Óbuda University

PhD thesis

The use of extreme value statistics to develop new metrics for risk assessment in diabetes and trial data analysis

by

Mátyás Szigeti

Supervisor:

Tamás Ferenci

Doctoral Schoold of

Applied Informatics and

Applied Mathematics

Budapest, 2023

NYILATKOZAT A MUNKA ÖNÁLLÓSÁGÁRÓL, IRODALMI FORRÁSOK MEGFELELŐ MÓDON

TÖRTÉNT IDÉZÉSÉRŐL

Alulírott **Szigeti Mátyás** kijelentem, hogy a „**The use of extreme value statistics to develop new metrics for diabetes risks and trial data analysis”** című benyújtott doktori értekezést magam készítettem, és abban csak az irodalmi hivatkozások listáján megadott forrásokat használtam fel. Minden olyan részt, amelyet szó szerint, vagy azonos tartalomban, de átfogalmazva más forrásból átvettem, a forrás megadásával egyértelműen megjelöltem.

Budapest, 2023.03.22.

……..
(aláírás)

Acknowledgments

First and foremost, I would like to express my profound gratitude to Tamás Ferenci, who served as my supervisor and has been my invaluable mentor for more than a decade, demonstrating extreme patience and understanding. It is through his expert guidance and unwavering support that I discovered the field of medical statistics, for which I am truly grateful.

I would also like to extend my heartfelt appreciation to Klára Horváth for demonstrating genuine interest and empathy towards my research. Her support was fundamental in overcoming obstacles during challenging phases.

I would like to express my thanks to György Eigner, who had an instrumental role in bringing the idea of this research to life.

I express my deepest gratitude to Deborah Ashby for placing her trust in me and providing invaluable guidance. Additionally, I would like to extend my appreciation to the entire ICTU team, whose support paved the way for me to begin this PhD program.

I am indebted to Gábor Borgulya, whose support enabled me to embark on this journey in the first place.

Lastly, I am grateful to Annamaria Rihmer and Marta Böhönyei for their unwavering support during challenging times. Their assistance were paramount in stay afloat amidst the waves of adversity.

To all of the mentioned individuals, I offer my heartfelt thanks for their invaluable contributions, mentorship, support, and belief in my abilities. This accomplishment would not have been possible without them, and for that, I am eternally grateful.

Contents

1 List of abbreviations

Abbreviation	Meaning
ANCOVA	Analysis of covariance
ANOVA	Analysis of variance
BG	Blood glucose
BM	Block maxima
BMI	Body Mass Index
CGM	Continuous glucose monitoring
CI	Confidence interval
CONGA	Continuous Net Overall Glycemic Action
CTD	Connective Tissue Disease
CV	Coefficient of Variation
CVGA	Control-Variability Grid Analysis
CYC	Cyclophosphamide
DKA	Diabetic ketoacidosis
DLCO	Diffusing capacity of the lung for carbon monoxide
EQ-5D	European Quality of Life Five-Dimension
EVS	Extreme value statistics
EVT	Extreme value theory
FVC	Forced vital capacity
GDA	Global Disease Activity
GV	Glycemic Variability
HR	Hazard ratio
ILD	Interstitial Lung Disease
IQR	Interquartile Range
KBILD	King's Brief Interstitial Lung Disease
MAGE	Mean Amplitude of Glycemic Excursions
POT	Peak over threshold
RTX	Rituximab
SD	Standard Deviation
SGRQ	St George's Respiratory Questionnaire
T1DM	Type 1 diabetes mellitus
TAR	Time above range
TBR	Time below range
TIR	Time in range

2 Introduction

Generally speaking, inferential statistics focuses on the analysis of a sample taken from a population in order to draw mathematically founded conclusions about the population. In the majority of the applications, our interest is focused on the central part of the population, i.e., the first, or perhaps the first two moments. This is true both for traditionally presented simple indicators, and more complicated models (e.g., regression, which focuses on the first – conditional – moment) too. The significance of skewed distributions or outlying observations is recognized, but the usual approach is to somehow eliminate them (e.g., by transforming the distributions, using other, robust statistics, or removing outliers). Rarely is the opposite attempted, i.e., getting rid or putting less weight on the values from the center and investigating extreme observations, despite the fact that "outliers" – when they're not results of data entry errors or sensor malfunction – could contain very valuable information.

Extreme value statistics (EVS) is a branch of statistics investigating the distributions of observations with unusually low or high values. These are not just simple outliers, like data entry errors, but real part of the data which are far from the central tendency and also occur rarely, yet, have relevance and sometimes serious impact. Thus, in many application, they can't be simply neglected. This could find its applications in biomedical science [1] as in medicine, where extremes naturally have an important role, as to some extent diseases are usually non-normal conditions leading to some biomarkers reaching abnormally high or low levels. Of course if the sample contains ill people only, these values will not be extreme within the study group, but even the extremes relative to the sample or to a certain time period could be very meaningful in terms of the current status of the patient or the prognosis of the disease. Despite that, few examples exist for such approaches in medicine, in contrast to fields like architecture [2, 3, 4], weather and climate analysis [5, 6, 7], sports and finance statistics [8, 9, 10], where rare, extreme events have an overwhelming impact too [11].

2.1 Data need of EVS

Because of the rarity of these events or observations, they form just a fraction of the total sample, thus have much smaller effective sample size, therefore their analysis could be exceptionally difficult. EVS is rarely used in medical statistics, the main reason being the lack of data on extremes and the cost of obtaining

them compared to natural factors like weather which is simple and cheap to observe and abundant data are publicly available, with daily or even higher frequency and recorded for decades even for the most remote places of Earth. In contrast, in most cases, important biomarkers are usually those that require some sample taken which is typically followed by a complex and expensive process to analyse that; thus it might be measured for patient maybe a couple or dozen times through their clinical history and that is often enough for clinical decision making. Although EVS was used to analyse cholesterol levels [12], pneumonia and influenza deaths [13], the lack off sufficient data is a serious limitation in the wider application of EVS in biomedical field.

Advances in measurement technology make diabetology an exception. (Intensive care represents another exception, but it is only relevant for a few patients.) With the widespread availability of continuous glucose monitoring (CGM), high-frequency (typically 5 minutes sampling time) and longer-term (up to weeks or months, even in routine clinical practice) measurements became possible and gone into clinical practice relatively long time ago [14, 15]. Electronic patient monitoring and electronic health records increase the amount of data in medical research as well. The amount of data recorded multiples by each year [16], however, different medical field are not benefiting equally from this.

2.2 Background of diabetes mellitus

In order to get a better understanding of the problems addressed, the methods used and the results of in this study, the important concepts about diabetes are summarized here.

The term diabetes mellitus describes a group of diseases characterized by high glucose levels and abnormal carbohydrate, fat and protein metabolism metabolism [17]. Carbohydrates are one of the three main nutrients found in food, along with proteins and fats. They are essential for providing energy to the body, as they are broken down into glucose (a type of sugar) and used by cells for energy. This process is regulated by the hormone produced by beta cells in the pancreas called insulin, which helps to lower sugar levels in the blood by facilitating the uptake of glucose into cells, where it can be used for conversion to energy [18]. Diabetes is associated with a relative or absolute impairment in insulin production, along with varying degrees of peripheral resistance to the action of insulin and characterized by hyperglycaemia (which means high blood glucose level) and extreme variation in the glucose level due to the loss of the usual, physiological regulation of it. This can sometimes

result in abnormally low glucose levels too as a consequence of improperly dosed treatment.

There are two main types called ”type 1” and ”type 2”. Type 1 diabetes (T1DM) is responsible for about 10-15% of the total diabetes cases and it is an autoimmune disease, meaning that the previously mentioned beta cells are destroyed by the immune system of the patient itself [19]. Thus it directly results in impaired insulin production, and an absolute deficiency of insulin. Its appearance is sudden, happens usually in childhood or in adolescence, that's why it was formerly known as ”juvenile diabetes” (and type 2 as adult-onset diabetes) but can occur at any age [20]. Compared to type 2 diabetes, type 1 is a much more severe condition but it has – after a presymptomatic stage – detectable and very characteristic symptoms so in general it is easier to diagnose it [21]. In type 2 diabetes, the body is either insensitive to the effects of insulin or doesn't produce enough insulin to maintain a normal glucose level. This can lead to high blood sugar levels and a range of associated symptoms and complications similarly to type 1 diabetes. Type 2 diabetes typically affects older adults, although the incidence of type 2 diabetes is increasing in younger people and children as well. It can often be managed through lifestyle changes, such as healthy diet, maintaining a healthy weight, and getting regular physical activity, but many people with type 2 diabetes may also need to take medication or insulin to help control their blood sugar levels. Hereafter in this document the statistical analysis and investigation will focus on type 1 diabetes and in the following text the statements applied on type 1 diabetes by referring to ”diabetes” unless it is stated otherwise.

2.3 The global burden of diabetes

The global prevalence of diabetes was estimated to be 463 million people (9.3% of the total population) in 2019 while it was 285 million (4.1%) in 2009. The prevalence was more than 2.5 folds higher in high-income (10.4%) countries than low-income countries (4.0%) [22, 23]. Moreover, diabetes is one of the most important cause of mortality and morbidity in the developed world, with the incidence of both type 1 and type 2 diabetes steadily growing and together with their complications the substantial individual and societal burden through disability and lost life years are also increasing [24]. In 2017, out of the 56 million deaths in the world in total, it was estimated that in about 1.4 million cases diabetes was the primary cause of death, which was a 34.7% increase compared to 2007. The distribution between the main two sub-types were the follow: 345.5 thousand deaths by type 1 which means

15.1% increase compared to 2007, and over a million by type 2 diabetes which means a – quite extreme – 43% increase compared to 10 years earlier [24]. The impact of premature mortality of a disease is often quantified with the number of "Years of life lost" (YLL) which is the difference of the age at death and standard life expectancy at that age. Altogether, to the two sub-types 29.3 million lost years were attributed, meaning 29.9% increase between 2017 and 2007 [24].

Additionally, it was estimated that the total number of cases of both type 1 and type 2 diabetes will raise to 578 million (10.2% of the estimated total) by 2030 and 700 million (10.9% of the estimated total) by 2045 [23].

It terms of costs, the economic burden of diabetes was 1.5 trillion USD in 2015 with an estimated increase to 2.1-2.5 trillion USD by 2030 which is around 2% of the total global GDP [25].

