UPPSALA
UNIVERSITET

Finding Influential Individuals in Drug Trial Data

Gunnar Eggertsson, Peili Guo, Magnus Larsson

Project in Computational Science: Report

January 2019

PROJECT REPORT

Institutionen för informationsteknologi

Abstract

Currently, Case Deletion Diagnostics (CDD) is the standard method applied to identify in uential individuals in drug trial data. We describe a new, alternative method to identify in uential individuals, using data from the bootstrap re-sampling method, with the aim of eliminating the need to run CDD. Our method uses objective function values from bootstrap runs to produce an estimate of individuals' in uence. The accuracy of our method varies between substances but is around 85% on average. Results suggest that maximum accuracy is reached within the range of bootstrap runs typically performed within pharmacometric analysis.

Contents

1 Introduction

1.1 Background

Pharmacokinetics (PK) is an area within the eld of pharmacology that studies how a substance is a ected by the organism it is entered into, for example, in a response-time pro le model with clinical trial data. The fate of a substance depends on how it enters the organism, how it interacts with the biological compartments it passes and how it is metabolised.

In order to optimise drug development and therapy, di erent mathematical models are used to study, investigate and understand drug and disease mechanisms. Non-linear mixed e ects models are commonly used for the analyses of population pharmacokinetics. The pharmacokinetics model is non-linear for the concentration with respect to the model parameters. Mixed e ects refer to a mix of xed e ects and random e ects, which are statistical terms, referring to parameters that are xed or varied between individuals, respectively. This project is within the eld of population pharmacokinetics that studies the variability of drug concentrations within a population at clinical trials.

1.2 Aim of the project

Traditionally, in non-linear mixed e ect models, the methods Case Deletion Diagnostics (CDD) and bootstrap are used. CDD is run to identify in uential individuals and bootstrap to test the robustness of the underlying model. CDD is currently the standard method for nding in uential individuals. Its advantage is its accuracy but it is a computationally intensive method, especially with large data sets.

This leads us to experiment with alternative methods that focus on the bootstrap data to try to identify in uential individuals in clinical trial, with the aim of possibly eliminating the need to run the CDD method.

1.3 Influential individiuals

In uential individuals are the individuals with large impact on the model parameters. They are the outliers for the pharmacokinetic model and often are outside the bounds of normal variability and have signi cant in uence on the response-time pro le model where the concentration of the substance is signi cantly altered. Deviations in the kinetics of a substance can be caused by extreme weight, altered metabolism, other interfering pharmaceuticals or diseases.

1.4 Case Deletion Diagnostics (CDD)

Currently the golden standard method for nding in uential individuals is the CDD method. The idea behind it is to get a set of parameter estimations based on the full set of data and then remove the individuals one by one from the data set to perform parameter estimation and compare the di erence between the parameters based on the full set.

When evaluating the model, Cook's score and the covariance ratio are typically used as metrics for identifying in uential individuals [8]. Cook's score is commonly used to estimate the in uence of an observation when performing least-squares regression analysis. It takes into consideration of both residuals and leverage. The de nition of Cook's score is, as follows:

$$\sqrt{(P_i \quad P_{orig})^T Cov(P_{orig}) \ ^1(P_i \quad P_{orig})},$$

where P_{orig} and P_i are the estimated parameter vectors for the original run and the run with individual i removed respectively.

The detailed description for CDD is given below.

1. Compute the parameters (p) with all the observations in the data set.
2. Enumerate the observations in the data set.
3. For each observation i, delete that observation and compute the parameter $p^{(i)}$ based on the remaining observations.
4. Compute the difference between the parameters p and $p^{(i)}$.
5. Place back observation i to the data set and repeat the above steps 2-4 with observation $i + 1$.

Let θ be the maximum likelihood estimate of some parameter based on the complete data set and $\theta_{(i)}$ be the maximum likelihood estimate of the parameter without observation i. When deleting one observation, $\theta - \theta_{(i)}$ is computed. A large change in the maximum likelihood of $\theta - \theta_{(i)}$ for any parameter estimates indicates that observation i is an influential observation. The CDD method is time consuming because for a data set of size n, $n + 1$ maximum likelihood estimates need to be made.

1.5 Bootstrap

Bootstrap is a re-sampling method that can be used to estimate the robustness of a model by establishing a confidence interval for model predictions. The basic idea is to generate new data sets from the available individuals by using random selection with replacement [9]. In the resulting synthetic data set an individual can thus have zero or multiple occurrences. The model is then optimized for each data set to obtain measures of fit such as Objective Function Value (OFV).

A large number of generated data sets are needed to provide a good representation of the parameter distributions. The exact amount of required data sets depends on the underlying model. The confidence interval can then be estimated based on the results from running the bootstrap method.

2 Data and Methods

2.1 Data

The data used in our project are real patient data from clinical trials, provided by the project's supervisor. They consist of raw results from applications of the CDD- and bootstrap-methods for five different models, each corresponding to a different injected substance. In total, data are available for the following substances:

Albumin (Al). A type of plasma protein [4].

Digoxin (Di). Commonly used as a treatment for various heart conditions [5].

