | 009       | 1_2_1 | 10/1/2019<br>10:35 | 124 | 76  | 76 | 0.158155 | 125.62 | 77.80 | 76.23 |

# Methodology

- Calculate the Mahalanobis Distance(MD) between pairs of subjects as applicable to each of the approaches.
- Calculate  $MD^2$  ( as it empirically follows the chi-square distribution). Number of parameters for the calculation of the MD will be considered as the degrees of freedom.
- Using the DF at 0.025 and 0.05 we derive the chi square statistic(x) value to use as threshold.

| df | A=0.005  | 0.01    | 0.025   | 0.05   |
|----|----------|---------|---------|--------|
| 1  | 3.90E-05 | 0.00016 | 0.00098 | 0.0039 |
| 2  | 0.01     | 0.0201  | 0.0506  | 0.103  |
| 3  | 0.0717   | 0.115   | 0.216   | 0.352  |
| 4  | 0.207    | 0.297   | 0.484   | 0.711  |

For example: if DF=3 then the lower threshold will be 0.216, corresponding to A=0.025 and upper threshold will be defined as 0.352, corresponding to A=0.05

# Methodology

- Flag the values as below(if DF=3):
  - If  $MD^2 = 0$  then flag it as “Zero”
  - Else If  $MD^2 \leq 0.216$  flag it as “Low”
  - Else if  $MD^2 > 0.352$  flag it as “High”
  - Else flag it as “Medium”
- For each site count the number of “Zero”, “Low”, “Medium” and ”High counts.
- The final risk for each site is derived as follows:

Risk score=  $(Zero * w1 + Low * w2 + Medium * w3 + High * w4) / ((Zero + Low + Medium + High) * w1)$

with  $w1$ ,  $w2$ ,  $w3$  and  $w4$  are the associated weights

$w1 > w2 > w3 > w4$ ;

$w1 + w2 + w3 + w4 = 1$ .

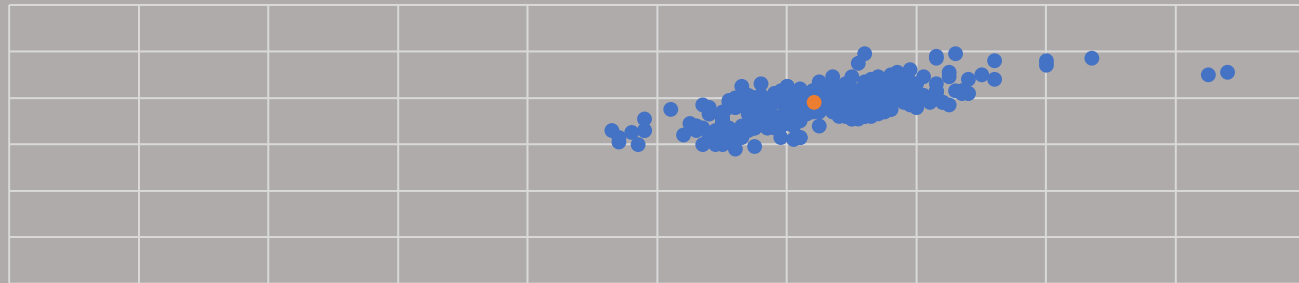
# Methodology

The weights were derived using rank exponent method. Since there is no thumbs rule available at this moment, the weight of the exponent is kept as 3,4,5 in the below formula

$$w_j(RE) = \frac{(n - r_j + 1)^p}{\sum_{k=1}^n (n - r_k + 1)^p}$$

In case of “*Site Specific Central Tendency of Parameters*” the “Zero” category is not used as mean more often will come as a non-integer, followed up with no comparison yielding the MD as “0” leading to have unused weightage, which may disrupt the clarity in the picture for the sites.