2.4 Complications

2.4.1 Hypo- and hyperglycaemia

Blood glucose is typically given in one of two units of measurement, either milligrams per deciliter (mg/dL) or millimoles per liter (mmol/L). While mmol/L is used in the most of the European Union, UK, China, Canada, and Australia, mg/dL predominantly used in the USA and Japan, Israel, India and also in some European countries, for example France. The conversion factor between mg/dL and mmol/L is 18, so the commonly used cut-off level for hyperglicaemia of having higher than 180 mg/dL is equivalent to 10 mmol/L and the same for hypoglicaemia is lower than 70 mg/dL and 3.9 mmol/L [26]. Hyperglycaemia is much more common as it is a direct consequence of the disease, while hypoglicaemia is often the result of its inadequately used treatment [27].

Both the acute (i.e., sudden or quickly worsening) symptoms and long-term complications of diabetes can be serious and have a significant impact on a person's quality of life. Acute symptoms of diabetes, caused by both high or low blood sugar levels, are, in most of the cases, relatively mild symptoms such as dizziness, blurred vision, fatigue, and difficulty concentrating [28]. These symptoms can be unpleasant and disruptive, but they are generally temporary and can be managed through proper treatment and management of the condition. Further symptoms of hyperglycaemia include thirst, frequent urination, dry, itchy skin and in severe cases, nausea, vomiting, confusion, disorientation, rapid breathing and fruity or sweet-smelling breath [28]. The latter is caused by a condition called diabetic ketoacidosis (DKA), which is a serious complication of diabetes that occurs when the body does

not have enough insulin to control blood sugar levels and it stays high (above 250 mg/dL) for a sustained time. The lack of obtaining energy from glucose metabolism causes the body to break down fat for energy instead. When this happens, the body produces a byproduct called ketones, which can build up in the bloodstream leading to lowering of blood pH [29, 30].

In more severe cases, acute extreme hyperglycaemia (with a blood glucose level of 600 mg/dL or more), could, even after a short time, lead to life-threatening conditions. The high blood sugar causes the body to lose large amounts of water through frequent urination as the body tries to get rid of the excess sugar, which can lead to dehydration. This is why it is also called hyperosmolar state. This can lead to diabetic hyperosmolar syndrome which is also known as hyperglycemic hyperosmolar nonketotic syndrome [31, 32]. This leads to coma in 25-50% of the patients affected by these syndromes and death in between 10 and 20%, which is roughly 10 times higher than the mortality rate in patients with diabetic ketoacidosis [33].

2.4.2 Long term complications

The long term complications of diabetes are the consequences of the damage to the blood vessels and impairment of the blood flow caused by the elevated levels of blood glucose that takes affect in three main ways. High levels of glucose in the blood can cause damage to the walls of the small blood vessels, which can lead to a buildup of fatty deposits and inflammation. This can make it harder for blood to flow through these vessels. In addition to this, high blood sugar can also damage the nerves that control the blood vessels, which can lead to changes in blood pressure and blood flow. Lastly, diabetes can increase the risk of clots forming in the blood vessels, which can further increase the risk of complications. As a consequence of these, the most affected areas are the ones where small blood vessels have crucial role such as the eye and kidney [34, 35, 36].

Diabetic retinopathy (damage of retina) is an important cause of visual impairment and blindness occurring as a result of long-term accumulated damage to the small blood vessels in the retina. In 2020, there were about 1 million patient blind and about 4 million living with moderate and severe vision impairment due to diabetes [37]. Additionally, through the above mentioned pathways, diabetes is among the leading causes of nephropathy (deterioration of kidney function) and kidney failure [36, 38]. Together with reduced blood flow, neuropathy (nerve damage) can lead to problems such as numbness and pain in the limbs and increases the risk of foot ulcers, infection and eventual need for limb amputation and thus a drastically

decreased quality of life [39].

Moreover, adults with diabetes have a 1.5-2.3 fold increased risk of heart attacks and strokes [40]. Additionally, people with diabetes are more likely to have several other diseases including cognitive impairment, Alzheimer disease, hypertension, depression, anxiety, wide range of skin complications and poorer outcomes for several infectious diseases, including COVID-19 [19, 41, 42, 43, 44].

2.4.3 Highest possible blood glucose level

Under very rare and extreme circumstances blood glucose level can go way beyond the above mentioned, already extreme levels. There are well documented cases in the scientific literature where patients got close to (and survived) 2,000 mg/dl blood glucose levels [45]. It is worth mentioning that there are some – albeit rather anecdotal – evidence that the highest blood glucose level ever recorded and survived was 2,656 mg/dl (147.6 mmol/L) on 23 March 2008 at the Pocono Emergency Room in East Stroudsburg, Pennsylvania, USA and belongs to Michael Patrick Buonocore as recognised by the Guinness Book of Records [46, 47] for a rather dubious world record.

2.5 Treatments in practice

To address this problem, external control of insulin (and possibly glucagon) – the hormones that control the metabolism of glucose – is the best available solution (for this reason type 1 diabetes was formerly also known as insulin-dependent diabetes). This involves taking insulin either through injections or an insulin pump which delivers insulin to the blood stream through a small tube under the skin. However, maintaining their normal level is difficult to achieve as not only the effects of treatment but many other factors (activity, calorie intake) have major roles in this process and are different from person to person.

Real-time, autonomous control of glucose level is the aim of the artificial pancreas (AP) systems. It consists of an insulin pump and a continuous glucose monitoring sensor, which measures the person's blood sugar levels continuously and adjusts the insulin delivery based on the blood sugar levels throughout the day in a closed loop control fashion [48]. The CGM's sensor measures blood glucose levels indirectly, through measuring the glucose concentration in the interstitial fluid. This is the liquid between the body's cells and blood vessels transporting oxygen and nutrients (including glucose) from the blood to the cells. The result is that CGM's measurements lag behind the actual blood glucose levels [49]. The AP systems reduce the

burden of managing diabetes by automatically adjusting insulin delivery based on the person's blood sugar levels, rather than requiring them to manually adjust their insulin doses, which is not just inconvenient but also important for the reason that high proportion of type 1 diabetes patients are children, for whom the adherence to such a rigorous treatment can be a serious issue [50]. The proper control is still an issue yet to be solved and there are a lot of different algorithms and developments, the robust comparison of which would be important.

2.6 Glycaemic variability and associated risks

As the above discussion shows, the major risk factor in the development of complications is the quality of glycaemic control, i.e., the incidence of hypo- and hyperglycaemic events and the overall glycaemic variability (GV), that is how unstable glucose levels are over time [51, 52]. Characterization of the GV is therefore of paramount importance.

Metrics to measure GV from blood glucose curves provided by CGM are still not optimal and widely agreed upon [53, 54, 55], although there are several efforts underway to improve these [56]. Additionally, the lack of accuracy and reliability limits their use in clinical practice [57].

Focusing on the risk of hyperglycaemia itself, it should be noted that while GV and risk of hyperglycaemia are likely correlated, traditional GV metrics are inherently limited as hyperglycaemia risk metrics as they are very insensitive to high values if there are only a few of them. This is in contrast both to intuition (even a single or very few measurements above, say, 400 mg/dL raises the fear that the patient has a high risk of hyperglycaemia) and to the mathematical behavior of extreme values as described by EVS.

2.7 Traditional metrics

The current practice of summarizing continuously measured blood glucose curves uses several indicators, such as the Mean Amplitude of Glycaemic Excursions (MAGE) [58], using glycaemic excursions in excess of one standard deviation (SD) above the mean, the Continuous Net Overall Glycaemic Action (CONGA) [59], which is the SD of the differences between measurements taken at regular time intervals, simple coefficient of variation (ratio of the SD to the mean), interquartile range, or percentage time spent above or below a standardized clinical target glucose ranges [26] or the control-variability grid analysis (CVGA) plot [60] (which is

essentially the same, but in graphical form) and other type of graphical tools [61] and composite metrics [62] which enable the rapid evaluation of the CGM measurements collected for several days or weeks. These metrics, however, mostly focus on overall variability, not specifically on extremities which is not necessarily the same. A patient's variability can be very high, even if the blood glucose level is never in the extreme range, or the opposite could also occur (although this is highly unlikely), with the patient spending a lot of time in extreme range with very low variability. Thus, these metrics are not really appropriate to capture this aspect and the associated (and hidden) risks of hyperglycaemia. Even those metrics that do account for extremities (such as time spent above range) are usually very simply – and ad hoc – indicators mathematically speaking, which do not incorporate the statistical knowledge on the behavior of extremities.

2.8 Suitability for EVS

In contrast to the above presented metrics, EVS allows the estimation of the probability that the measurement exceeds a certain threshold (which is the relevant factor for hyperglycaemia), even if such values were never observed in the sample. (Note that time spent above range can never do this, as it will always estimate such probabilities to be zero.) By taking the sampling frequency into account, this can be used to calculate the probability that the patient's blood glucose will be above a threshold in a given time span (e.g., in 1 year) and the expected time spent above the threshold in the interval. The concept of return level is also often used in EVS: this is the level, blood glucose value in the present case, that is expected to be exceeded exactly once in every year (or any other time interval specified, called the return period). Taken together, these raise the possibility that metrics based on EVS are more useful to accurately capture the risk of hyperglycaemia. This approach is based on a much more sophisticated statistical foundation, addressing the extremes directly [63, 64, 8, 13].

3 Objectives

The main objective of the current study is to develop a novel approach that focuses on the maximums of blood glucose measurements using EVS methods instead of the traditional metrics used in the current practice for the assessment of CGM data. In order to do this, one of the first steps will be to map the capabilities of EVS in regards to such analysis of CGM measurements and other diabetes related data. This includes the examination and presentation of its theoretical background and the assessment of the possible EVS approaches such as "the peak over threshold" (POT) and "block maxima" (BM) approaches though preliminary studies. This should include their comparison to each other and to the currently accepted traditional metrics in terms of risk assessment of hyperglycaemia and quality of blood glucose control on patient and population level and to present their advantages and disadvantages and find the needs of further analysis especially in terms of data needed. For such analysis, obtaining the right amount of data is crucial, and might require lot more data than regular methods or the assessment through the traditional metrics because EVS focuses more on extreme – thus rare – events. In order to overcome this issue, accessing large amount and high quality CGM data should be a priority. Ideally, this data should include high and extreme measurements so it would be the best to get it from real-life patients with relatively severe diabetes. Alternative solutions such as simulated data or use multiple sources should be considered and critically assessed. Additionally, the preliminary work should also include the mapping of currently available statistical software solutions for EVS calculations and analysis especially packages available to R [65] and assess their capabilities in terms of applicability of a large scale analysis of CGM data.