Nevirapine (Ne). Commonly used for treatment and prevention of HIV/AIDS [6].

Paclitaxel (Pa). Commonly used as a treatment for cancer [7].

Phenobarbital (Ph). Commonly used to treat seizures [3].

Table 1: Characteristics of the data sets

	Ph	**Al**	**Di**	**Ne**	**Pa**
Total number of individuals	59	5	227	58	66
Number of skipped individuals	59	5	225	53	45
Number of model parameters	6	7	9	23	21

2.1.1 CDD

The dimensions of the di erent data sets di er between the models. Table 1 shows the total number of individuals included in the data for each substance as well as the number of individuals that are skipped at some point in the CDD-data. The rst row in the table shows that the total population size in the data sets ranges from 5 (Albumin) to 227 (Digoxin). In the case of Albumin the size of the population is deemed too low for us to be able to obtain relevant results and we therefore focus on the other models in our study. The second row in the table shows that for Digoxin, Nevirapine and Paclitaxel some individuals included in the population are never skipped when the CDD-method is applied. We have found that those same individuals are also never included in the application of the bootstrap-method for the same medications. The reasons for those individuals not being skipped/included are unknown to us.

The models also di er in terms of the number of model parameters, which can be viewed as a crude estimate of the model complexities. The third row in the table shows that Nevirapine and Paclitaxel include more than twice as many model parameters as the other substances.

2.1.2 Bootstrap

The data sets contain raw results from 10000 runs of the bootstrap-method, for all the di erent substances. In the case of Phenobarbital and Albumin we also have raw results from 100 and 1000 runs but in general when we investigate di erent number of bootstrap runs we use permutations of the data sets containing the results from 10000 runs.

2.2 Method

For each of the di erent models we start by analyzing the CDD-results to identify the in uential individuals. They are identi ed as the individuals that, when skipped in the CDD-data, produce a Cook's score above 0.8. Figure 1 shows the objective function value plotted against the Cook's score for Phenobarbital. In this example, one individual is classi ed as in uential. Two individuals also produce a high Cook's score and high objective function value compared with the rest and have therefore got a high in uence despite not being classi ed as in uential.

Next we analyse the bootstrap-results alone with the aim of nding characteristics that identify the in uential individuals found in the rst step. For this purpose we investigate the distribution of the total objective function values produced in each bootstrap run and speci cally focus at the two ends of the distribution, labelled with A and B in Figure 2. Figure 2 shows the distribution of the total objective function values from the bootstrap data for Phenobarbital with the vertical lines showing example cuts that distinguish the two ends of the distribution. The next step in our method is to look closer at the bootstrap runs that produce total objective function values in the two ends of the distribution (A and B). We are interested in computing each individual's relative frequency of inclusion in these runs to investigate if some individuals' presence drives the total objective function value towards either high or low objective function

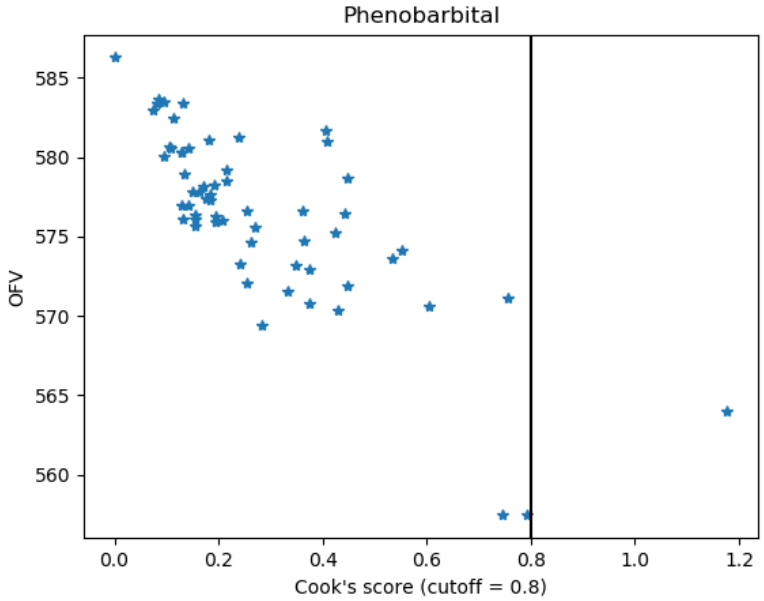

Figure 1: Objective function value plotted against Cook's score.