An example of the “Site Specific Central Tendency of Parameters”



# Methodology

As per the conventional categorization a site can be flagged as following:

Red=High risk, often requires immediate action

Amber= Medium risk, requires constant monitoring

Green= Low risk, best performing sites

Using the final risk scores thresholds can be derived using

- Mean  $\pm$  SD
- Median  $\pm$  SD
- Q1 ~ Q3

and based upon the risk score and the selected threshold the sites will be flagged as either Red, Amber or Green.

Choice of threshold requires detailed discussion between the Central Monitoring and Statistics team to understand how much leniency or stringency is expected to flag a site, for that phase of the study.

# Results

Using the formula Risk score= (Zero\*w1+ Low\*w2+ Medium\*w3+ High\*w4)/((Zero+ Low+ Medium+ High)\*w1) with different weight(3,4,5) in the formula we get the result as follows:

| site | h    | hper | m  | mper | l  | lper | z  | zper | rbm3  | score3   | rbm4  | score4   | rbm5  | finscore5 |
|------|------|------|----|------|----|------|----|------|-------|----------|-------|----------|-------|-----------|
| 305  | 1118 | 91.9 | 0  | 0    | 86 | 7.1  | 12 | 1    | RED   | 0.054071 | RED   | 0.035837 | RED   | 0.027549  |
| 116  | 33   | 91.7 | 0  | 0    | 3  | 8.3  | 0  | 0    | RED   | 0.049479 | RED   | 0.029948 | RED   | 0.020671  |
| 125  | 31   | 86.1 | 3  | 8.3  | 2  | 5.6  | 0  | 0    | RED   | 0.047309 | RED   | 0.02615  | RED   | 0.016629  |
| 600  | 579  | 97.8 | 7  | 1.2  | 5  | 0.8  | 1  | 0.2  | AMBER | 0.022012 | RED   | 0.008921 | RED   | 0.005018  |
| 121  | 1835 | 97.6 | 22 | 1.2  | 23 | 1.2  | 0  | 0    | AMBER | 0.021875 | RED   | 0.008415 | RED   | 0.004222  |
| 102  | 42   | 100  | 0  | 0    | 0  | 0    | 0  | 0    | AMBER | 0.015625 | GREEN | 0.003906 | GREEN | 0.000977  |
| 104  | 78   | 100  | 0  | 0    | 0  | 0    | 0  | 0    | AMBER | 0.015625 | GREEN | 0.003906 | GREEN | 0.000977  |
| 114  | 265  | 98.1 | 4  | 1.5  | 1  | 0.4  | 0  | 0    | AMBER | 0.01875  | AMBER | 0.005932 | AMBER | 0.0023    |
| 703  | 394  | 98.5 | 4  | 1    | 2  | 0.5  | 0  | 0    | AMBER | 0.01875  | AMBER | 0.006055 | AMBER | 0.002461  |
| 112  | 133  | 99.3 | 0  | 0    | 1  | 0.7  | 0  | 0    | AMBER | 0.018657 | AMBER | 0.006238 | AMBER | 0.00274   |
| 139  | 59   | 98.3 | 1  | 1.7  | 0  | 0    | 0  | 0    | AMBER | 0.017448 | AMBER | 0.004883 | AMBER | 0.001481  |

# Conclusion

Although there are many others way, but the inliers approach, discussed in this presentation will be helpful to the below points:

- It could help one understand that how sites could be identified & distinguished with probable fraudulent activities, with possible false data, hiding & wearing the mask of general values within the data itself.
- This could help the central monitors to flag and create action items to improve the data quality in the site and helps to get more of true data for the analysis purpose.

# Reference

1. Risk Based monitoring and Fraud Detection in Clinical Trials Using JMP and SAS by Richard Zink
2. Rank Ordering Criteria Weighting Methods – A Comparative Overview by Ewa Roszkowska
3. The Relationship between the Mahalanobis Distance and the Chi-Squared Distribution <https://markusthill.github.io/mahalanbis-chi-squared/>

# Q/A



Thank you for your patience