Furthermore, after the best approach was selected based on the preliminary investigations, its applicability should be demonstrated in a large, real-life dataset and a complete analysis of it with both the EVS approach and the traditional metrics should be undertaken. This would be the first time for such analysis, and it would allow the comprehensive comparison of the traditional and the EVS based approach. Also, sensitivity analysis have to be performed with different assumptions or restricted datasets simulating different scenarios to get a better understanding of the behavior of the methods under different conditions. Using these results, the patient level risk assessment should include the selection of the patients with the highest risk of hyperglycaemia. The level of risk should be characterised and be informative, relevant and interpretable from clinical point of view. This should include the analysis and assessment of probability of reaching and the time spent

above clinically important levels of blood glucose, such as the threshold of the above presented level for diabetic hyperosmolar syndrome. The group of patients with the highest risk identified through the EVS analysis should be compared with the group with the highest risk through the traditional metrics. If the patients in these groups are different, then this is a key difference between the two approaches and should be investigated in depth and the possible reasons for the differences should be analysed and presented, including the examination of the raw CGM data of these patients. The analysis should be extended to more comprehensive analyses with the use of non-stationary models, which are to some extent similar to regular regression type of models as it is possible to add patient level baseline clinical characteristics to the model with them and to estimate their effects on the outcomes. Thus, ideally, the dataset used for this analysis should also contain such clinical data of the patients. If possible, the analysis should be extended to the validation of the data as well. The question of statistical independence should be also investigated in regards of the EVS in such time-series type of data as the CGM. This issue and its importance should be compared between the EVS and regular regression modelling methods. Of course, the strengths, weaknesses and limitations of the whole analysis should be thoroughly explored and presented.

Apart from the EVS related analysis , the statistical aspects of another, advanced regression analysis of a clinical trial will be presented. This is the primary analysis of the "Recital" trial which was a contemporary, randomised, controlled trial to comparing the effectiveness and safety of a novel treatment with standard care for a potentially life-threatening lung disease.

4 Proof-of concept studies

4.1 Introduction

At the time when this research began both STATA and R [65] had the capability of EVS calculations but STATA got it not so long ago with the introduction of David Roodman's "extreme" package in 2015 [66]. This seemed to be limited and less flexible compared to the relatively wide range of available EVS related packages available in R such as "fExtremes" [67] and "extRemes 2.0" [68], so this programming language and environment was chosen.

Obtaining CGM data seemed a difficult obstacle, as it focuses on extreme thus rare events requiring more data than other, regular statistical methods. For the purpose of a proof of concept analysis it seemed a good solution to first use simulated data, which provides potentially infinite amount of data, but – as it was discovered later on – was sensitive to its parametrisation and produces less real-life like data towards the tail of the distribution which would be the most important for the EVS analysis. In subsequent analyses, different real-life datasets were used (as discussed in the pertaining sections).

4.2 Extreme Value Theorem

In this chapter the theoretical background of EVS and its two main approaches will be presented. These have many similarities but the most important is that they both use a secondary sample that is taken from the raw data which gets analysed and their names reflect the way this secondary sampling takes place.

Namely, the Peak Over Threshold (POT) approach uses a chosen cut-off level and takes only the observed values over that level as the secondary sample, the block maxima (BM) approach splits the data to equal sized non overlapping blocks of observations and takes the maximum (or minimum) of each block as the secondary sample to analyse.

Historically the BM approach was discovered and used first when the behaviour of extreme values was formally described, initially by Ronald Fisher and Leonard Henry Caleb Tippett in 1928 [69]. Their findings were later proven by Boris Vladimirovich Gnedenko in 1943 [70]. Together these form the so-called Fisher–Tippett–Gnedenko theorem which establishes that if there are constants with which the maximum of independent and identically-distributed random variables can be linearly transformed so that this renormalized variable converges to a non-degenerate distribution, then this distribution must be one of the following:

$$F(x) = \begin{cases} \exp\left[-\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}_+^{-1/\xi}\right] & \text{if } \xi \neq 0 \\ \exp\left[-\exp\left\{-\left(\frac{x-\mu}{\sigma}\right)\right\}\right] & \text{if } \xi = 0 \end{cases}$$

Here $\mu \in \mathbb{R}$ is the location, $\sigma > 0$ is the scale and $\xi \in \mathbb{R}$ is the shape parameter. Fortunately these distributions are closed under linear transformation, so instead of saying that the transformed maximum converges, we can say with the above formulation that the maximum converges to one of the above distributions without loss of generality, as location and scale parameters have to be estimated from the sample anyway.

The distribution presented above is called the Generalized Extreme Value (GEV) distribution. It covers three special cases based on the value of ξ, which are the following:

1. Frechet distribution ($\xi > 0$):

$$F_\alpha(x) = \exp\left\{-\left(\frac{x-u}{\sigma}\right)^{-\alpha}\right\}$$

2. Weibull distribution ($\xi < 0$):

$$F(x) = \exp\left\{-\left(-\left(\frac{x-u}{\sigma}\right)^{\alpha}\right)\right\}$$

3. Gumbel distribution (also called as largest extreme value distribution) ($\xi = 0$):

$$F(x) = \exp\left\{-\exp\left(\frac{x-u}{\sigma}\right)\right\}$$

While this formulation is suitable for analysis from the probability theory point of view, in statistical investigation the maxima has to be estimated. A sample is needed, so we can't simply take the maximum of the whole series, and that's where the BM approach first appeared, i.e., block maxima were used to capture maxima.

The distribution is described by three parameters: shape, scale and location. When we fit a model to the available empirical data (sample) the values of these parameters – and its uncertainty – has to be estimated. Several statistical methods can be used, but most commonly the maximum likelihood estimator (MLE), the L-moments and the Bayesian method are in used in practice[71, 72]. The MLE is an often-used method due to its reliable results and its simplicity, thus can be used

for large data sets where other more computationally intensive methods are not efficient. The L-moment method is based on the linear combinations of probability weighed moments, while the Bayesian is a complex method, which uses the initial data and offers the opportunity to use supplementary information about it from external sources through the prior distribution.

Using the other main approach, the secondary sample of Peak Over Threshold (POT) would asymptotically follow a so-called Generalised Pareto distribution (GPD) that was first introduced by James Pickands III half a century after Fisher and Tippett's work, in 1975 [73]. He has shown that the behavior of these extreme values after the POT re-sampling follows the following probability law:

$$P(X - u < y | X > u) \approx 1 - \left(1 + \frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi},$$

where ξ is the shape and σ is the scale parameter.

The role of extreme value theorem for block maxima or for peak over a threshold is very similar to the role of central limit theorem as both theories involve limiting behaviors of distributions of independent and identically distributed random variables, except two remarkable differences.

Firstly, the extreme value theory only concerns the behavior of a subset (the tails) of those distributions while central limit theorem concerns the behavior of entire distributions of random variables. More precisely, while the extreme value theory describes the limiting behavior of the extremes $max(X_1, \ldots, X_n)$ or $min(X_1, \ldots, X_n)$ the central limit theorem describes the limiting behavior of $X_1, X_2, \ldots, X_n$ [72].

Moreover the central limit theorem applies to the mean of a sample from any distribution with finite variance, while the Fisher–Tippet–Gnedenko theorem states that if the distribution of a normalised maximum converges, then the limit must be one of the distributions described above [74].

4.3 Statistical dependence and EVS, relation to regular regression models

It should be noted that in most cases, where EVS is applied, the observations, such as CGM measurements, are time series data meaning it is a sequence of data points of the same variable, collected over time. This means that their independence is usually very questionable. In case of CGM data, the observations are likely positively autocorrelated, meaning the observations are positively correlated between two successive time intervals.

The two approach, BM and POT, behave very differently when faced with dependent observations. In the POT approach, dependence means a deeper problem as extremes tend to cluster, which will show up in the secondary sample of the POT, as many observations around the extremes are likely to be also relatively high and also above the threshold. This necessitates the application of more complicated techniques, the most widely used of which is declustering.

The BM approach greatly alleviates these problems: with that, local dependence is much less of concern, and even for the long-term dependence only a rather weak condition, the so-called D(un) condition of Leadbetter [75], is required to ensure that block maxima of the dependent data will have the same distribution as independent data would have [76]. The drawback of the BM approach is the less efficient use of sample size, but with large enough samples this is of less concern.

Another form of dependence occurs if the observations come from different sources, for instance, different patients, in which case the dependence is induced by the clinical characteristic of the patients, such as age or sex. This gives rise to non-stationary series. In this case, one could either analyse every patient separately, or use a single model for all patients, but with parameters that are allowed to depend on patient characteristics. These are called non-stationary models and demonstrated in Chapter 5.12. This is very useful feature of EVS as it allows a deeper investigation if such data is also available and enables us not just the evaluation of individual risk but also inference on the impact of the covariate, similarly to regular regression models. If these are applied to EVS models, coefficients will be attributed to the three parameters of the distribution separately or in any combination, but not directly to the outcome. The effect on the outcome can be calculated using the estimated parameters, but this is a two step process as the estimated parameters describe not the behaviour of the data itself but rather the behaviour of the distribution of the maximums of the data.

4.4 Peak-over-threshold (POT) approach

4.4.1 Data

The sample used for this analysis was generated using the simulation from the UVA/PADOVA Type 1 diabetes simulator [77]. This contained 1440 measurements of 99 simulated patients, which – having 1 minute between the measurements – covered a whole day for each simulated patient. The simulated measurements of the 99 patients are shown in Figure 1. A couple of the weaknesses of the simulated data is easily observable with a brief graphical assessment. The starting value for

each patient were the same, their main trends during the ”day” are very similar and interestingly the ending of each trajectory became more similar than the mid-day values of the simulated patients as they trending towards a common value. The median glucose level across the whole dataset was 123 mg/dl with a range of 57-343 mg/dl so there were some extreme values and overall the results of the simulations seemed good enough for a proof of concept type analysis.