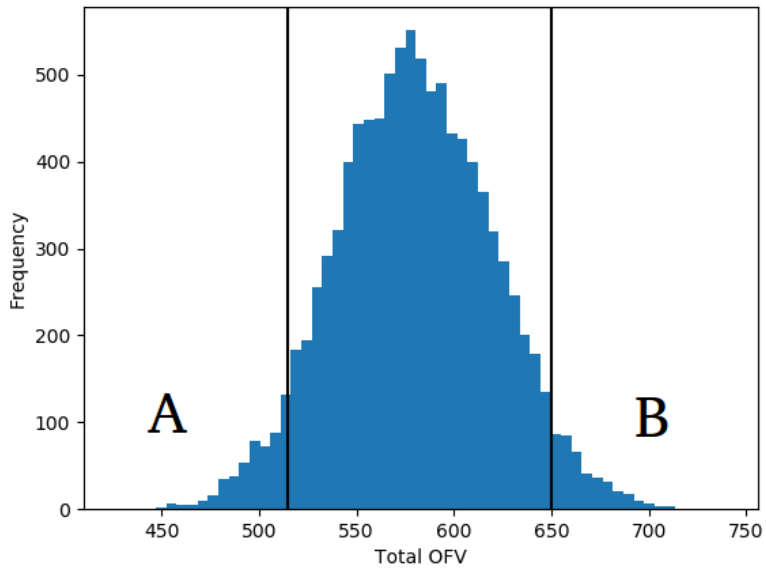

Figure 2: Distribution of total objective function values from data for Phenobarbital.

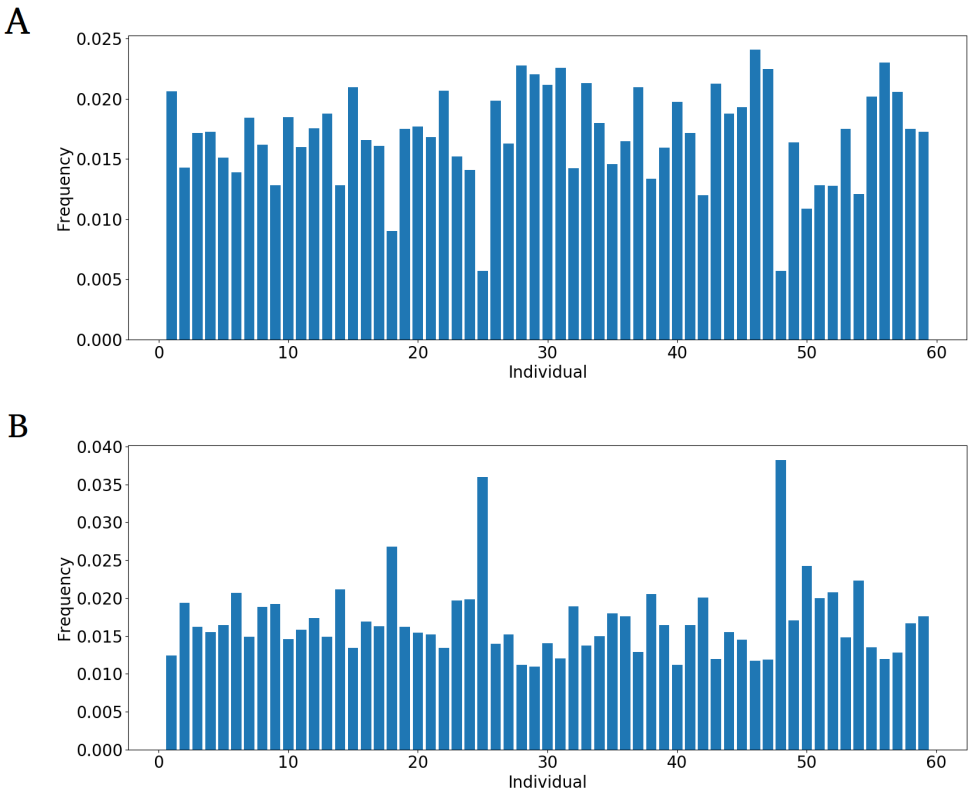

Figure 3: Relative frequency of inclusion for each individual in the two ends of the distribution.

values. In other words, the aim of this step is to analyse how much each individual contributes to the total objective function values in the two ends (A and B). Figure 3 shows the relative frequency of inclusion for each individual in the two ends. By visual inspection we can see that three individuals appear to be highly frequent in the high end and infrequent in the low end. These are individuals nr. 18, 25 and 48; the same individuals that were found to have the highest in uence in the analysis of the CDD-data.

Based on this positive initial result we are motivated to introduce the di erence between each individuals frequency in the high- and low end as an estimate of in uence. In other words we subtract each individual's frequency in the lower end from its frequency in the higher end. Figure 4 shows a histogram for the di erences between the two frequencies shown in Figure 3. Here, the most in uential individuals stand out as the highest values.

The nal step in our method is an optimisation process that aims at maximising the method's accuracy. We start this process by determining optimal cuto values for the total objective function value distribution from the bootstrap data (Figure 2) and the eventual di erence distribution (Figure 4), described in previous steps. As a measure of accuracy we use the following metric:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Population Size}}.$$

The accuracy is evaluated at various di erent cuto s. This is done by running our method for a wide range of cuto values in the total objective function value distribution and for each such, running the method for a wide range of cuto values in the eventual di erence distribution. Based on the obtained accuracies, optimal values for the cuto s are determined.

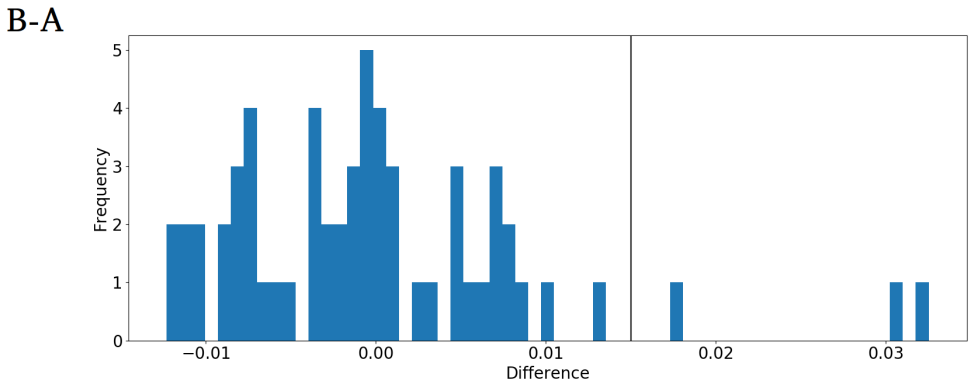

Figure 4: Histogram of relative frequencies in A subtracted from relative frequencies in B.

Finally we investigate the dependence of the accuracy on the number of bootstrap runs. We do this by evaluating the accuracy for a successive increase in the number of bootstrap runs. Our aim is to get an estimate of how many bootstrap runs are required to obtain acceptable results. This has signi cance because ideally we want to reach acceptable accuracy within the number of bootstrap runs traditionally performed within pharmacometric analysis.

3 Results

3.1 Identification of influential individuals - CDD

Initial evaluation by the CDD method is performed for four substances to establish in uential individuals based on a Cook's score cuto of 0.8. These individuals are treated as real in uential individuals and serve as a reference in the evaluation of our method. Figure 5 shows the objective function value plotted against the Cook's score for all four substances. The gure shows that the number of in uential individuals varies considerably between the substances.

In the case of Phenobarbital there is one individual who is classi ed as in uential, but also two other individuals who have a really high in uence compared with the rest and are very close to being classi ed as in uential. The data set with the highest number of individuals, Digoxin, does not include any in uential individuals. Nevirapine, on the other hand, includes more than twenty in uential individuals while Paclitaxel includes around ten.

3.2 Identification of influential individuals - Bootstrap populations

To optimise the new method, optimal cuto s to generate bootstrap sets A and B from the bootstrap distribution of OFV values are identi ed. Each cuto potentially generates an alternative distribution of relative di erence values (B-A) which also have an optimal cuto that produces the highest accuracy.

As described in detail in Section 2.2, di erent cuto values as number of standard deviations (SD) are generated for the bootstrap distribution and the associated relative di erence distribution is generated. The di erence distribution is then cut in a similar way and the accuracy is computed. The result is presented in Figure 6 where the cuto s for the bootstrap distribution are in the vertical direction and the cuto s in the relative di erence are on the horizontal axis. This produces a heat map where the accuracy is color-coded with dark blue being the most accurate and white least accurate. For the four substances under investigation the optimal cuto for bootstrap OFV-values shows that the accuracy is practically independent of cuto

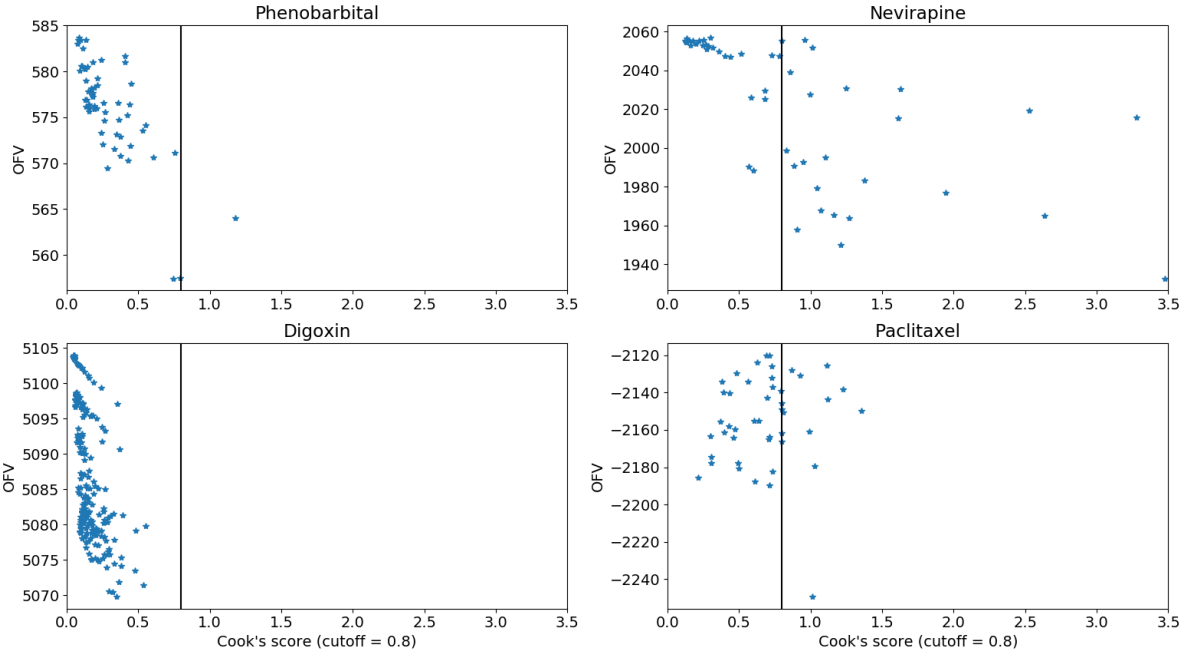

Figure 5: Objective function value plotted against Cook's score for four di erent substances.