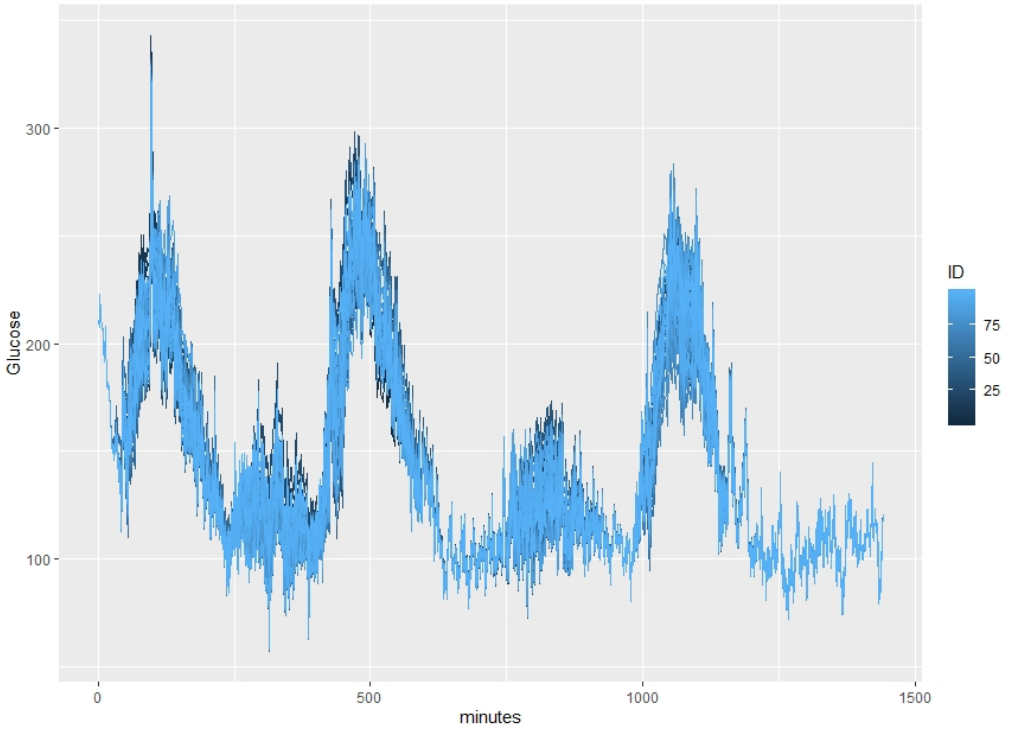

Figure 1: Plot of 1440 simulated blood glucose measurements of 99 patients

4.4.2 Choosing the threshold level

First, a threshold level was chosen for the secondary sample. This is a key element of the analysis because on the top of the purely statistical considerations, the investigated level has to be clinically significant and meaningful level in order to enable us to characterise a practically important region of glucose levels. Additionally, if the threshold is too high then we lose too much data and the remaining – while characteristic for the extremes – might not be sufficient for a reliable statistical analysis, on the other hand, if the threshold is too low then the model will be fitted more and more to the non-extreme parts of the data.

This problem was resolved using the mean excess plot (Figure 2) and the threshold range plot (or also called as Parameter Stability Plot or Threshold stability

plot) (Figure 3). The mean excess plot or as often it's called, the mean residual life plot (MRL-plot) was introduced by Davison and Smith in 1990 [78]. The MRL-plot is used the present expected value of the GPD excesses. For an u threshold the expectation of the excesses will be

$$E(X - u|X > u) = \sigma_u/(1 - \gamma),$$

where σ_u is the scale parameter for a given threshold of u and γ is the shape parameter which needs to be set for $\gamma < 1$ to ensure that the mean exists. Given its linear property, the expected value of the excesses for any threshold $v > u$ will be

$$E(X - v|X > v) = \frac{\sigma_u + \gamma v}{1 - \gamma},$$

which can be shown to be linear in v with the gradient $\gamma/(1-\gamma)$ and intercept $\sigma_u/(1 - \gamma)$ [79].

Thus the mean excesses can be plotted and the assumption is that where the plot starts showing a linear behavior a suitable threshold can be estimated. In theory the plot is very likely to lose its linear behavior when the threshold gets too large due to the variance of a few extremes left will cause the plot to jump. It might happen that for some data the plot is completely or never linear and little to no information may be collected from observing the plot [79]. Fortunately in this case the mean excesses plot, as shown in Figure 2 shows an ideal (flat) region starting at 250 mg/dl which is also in the region that can be considered clinically important threshold while remaining low enough to utilize as much of the data as possible.

The other useful tool to determine the threshold level is the threshold range plot introduced by Scarrot and MacDonald [80]. This is used to see how sensitive are the shape and scale parameters to the value of the threshold by calculating and plotting them for a range of u thresholds assuming a constant shape parameter when calculating the different scale parameters and assuming a constant scale parameter for calculating the different shape parameters. These estimates with their 95% CI can be plotted against u. A suitable candidate for threshold is at the lowest value where these plots are approximately constant. Depending on the data there might be cases where the parameter stability plot doesn't give any further information on the situation just like with the mean excesses plot [79]. In this case Fig. 3 also supports the choice of 250 mg/dl as threshold. The choice of the proper threshold is always a subjective decision to some extent.

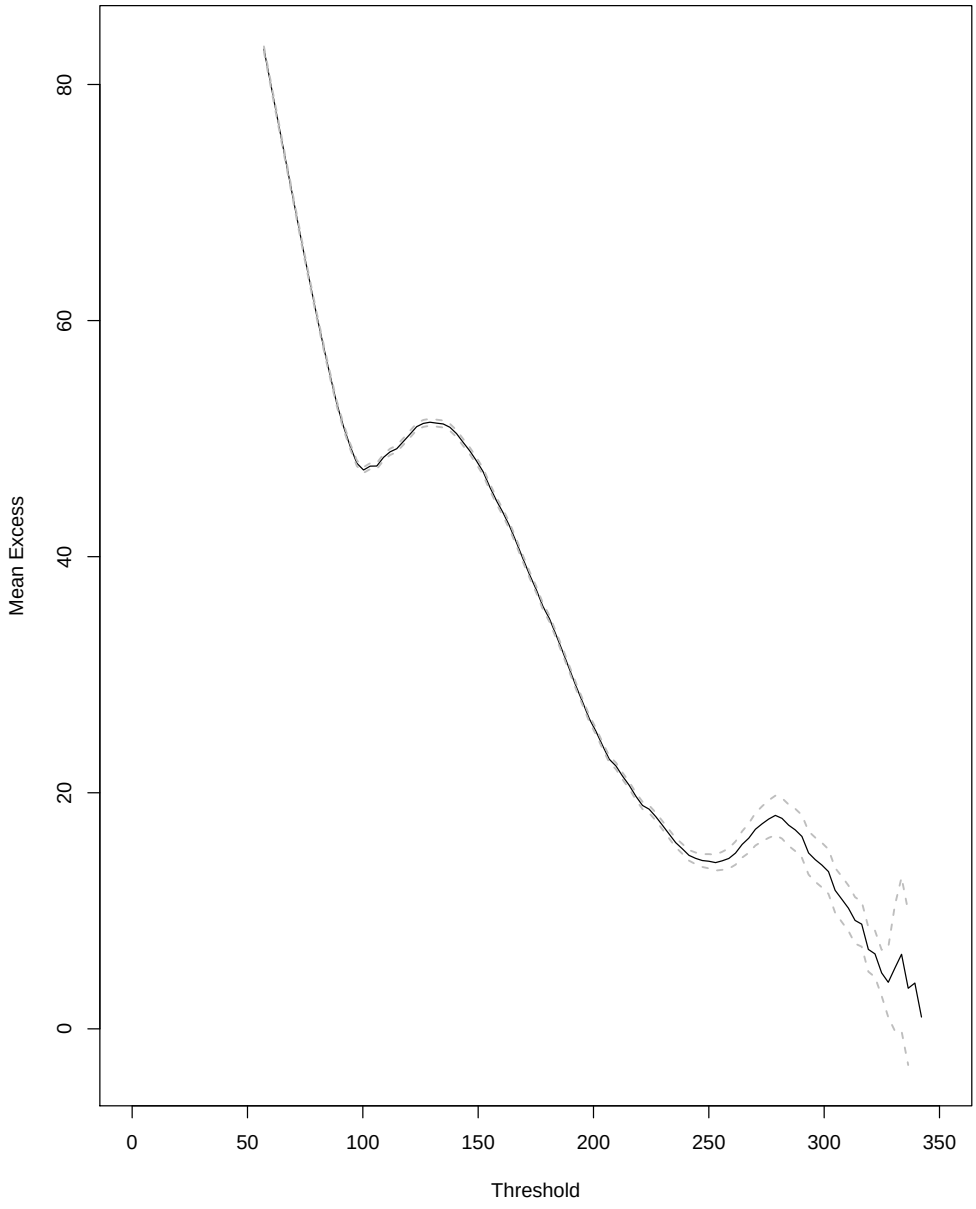

Figure 2: Mean excess plot show that 250 mg/dl was an ideal threshold level.

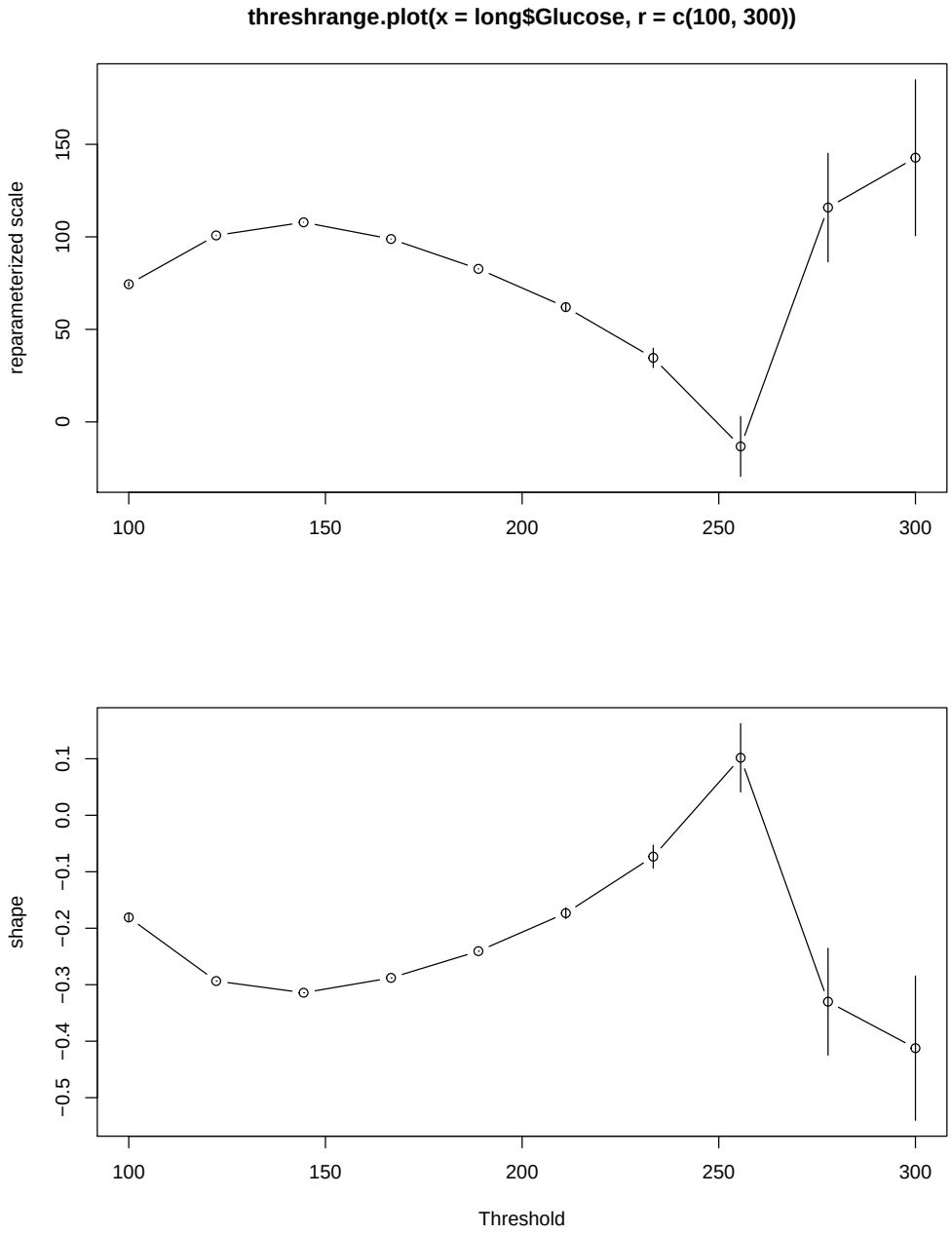

Figure 3: Threshold range plot shows the scale and shape parameter values and their 95% CI for different thresholds.