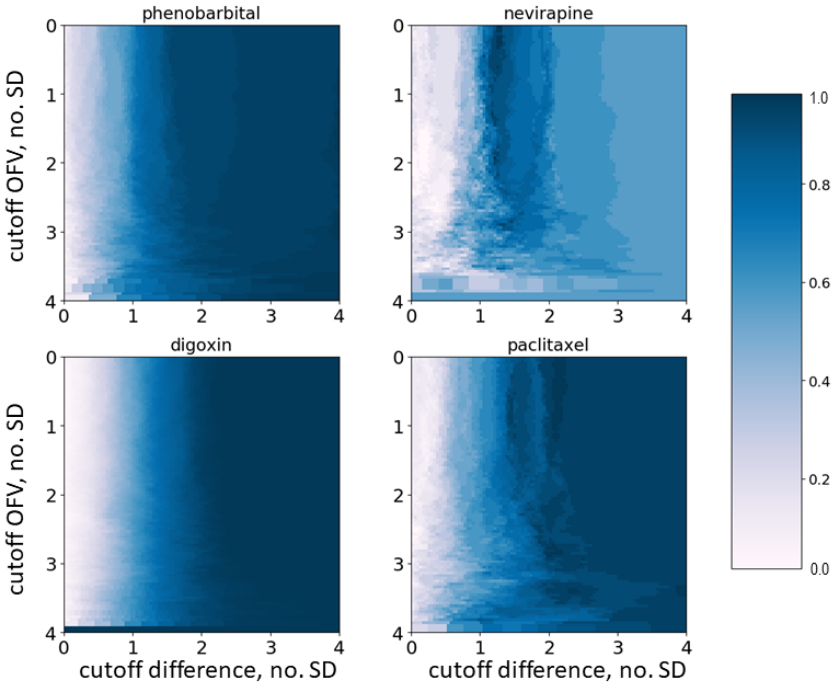

Figure 6: Identi cation of optimal cuto s.

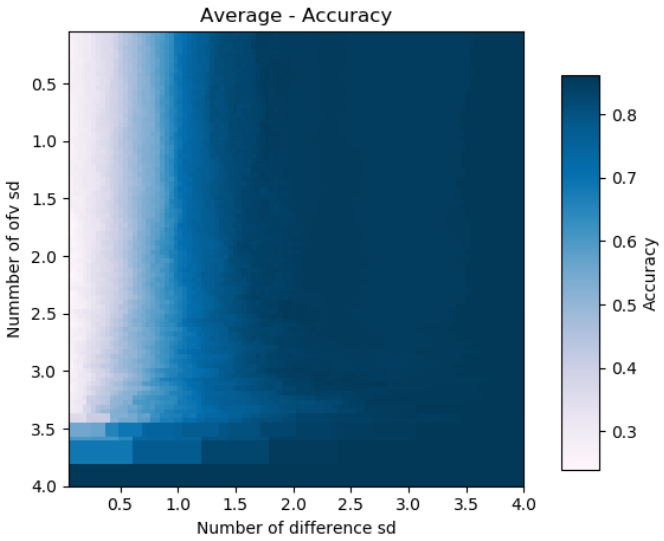

Figure 7: Identi cation of optimal cuto s from composite accuracy.

within the range 0.05 to 4 standard deviations. This indicates that the bootstrap runs could basically be divided based on the mean OFV value for the distribution. These results holds for all studied substances.

The accuracy as a function of cuto s for the di erence distribution displays a more complex pattern with streaks of high accuracy at one or two cuto s typically between 1 and 2 standard deviations.

A plot of the composite accuracy for the four substances is shown in Figure 7. Based on the composite analysis an average of 1.75 di erence standard deviations is selected as the optimal value for the di erence. Since there is almost no dependence of accuracy on cuto for OFV distribution, the lowest tested value of 0.05 OFV standard deviations is used. This allows for more bootstrap runs to be included which could be bene cial when only a small number of bootstrap runs are available.

3.3 Dependence on number of bootstrap runs

To evaluate the dependence on number of bootstrap runs our method is run for di erent number of bootstrap runs, ranging from ve to the whole ten-thousand with a spacing of ten between them. For each number, fty iterations are performed and the accuracy is computed. Figure 8 shows the results for the four substances, up to 4000 bootstrap runs. The error bars in the gure represent standard deviations.

The gure shows that the accuracy di ers between the substances. Phenobarbital and Digoxin are the substances that exhibit the highest maximum accuracy, at around 95%. Paclitaxel has maximum accuracy at around 85% while Nevirapine has maximum accuracy at around 65%. It is of interest that the substances that exhibit the lowest maximum accuracy (Paclitaxel and Nevirapine) are the substances whose models are the most complex ones in terms of number of model parameters. They include more than twice as many model parameters as the other models.

Another interesting feature visible in Figure 8 is the fact that all models converge to an accuracy value very close to maximal accuracy within around 700-2000 runs. In typical pharmacometric analysis, the number of bootstrap runs performed is usually less than 2000. This implies that our method converges to almost its maximum accuracy within the range of runs,

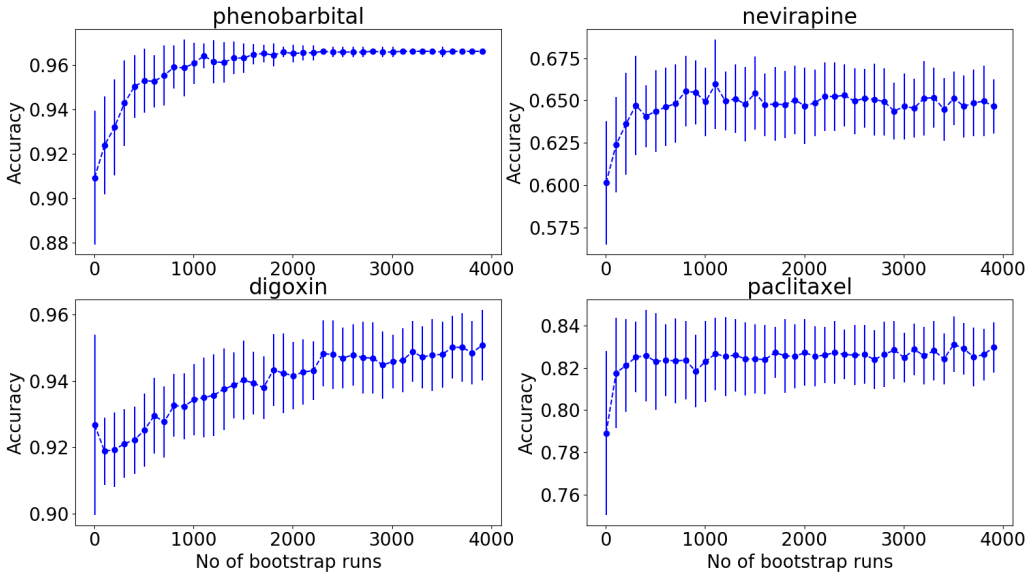

Figure 8: Accuracy for di erent numbers of bootstrap runs.

typically performed within the eld of pharmacometrics.

3.4 Correlation between Cook's score and new measure

Since CDD, as the current standard method, uses a Cook's score cuto to identify in uential individuals it is of interest to investigate the correlation between the di erence measure we used and Cook's score. As a measure of linear correlation the Pearson's score $\rho(X, Y)$ is used where X and Y represents the Cook's score and di erence measure, respectively and σ is their standard deviation. Pearson's score is de ned as

$$\rho(X, Y) = \frac{cov(X, Y)}{\sigma_X \sigma_Y},$$

where $cov(X, Y)$ is the covariance.
The Pearson's correlation between Cook's score and the di erence measure is shown in Figure 9, where the green line represents the optimal linear t of the data. Furthermore, to investigate possible monotonic non-linear correlation Spearman's correlation $r(X, Y)$ is used [10]. The formula is similar to the Pearson's score but use ranks (rg_X, rg_Y) instead of the actual values.

$$r(X, Y) = \frac{cov(rg_X, rg_Y)}{\sigma_{rgX} \sigma_{rgY}},$$

The Spearman's correlation between Cook's score and the di erence measure is shown in Figure 9 where the red curve is a t of the data to a 2^{nd} degree polynomial.

The correlation values for the investigated substances are summarized in Table 2. The substance Paclitaxel has a very weak or non-existing correlation for both Pearson and Spearman. Substances Phenobarbital, Dioxin and Nevirapine all displays medium to high correlations using both methods. This hints to a connection between the Cook's score and the new di erence measure. The reason for the deviant behavior of the substance Paclitaxel is not known. A