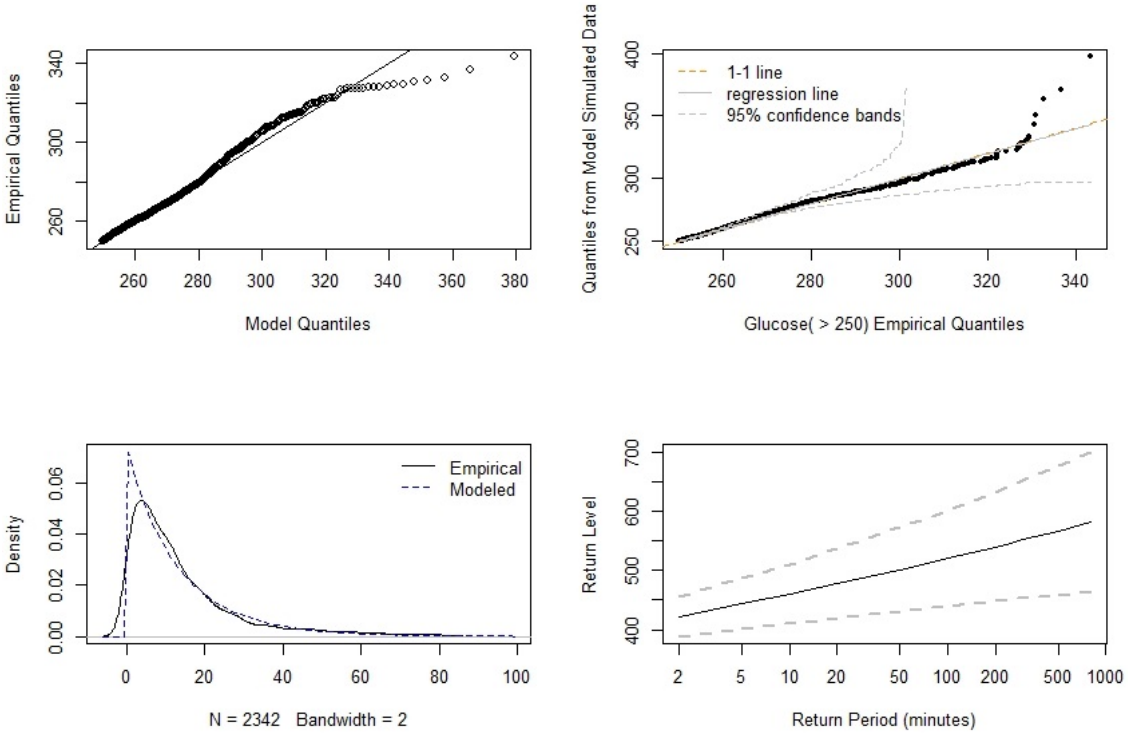

Figure 4: Model diagnostics plots - Peak over threshold

4.4.3 Results

A POT model was fitted to the simulated dataset. The analysis itself was conducted using the R program package version 3.6.2 [81] and with the use of the ExtRemes 2.0 package [82].

The fitted model assumes a Generalised Pareto distribution for the values above the threshold with 13.46 (95% CI: 12.66, 14.27) as estimated scale and 0.0530 (95% CI: 0.0088, 0.0973) as estimated shape parameters.

The goodness of fit was assessed graphically by various diagnostic plots shown on Fig. 4. The top two graphs are quantile-quantile plots (Q-Q plots), where quantiles of a sample drawn from the fitted GEV distribution and the quantiles of the empirical data are plotted against each other with 95% confidence bands (top right plot). Density plot of empirical (observed) data and the density plots fitted GEV distribution presented as well (bottom left) and the return level plot (which is not really a model diagnostic plot but presenting some part of the results so it will be assessed later) with its 95% (pointwise normal approximation) confidence intervals is also shown (bottom right).

The model fitted well to a certain level around 320 mg/dl and above that level the Q-Q plots suggest that our data was less extreme than what modelled

GP distribution would suggest. This was expected because of the nature of the simulated data. Additionally, as the very wide confidence bands suggest, there were just a very few observations available above 300 mg/dl. For this reason 270 mg/dl was chosen as an investigated threshold which is within the available data's range and symptoms as cognitive impairment are expected at this level [83]. We chose 600 mg/dl as a secondary limit which is way beyond this range, but is an important physiological limit for diabetic hyperosmolar syndrome which leads to coma for 25-50% and death in between 10 and 20% of the patients [33]. For these levels the probabilities to be exceeded for a new peak were calculated as a clinically meaningful outcome. According to the fitted model there is 23.9% probability per day to exceed 270 mg/dl and $8.0 \cdot 10^{-6}$% probability per day to exceed 600 mg/dl on average for a new peak above 250 mg/dl.

Note that at this stage a stationary model was used only and the data of all simulated patients were used as a whole, with no attempt to be made for patient level risk assessment.

4.4.4 Conclusions

As this was a proof of concept type of analysis at this stage, it had limited aims. The idea that EVS can be used to assess risk or the quality of blood glucose control based on CGM data was successfully proven, however important limitations were also discovered paving the way for further and different analyses which lead to focus on BM approach as a next stage.

First, POT analysis only reflects values above a certain threshold and within that tail of the data it enables us to calculate estimates for the probability of reaching certain extremes. This is useful to compare e.g. different antidiabetic drugs or AP controllers comprehensively, but does not take into account the extremes' occurrence in time, therefore it is less suitable for patient-level characterization of glycaemia. Thus the outcome itself is more suitable for population level comparison rather than patient level which is useful on one hand but does not really materialise the true potential of EVS and does not provide an alternative to the traditional metrics used for patient level risk assessment of hyperglycaemia based on glycaemic variability. The block maxima approach focuses on the distributions of maxima in time so it might be more relevant clinically to give such an estimate for each patient.

One of the other important limitation was that the threshold of diabetic hyperosmolar syndrome (600 mg/dl) was very far from the upper boundaries of the data (maximum: 343 mg/dl) thus the results might be uncertain and unreliable for the higher investigated threshold. Additionally, although the simulated data had many sophisticated features, those capabilities were not too useful for the current analyses. By the nature of the simulated data, there were clinically unrealistic similarities between the simulated patients which were discovered during the analysis. Their starting value was the same and if there were no (simulated) glucose or insulin input they reached the same reference level (middle and last section of Fig. 1). The periodicity of glucose inputs (" simulated meals") were the same and happened at the same time for all patients. It was also expected that the model would fit better on non-simulated data as the simulation did not use GEV distribution to draw a sample from, but real-life data might have such characteristic. So an important and difficult further task was to obtain significant amount of real-life data to try EVS methods with BM approach on.

4.5 Block maxima (BM) approach

4.5.1 Data

For the reasons outlined above, the next phase aimed to do a similar proof of concept type analysis using the experience gained from the POT study, but this time applying the BM approach with real life data.

For that end, the sample dataset from the R package 'gluvarpro' was used; this preprocessed the measurements of a single type 1 diabetic patient acquired through Abbott Freestyle Libre continuous glucose monitoring sensor. The dataset contains 476 complete days of data taken with sampling time of 15 minutes, this means 45696 observations in total [84]. This enabled us to conduct a proof of concept EVS analysis using BM approach with some obvious limitation inherited from the relatively large but homogeneous dataset. For all analysis, the same 3.6.2 version of the R statistical program package [81] and ExtRemes 2.0 [82] package was used supplemented by gluvarpro [84] package.

The plotted time series of a week of the blood glucose measurements of the sample are shown in Figure 5. The median glucose level across the whole dataset was 135 mg/dl, with a range of 25-429 mg/dl. Luckily from the analysis' point of view even this single patient's data had larger range and thus more extreme values than the previously used simulated data.

4.5.2 Selecting the Block Size

A BM model was fitted to the dataset with the block maximum calculated with a block size of 4 – of 15 minutes long – periods, i.e., hourly maxima were calculated. This was a somewhat arbitrary choice and like most cases in statistics, it involves making a trade-off between bias and variance. Too small block sizes lead to bias in estimation and extrapolation while large blocks generate few block maxima/minima, leading to large estimation variance and less efficient use of data. An important thing to keep in mind is the trends or seasonality of investigated signal. Most physiological processes have a daily, repeated pattern thus a block size larger than a day would be very inefficient. Blood glucose level can change relatively quickly compared to most of physiological factors depending on many factors but most importantly on glucose intake (eating) and physical activity. In general, practical considerations usually lead to the use of blocks of length one whole unit of time i.e., hour, day, week or year [76].