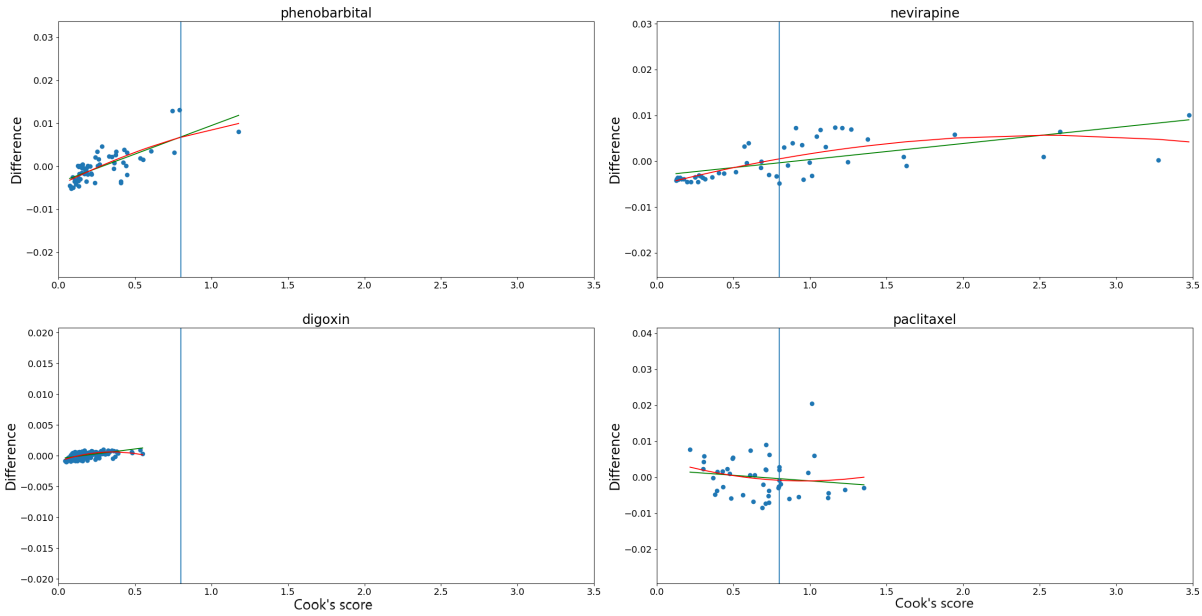

Figure 9: Scatterplot of di erence vs Cook's score with data tted to line and curve.

Table 2: Summary of correlations

Substance	Pearson correlation	Spearman correlation
Phenobarbital	0.75	0.72
Digoxin	0.58	0.68
Nevirapine	0.63	0.78
Paclitaxel	-0.15	-0.22

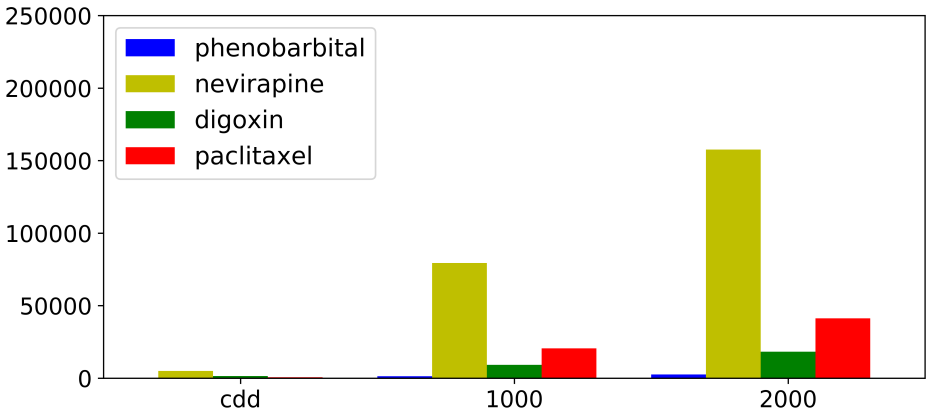

Figure 10: comparison of run time with CDD, 1000 and 2000 bootstrap runs

Table 3: Command line arguments

Argument	Description
-help	Prints usage message
-in	Input le (bootstrap result)
-out	Output le with in uential individuals
-cuto _ofv	Manually set cuto for bootstrap (0.05)
-cuto _di	Manually set cuto for di erence (1.75)
-verbose_di	Output execution information

possible reason could the high percentage of skipped individuals. The removal of a large set of individuals could turn a correlated data set into apparently uncorrelated.

3.5 Run-time analysis

The run times for di erent substances are shown in Figure 10. The di erence among the substances depends on the number of trial size and di erent pharmacokinetics model and how complex the model is. As it can be observed in Figure 10 that Nevirapine is more computational intense than other substances.

CDD does not take a long time for analysis due to the fact that the sample size is relatively small compared to the numbers of bootstrap runs. The dimensions of the data used in this project ranges from 5 to 227. Compared with a bootstrap run of 1000, CDD only needs to run $i + 1$ sets of data, where i is the trial size. It can be observed together with Figure 8 that the accuracy rate converges to the maximum accuracy rate very fast. Within between 700-2000 runs, we can achieve the maximum accuracy rate, which is substance dependent.

3.6 Software

A standalone python software was developed based on the developed method for identi cation of in uential individuals. The software use les from bootstrap runs as input and outputs the identi er for those individuals that have been identi ed as in uential. Furthermore, there are a number of command line arguments available for the user to adjust the execution and output format. The list of arguments are presented in Table 3.

4 Conclusions

We have developed a new method that makes it possible to identify in uential individuals using only data from the bootstrap method. This suggests that eliminating the need to run the CDD method in traditional pharmacometric analysis may be a real possibility. The accuracy of our method varies between di erent substances but on average it is around 85%. One possible reason for di erent accuracy between the substances is in terms of the complexities of the underlying models. This is supported by the fact that the substances which exhibit the lowest accuracy are also the ones that include the highest number of model parameters, which can be thought of as a crude estimate of model complexity. They also correspond to the data sets that contain the highest relative abundance of in uential individuals. This makes the accuracy more sensitive to misclassi cation of true positives.

Our results suggest that the exact position of the cuto s in the total objective function value distribution from the bootstrap data (Figure 2) only vaguely a ect the accuracy. This means that the cuts can be performed in almost any positions, on either side of the mean, or simply one cut in the centre. For the eventual di erence distribution (Figure 4) the combined optimal cuto was determined to be around 1.75 standard deviations away from the mean. A common characteristic of our results is that the accuracy converges to a value very close to maximum after between 700 and 2000 bootstrap runs, for all four substances. This has real signi cance because in typical pharmacometric analysis, the number of performed bootstrap runs is within this range. This means that maximum accuracy of our method is reached within the limits of typical pharmacometric analysis.

Medium to high correlation between Cook's score and our method for nding in uential individuals is observed for three of the four substances. This applies for both Spearman's- and Pearson's-correlation and suggests a possible connection between Cook's score and the method we have developed. For one substance, Paclitaxel, weak or non-existing correlation is observed for both Spearman's- and Pearson's-correlation. The reason for the low correlation from Paclitaxel is unknown to us but a high number of individuals who are never included in the bootstrap runs for Paclitaxel (see Table 1), may be part of the reason.

Using the method we have developed, standalone Python software that takes bootstrap data as input and returns identi ers for the in uential individuals, was created. The software's execution and output format can be adjusted by the user through di erent command line arguments.

5 Discussion and Recommendations

In this project, we analyse ve sets of clinical trial data. The results are positive in terms of successfully identifying in uential individuals. However, in order to validate our alternative method to nd in uential individuals, there is the need to experiment on more, di erent substances.

The output of the software for bootstrap analysis has more metrics available. Here we used the total OFV of a bootstrap run. Within one bootstrap run, there is also an OFV contribution available for each individual. This would be of signi cant value for future investigation.

The characteristics of the bootstrap sampling can also in uence the results. For example, if one observation appears many times in all the bootstrap runs and another observation appears rarely, that can also limit the information we can retrieve, resulting in a reduced accuracy in identifying in uential individuals.

When classifying the individuals they are marked only as either in uential or non-in uential. Information about the e ects an individual has on a model is not included. Sometimes an

individual has a high in uence on model parameters without the in uence being high enough for the individual to be classi ed as in uential. Thus, a metric to describe the degree of in uence could be informative to evaluate the model and individuals.

In our alternative methods, we can successfully identify the in uential individuals, though the accuracy is not 100%. It is worth discussing whether it is important to identify in uential ones that sometime includes the false-positive non-in uential individuals. Or in the other case that it is very important to identify both the in uential and the non-in uential individuals.

6 Acknowledgements

We would like to give special thanks to our project's supervisor, Rikard Nordgren, for his assistance and guidance throughout the project work.

References

[1] D.R. Mould and R.N. Upton: Basic Concepts in Population Modeling, Simulation, and Model-Based Drug Development - Part. 2: Introduction to Pharmacokinetic Modelling Methods, CPT: Pharmacometrics & System Pharmacology, Number 2, 2013.

[2] B. Efron: Bootstrap Methods: Another Look at the Jackknife, Annals of Statistics, Volume 7, Number 1, 1979, pp. 1-26.

[3] T. H. Grasela and Jr. S. M. Donn: Neonatal Population Pharmacokinetics of Phenobarbital Derived from Routine Clinical Data, Dev. Pharmacol. Ther., Volume 8, 1985, pp. 374-383.

[4] O. Alskar, J. Korell and S.B. Du ull: A pharmacokinetic model for the glycation of albumin, J Pharmacokinet Pharmacodynk 39(3), 2012, pp. 273-282.

[5] Wikipedia contributors: Digoxin. In Wikipedia, The Free Encyclopedia. Retrieved 13:14, January 23, 2019, from https://en.wikipedia.org/w/index.php?title=Digoxinoldid=876262403

[6] D. Elsherbiny, K. Cohen, B. Jansson, P. Smith, H. McIlleron, U.S.H. Simonsson: Population pharmacokinetics of nevirapine in combination with rifampicin-based short course chemotherapy in HIV- and tuberculosis-infected South African patients, Eur J Clin Pharmacol, 2008.

[7] A. Henningsson, A. Sparreboom, M. Sandstrom, A. Freijs, R. Larsson, J. Bergh, P. Nygren, M.O. Karlsson: Population pharmacokinetics modelling of unbound and total plasma concentrations of paclitaxel in cancer patients, European Journal of Cancer 39, 2003, pp. 1105{1114

[8] Cook, R. Dennis. \Detection of In uential Observation in Linear Regression." Technometrics, vol. 19, no. 1, 1977, pp. 15{18. JSTOR, www.jstor.org/stable/1268249.

[9] Efron, B. Bootstrap Methods: Another Look at the Jackknife. Ann. Statist. 7 (1979), no. 1, 1{26. doi:10.1214/aos/1176344552. https://projecteuclid.org/euclid.aos/1176344552

[10] Wikipedia contributors. Spearman's rank correlation coe cient. In Wikipedia, The Free Encyclopedia. Retrieved 10:46, February 5, 2019, from https://en.wikipedia.org/w/index.php?title=Spearman%27s_rank_correlation_coe cientoldid=880188660