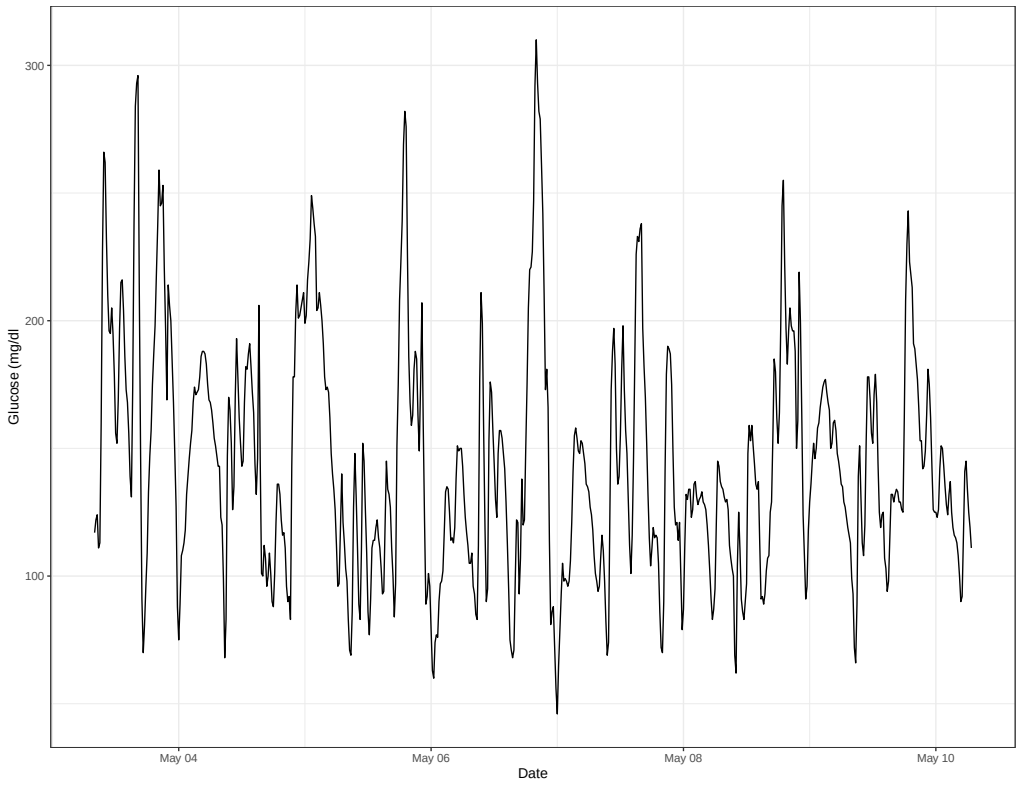

Figure 5: Plot of a week of glucose measurements

4.5.3 Results

The estimated parameters of the fitted GEV model and their 95% confidence intervals were: $\widehat{\mu} = 133.74$ (95% CI: 134.64, 135.54) as location, $\widehat{\sigma} = 42.93$ (95% CI: 43.58, 44.23) as scale and $\widehat{\xi} = -0.058$ (95% CI: -0.045, -0.031) as shape parameters. The shape parameter is negative and its CI does not contain 0, so the fitted GEV distribution is a Weibull distribution. This is also confirmed by using a likelihood ratio test to compare this model to the restricted model with $\xi = 0$, i.e., Gumbel model; this results in $p < 0.0001$. The goodness of fit of the model was assessed graphically similarly to the previously presented POT model by the same diagnostic plots shown on Fig. 6 which were described in Subsection 4.4.3. The model fitted very well to a certain level around 320 mg/dl but above that the observed data is seemingly slightly less extreme than what the Weibull distribution would suggest. Nevertheless, the overall fit is much better than what we observed when POT model was fitted on simulated data in the previous analysis in Subsection 4.4. Despite the fact that only a very small proportion of the observations exceeded 270 mg/dl (1.97%), leading to a suddenly widening confidence interval of the estimated values in that region, the overall goodness of fit is very good.

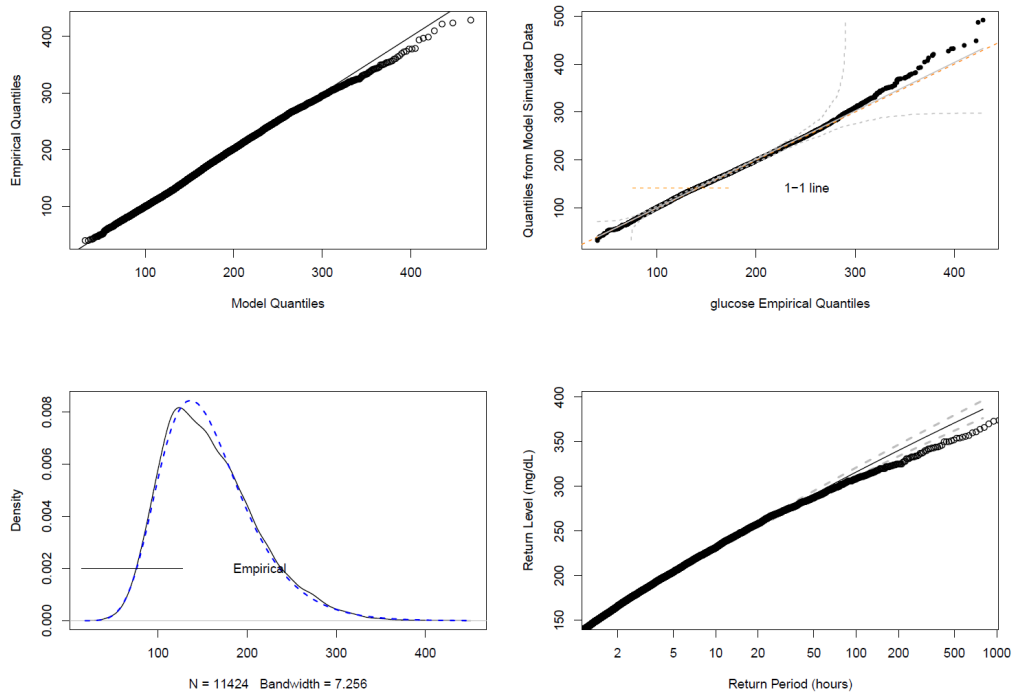

Figure 6: Model diagnostics plots - Block Maxima

After fitting the model, return levels were calculated which are the maximums which occur once on average during the given time period. These are presented in Table 1. It can be seen that, according to the estimated glucose levels, this patient gets quite close (with 383 mg/dl) to the lower but still dangerous level we investigated in the POT analysis (400 mg/dl) once a month on average and once a year it goes way beyond with 459 mg/dl. For both estimates the confidence intervals are relatively narrow compared to the estimated value itself. Also note that with the 1 year return value the estimation is extrapolated beyond the range of observed values. This is an important additional feature compared to the traditional metrics, especially as diabetes is a chronic, life-long disease where continuous blood glucose monitoring is advised but usually 14 days long observations are used to assess the condition of the patient through the traditional metrics [85].

The other main feature and advantage of the BM approach is that it is also possible to calculate the probability that the blood glucose level exceeds a certain threshold level, as the appropriate quantile of the fitted distribution. In contrast with the POT method, this applies for the whole range of possible values not just for a subset above a certain level. So instead of getting a (conditional) probability to exceed a certain level for a new peak above the pre-specified threshold level applied

Table 1: Estimated return levels of the BM model for the example CGM of gluvarpro's example patient

Return level	**Estimate (mg/dl)**	**95% CI (mg/dl)**
1 day	262.95	260.23 - 265.67
1 week	334.11	328.00 - 340.22
28 days	383.02	373.38 - 392.67
1 year	459.82	442.53 - 477.10

to get the secondary sample, the estimated probability though the BM approach gives a direct estimation. Moreover, this probability multiplied by a certain time period of interest (e.g., a year) gives the time spent above that blood glucose level during the time period which gives an easy to understand answer to a different but more relevant research question.

Similarly to the POT analysis, as an illustration, 270 mg/dl was chosen as one of the limits, as it was within the range of the actual data of this patient, but also detectable acute symptoms as cognitive impairment are expected at this level [83] and 600 mg/dl was chosen as a secondary limit of interest which is way beyond the range of the observed available data, but it is an important physiological and dangerous limit for the above mentioned diabetic hyperosmolar syndrome which leads to coma for 25-50% of the patients and the mortality is between 10 and 20% [33]. According to the fitted BM model presented above, there is 3.47% probability to exceed 270 mg/dl (which means 30.4 hours per year) and $4.96 \cdot 10^{-7}$% to exceed 600 mg/dl (which means 15.6 seconds per year in total) for this patient.

4.5.4 Conclusions

With this analysis, the concept of using EVS with BM approach to assess the quality of glucose control based on a CGMS curve, thus the attributed risk to some extent, been successfully proven. This enabled the comparison with the POT approach, however there were many important points yet to be answered.

The main difference and advantage of the BM approach is its ability to directly reflect the probabilities of reaching certain extremes and the amount of time spent above certain blood glucose levels. This is applicable to compare not just different treatments such as antidiabetic drugs or AP controllers comprehensively, but as it also takes into account the extreme's occurrence in time as return levels or the directly estimated probabilities, it is also suitable for patient-level characterization

of glycaemia and the associated risks through the absolute estimates.

Thus, the block maxima approach focuses on the distributions of maxima in time so it might be more relevant clinically to give a metric for each patient than the POT approach.

It is important to note that this analysis was limited by the fact that it used data from a single patient. Although the amount of data was relatively high given the more than a year follow up of this patient, this was a limitation in using the full potential of the BM approach and very far from the usual clinical practice of monitoring glucose levels.

Another notable issue – partly inherited from the limited amount of data – was that the chosen higher threshold (of the diabetic hyperosmolar syndrome, 600 mg/dl) was very far from the upper boundaries of the data (range of 25-429 mg/dl) thus the results were very likely unreliable for the higher threshold. Yet, we were able to demonstrate the theoretical possibility of calculating estimates beyond the range of the available data, which could be a very important advantage compared to the traditional metrics used to characterise blood glucose measurements. Higher sample sizes would make these estimations more reliable, despite being beyond the range of the data.

Although the lower investigated threshold (270 mg/dl) was not too close to the upper limit of the data, yet only the 1.97% of the observations were above this threshold. According to the BM model, there was a 3.47% probability for an observation to fall beyond this threshold and it's worth to note this also includes the probability of exceeding the upper limit of currently observed data. Overall, as a proof of concept analysis, it was a success and was able to achieve its goals, but the next steps also seemed very straightforward: to obtain much larger amount of data and with the exploration of the additional capabilities of EVS, conducting a more comprehensive analyses including the comparison with the traditional metrics and using non-stationary models.

4.6 Thesis group 1

Thesis 1.1

I developed a novel approach that focuses on the maximums of Continuous Glucose Monitoring (CGM) measurements rather than the variability used by the traditional metrics of diabetology with the use of extreme value statistics (EVS). I have shown using a simulated dataset that the EVS with the peak over threshold (POT) approach can be used to characterise CGM curves and produce clinically relevant metrics to describe patient level risks by estimating the probability for a new peak to exceed a certain threshold, however, the choice of the threshold might be problematic.

Related publication: [86].

Thesis 1.2

I used, for the first time, the block maxima (BM) approach of EVS to characterise CGM curves. I have proven that it can also provide clinically relevant estimates that can be used as metrics to assess patient level risks and have many advantages over POT method. The estimated probability and the estimated time spent over certain, chosen thresholds can be calculated. These thresholds and estimates can be beyond the range of the CGM measurements. The analysis was conducted on real-life dataset.

Related publications: [86, 87].

5 The EVS analysis of a clinical trial population

5.1 Introduction

After the assessment of the analysis, difficulties and results of the preliminary works, it was clear that EVS and especially BM approach is suitable for the analysis of CGM measurements. This approach is fundamentally different compared to the traditional metrics which assess glycemic variability and not the extremities, which might be more suitable to accurately capture the risk of hyperglycaemia. So the next step was a comprehensive assessment and comparison of the results and metrics of the EVS with the traditional metrics. It was obvious from the previous works that simulated data is not suitable for such analysis and a single patient's measurements means far less data than what is needed for this analysis. Additionally, these results were not generalisable at all, so it was not simply more data was needed but data from more patients was required. Moreover another important feature of EVS, the use of non-stationary models was not demonstrated. In these models any or all of the three parameters of the distribution of the extremes – thus the associated return level or probability, or time spent above a certain level – is a function of a variable such as time or in this case a potential clinical characteristics. This is somewhat similar to the approach of (linear) regression models.

5.2 T1D Exchange

The need for more data was solved through the T1D Exchange [88], using the T1D Exchange Registry which mainly connects people with type 1 diabetes with ongoing studies that they might be interested to participate thus helps with the recruitment and enables to access to specific subgroups, but it also gives access to historical datasets. These are shared as aggregate and de-identified datasets upon request which has to include a brief overview of the research proposal and aims. When my request was accepted there were just a couple datasets available but one of them, the REPLACE-BG [89] seemed the best source to work with.

5.3 REPLACE-BG trial

5.3.1 Data quantity and quality of REPLACE-BG

The dataset – after the exclusion of calibration measurements – contained 14.8 million CGMS measurements of 226 patients (median duration: 33 weeks) with type I diabetes. The sampling frequency was 5 minutes using Dexcom G4

Platinum CGM device [90] (Dexcom, San Diego, California) and CONTOUR NEXT (Ascensia Diabetes Care US, Parsippany, NJ) BGM was used for the confirmatory BGM measurements. Basic clinical data (except age due to make patients non-identifiable for the public data) were available for all patients, while aggregated baseline characteristics were available from the original publication of REPLACE-BG [89] results.

The REPLACE-BG study was a randomised controlled trial (RCT) which aimed to determine whether the routine use of continuous glucose monitoring CGM without confirmatory blood glucose monitoring (BGM) confirmation is as safe and effective as CGM used as an adjunct to BGM in adults with type 1 diabetes. The recruitment was conducted between 22nd May 2015 and 11th March 2016 and the results were published online first on 16th February 2017 in Diabetes Care (doi: 10.2337/dc16-2482) [89].

It was a non-inferiority trial, meaning that the null hypothesis which was to be confirmed or rejected was that the treatment in question is not substantially less efficient than the standard (control) treatment. Thus the new treatment is accepted as non-inferior if the confidence interval of the effect size excludes worse effect by a pre-specified, clinically acceptable level which is called the non-inferiority margin [91]. In this case the "treatment" was not really a direct clinical intervention such as medications or surgery but the above mentioned monitoring of blood glucose levels using CGM with and without the confirmatory measurements. Neither the fact that this was a randomised trial nor the interventions could not compromise the data for the planned EVS analysis. The difference between the arms might seem not too important but before December 2016, the CGM systems commercially available in the U.S. and also in the rest of the world were approved by the U.S. Food and Drug Administration (or the local authorities) for use only as adjunctive devices to information obtained from standard home blood glucose monitoring so before making an insulin dosing decision a confirmatory BGM measurement was required to check the CGM sensor glucose concentration according to the labeling of these CGM systems (but many CGM and insulin pump users were making insulin dosing decisions by CGM alone) [89]. According to T1D Exchange Clinic registry (a separate, but important part of T1D Exchange) cited by the authors of REPLACE-BG, only 26% of 999 surveyed CGM users indicated that they follow the instructions and always confirm the CGM glucose measurement with a BGM measurement before administering an insulin bolus, and 41% indicated that they dosed insulin based on only the CGM alone more than one-half of the time. Another survey has shown that 50% of the 222 CGM respondents indicated that during the night, they would

treat a CGM low-glucose alert without a confirmatory fingerstick BGM measurement and 34% would dose insulin for hyperglycaemia without a confirmatory BGM measurement [89, 92].

The regulatory decision presumably was made because the lack of accuracy of the CGM sensors was considered low, and deemed them to be inadequate for decisions of dosing insulin without BGM confirmation. But the development of sensors and their improved accuracy suggested that CGM might became sufficiently accurate to be safely implemented as a stand-alone tool for glucose monitoring and therapeutic decisions. In December 2016, the FDA expanded the indications of the Dexcom G5 sensor enabling it to replace the common fingerstick BGM testing for diabetes treatment decisions [89]. The REPLACE-BG study used a slightly older model, the Dexcom G4 Platinum CGM System [90] for both arms.

Additionally, a run-in period was conducted to ensure credible and accurate measurements and to assess the patients' willingness and ability to use the study CGM and BGM. Depending on their current CGM usage, patients might have required to do a 2 weeks long blinded CGM, when the system was configured to record glucose concentrations not visible to the participant and then a 2–8 weeks long period for CGM training. For the successful completion of the blinded phase, patients had to wear CGM for a minimum of 11 of 14 days and had to take BGM measurements an average of 3 time a day by the study BGM. During the unblinded phase they had to use CGM on 21 days or more during the last 28 days and take four or more BGM measurements on average on at least 90% of days [89].

Overall, not just the quantity of data created an opportunity to develop the EVS analysis further but also the quality of it was exceptionally good and as it was from a clinical trial, and detailed background information was available to support this. This was a great advantage compared to the preliminary works and compared to some other potential datasets of T1D Exchange as some extreme measurements and thus related findings of their assessment and analysis might raise the suspicion that these might be due to failure of the sensor, but in one arm of REPLACE-BG, and during the run-in period, confirmatory BGM measurements were taken which were also available as part of the shared dataset allowing confirmatory analysis on the reliability of the CGM measurements.

5.3.2 Outcomes of the REPLACE-BG trial

Interestingly the primary outcome used for assessing the efficacy was the time in the range of 70–180 mg/dL over the entire 26-week trial which is a very sim-

ple and not too sensitive indicator. The secondary outcomes related to measured blood glucose values were time spent above or below certain levels, area under curve 180 mg/dL, area above curve 70 mg/dL. Glycemic variability was assessed through coefficient of variation which is simply the ratio of the standard deviation to the mean. Extremities were only assessed through percentage of days with 20 or more consecutive minutes of glucose concentrations <60 mg/dL for the lower values (hypoglycemia risk) and percentage of days with 20 or more consecutive minutes of glucose concentrations >300 mg/dL for higher values (hyperglycaemia risk). Other secondary outcomes include change in HbA1c and safety outcomes.

5.3.3 Patient characteristics of the REPLACE-BG trial

The patients were relatively homogeneous due to the inclusion and exclusion criteria of the REPLACE-BG study. Patients were at least 18 years old and had type 1 diabetes for at least 1 year and were on insulin pump for at least 3 months prior to the starting of the measurements and were not using a low-glucose-suspend function [89]. From the point of view of the EVS analysis, the most important exclusion criteria included the occurrence of a severe hypoglycemic event resulting in seizure or loss of consciousness in the past 3 years or an event without seizure or loss of consciousness but requiring the assistance of another individual in the past 12 months. Diabetic ketoacidosis (DKA), also related to extreme blood glucose values, was part of the exclusion criteria too. This condition occurs when the body starts to run out of insulin and ketones build up, which can be life-threatening without prompt treatment [29, 30]. Patients with more than 1 episode of DKA in the past year were excluded, and it is also known that no patient had DKA during the study.

Another important aspect of the trial population was their general kidney function as poorly controlled diabetes can cause damage to small blood vessels especially in the eyes and kidneys. In the REPLACE-BG, patients with an estimated glomerular filtration rate (eGFR) of <30/min/1.73 m^2, an indicator of the kidney function, estimating how much blood passes through the glomeruli each minute, from a measurement obtained within the prior 12 months as part of usual care were excluded, as were kidney transplant patients. During the whole follow up 7 participants were hospitalized for a total of 8 times, including a single surgery; none of these were related to glucose metabolism [89].

During the study no serious events occurred that could have realistically led to renal dialysis or mannitol (a drug used to treat acute kidney failure) administration [89].

We had no information on the actual insulin pump usage, but due to the inclusion criteria of REPLACE-BG that patients not just had to use an insulin pump for insulin delivery for at least 3 months prior the trial, but they were also required to have no plans to discontinue pump use during the following 8 months after consent so it can be assumed that no long-term lack of pump usage occurred during the measurements.

The study population was middle-aged with the mean age of 44 years (SD = 14), dominantly white (92%) with 50.5% males and on average 23.7 years of diagnosed type 1 diabetes and their mean HbA1c was 7.0 (SD = 0.6) at baseline. 47% were current CGM user, 35% used CGM previously and 18% have never used before.

Overall these patients seemed ideal for EVS analysis as they were relatively high risk patients where some extreme measurements could be expected.

5.4 CGM measurements

The graphical assessment of the whole dataset brought the first interesting finding. As shown of Figure 7 which is the histogram of all CGM measurements, a skewed distribution can be observed with two distinct ”spikes” at the two ends of the graph with no CGM measurements appearing beyond these spikes. It seemed that the data is truncated at these levels and data is saturated there so if there was a measurement above 400 mg/dL it may appear incorrectly as 400 mg/dL instead. This indicated that there might have been lower and upper detection limits present leading to the loss of information. This was not mentioned in the REPLACE-BG paper, but after consulting with the Dexcom G4 Platinum's user's manual [90] it has been confirmed that the lower detection limit is 40 mg/dL and the upper detection limit is 400 mg/dL. Unfortunately, no literature was found that investigates the impact of the presence of a lower or upper detection limit on GV metrics.

5.5 Validation and accuracy of CGM measurements

Although the manual confirmed that there are detection limits it did not tell anything about how potential measurement above or below these are accounted. It seemed that these are recorded as 40 or 400 mg/dL depending which limit was exceeded causing the saturation (spikes on the histogram) at these levels, although we could not prove this. Fortunately, there were confirmatory BGM measurements that were available to use. This was very important for the EVS analysis as this

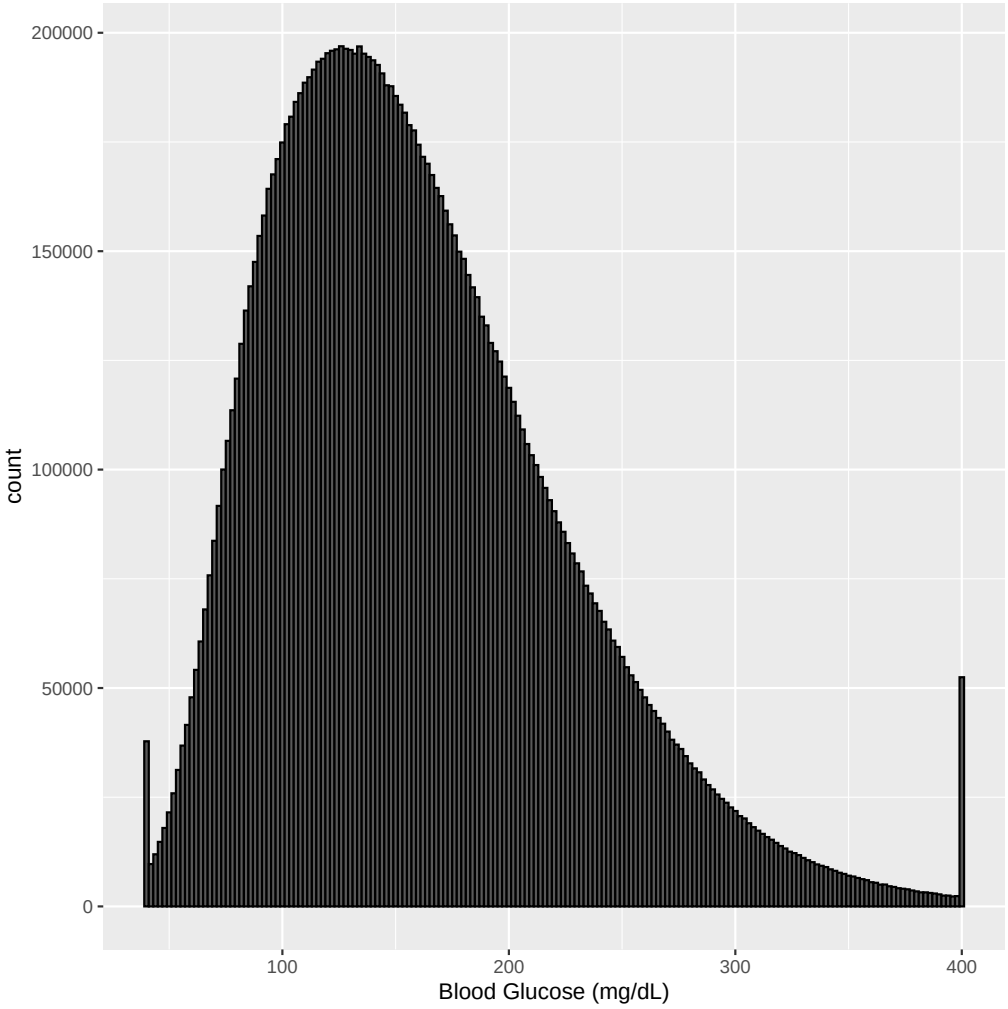

Figure 7: Histogram of all CGM measurements

anomaly was detected at the highest and lowest measurements that had the most important role, so it was very important to investigate this phenomenon.

First, CGM measurements were collected that are over 300 mg/dL and a corresponding BGM measurements within ±2.5 minutes time frame was available. Fortunately, this was still a relatively large sample with 15,965 measurement pairs where the mean CGM was 339 (SD=32) mg/dL and the mean BGM was 331 (SD=57) mg/dL. These were plotted against each other (Figure 8) where the red line represents the equality between the two measurements. It can be observed that BGM and CGM values randomly and somewhat symmetrically scatter around the equality line with few outliers. Lower BGM outliers are more frequent which explains why the mean BGM is somewhat lower compared to CGM which is limited by its upper detection limit; this also explains the CGM's much lower standard deviation. Due to the upper detection limit of the CGM, thus its truncated distribution, their

Spearman's rank correlation coefficient was calculated and statistically significant correlation was found between this subset of CGM and BGM measurements with $\rho = 0.606$ and $p < 0.001$.

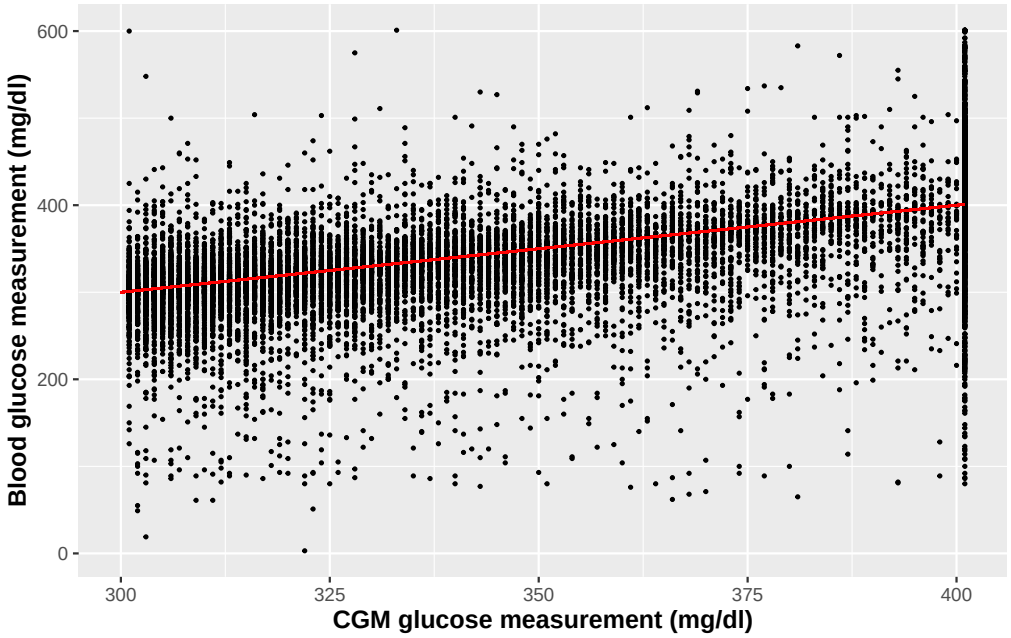

Figure 8: Pairwise comparison of BGM and CGM measurements above 300 mg/dL

The upper detection limit can be also easily seen in Figure 8, as the maximum value of CGM in the plot was 400 mg/dL, around which a lot of BGM measurements saturated. This graph alone was not enough to adequately assess the whole picture, as the saturated part had too many observations where all CGM measurements have a value of 400 mg/dL and the BGM measurements scatter around it. Those measurement pairs (N= 1,603) were assessed separately as well by plotting the histogram of the BGM measurements Figure 9. It shows that these also randomly scatter around 400 mg/dL with a roughly symmetrical distribution with the majority of the measurements rallying around a distinct peak at 400 mg/dL. The mean of these BGM measurements was 403 mg/dl (SD=80) while the median was 407 mg/dl. It can be also observed in Figure 9 that the BGM measurements at exactly 400 mg/dl are relatively high and the other higher round numbers (400 and 500 mg/dl) are also outlying which is suspicious and could indicate recording bias. According to the study protocol, CONTOUR NEXT BGM was used which has a detection range of 10 mg/dL – 600 mg/dL. [93] It is also odd that quite low BGM measurements, even under 300 mg/dl also occurred.

If those CGM measurements result from a malfunction that would not affect

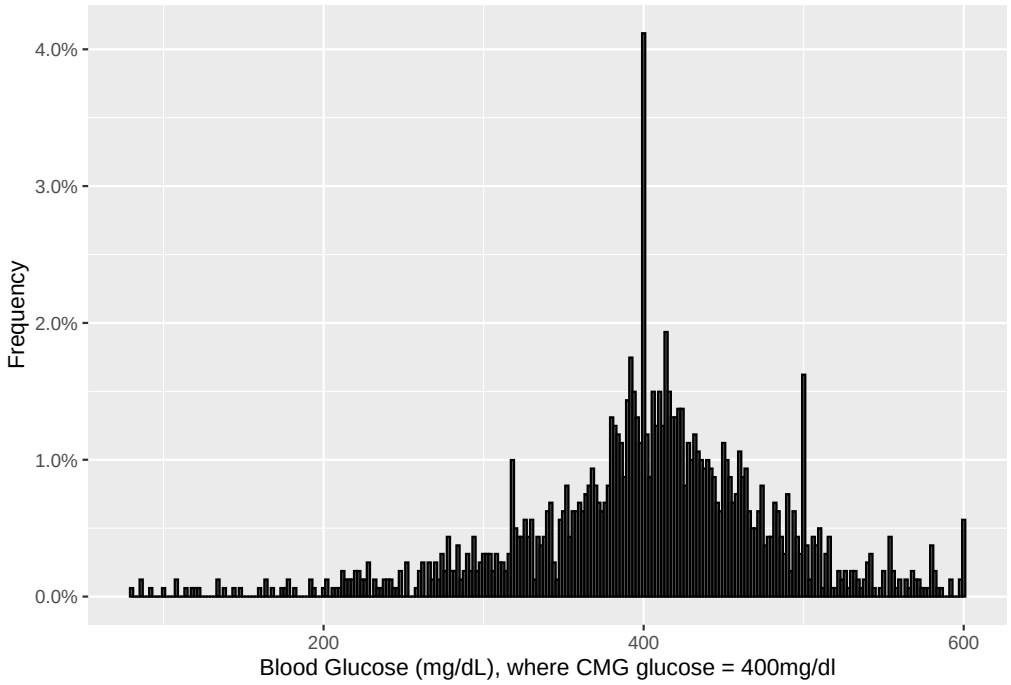

Figure 9: Histogram of BGM values where the CGM was 400 mg/dl

the BGM measurements so the results would not scatter around the CGM value, thus this finding rules out the possibility of a systematic error with high certainty. Additionally it was also confirmed that saturation at 40 mg/dL and 400 mg/dL are due to measurements exceeding detection limit are recorded as they were measured exactly 40 mg/dL or 400 mg/dL.

5.6 Classical metrics

5.6.1 Standardised ranges

The next step of the analysis was to calculate the values of each classical metric for the patients. The most common metrics were used which are briefly described in this chapter.

From the mathematical point of view, the most simple group is where the percentage of time spent in a certain range is given. These are standardised ranges which are part of a group of other factors to describe CGM results which included more general characteristics such as the number of days CGM worn or mean glucose. These were widely accepted as the International Consensus on Use of Continuous Glucose Monitoring [94] in 2017. These were amended by Battelino et al. [26] in 2019 and the initial 14 core metrics the panel selected were narrowed to 10 metrics that may be most useful in clinical practice. Five standardised blood glucose ranges were defined from which the middle was defined as the target range. These were the following:

- Time above range (TAR): % of readings and time >250 mg/dL (>13.9 mmol/L)
- Time above range (TAR): % of readings and time 181–250 mg/dL (10.1–13.9 mmol/L)
- Time in range (TIR): % of readings and time 70–180 mg/dL (3.9–10.0 mmol/L)
- Time below range (TBR): % of readings and time 54–69 mg/dL (3.0–3.8 mmol/L)
- Time below range (TBR): % of readings and time <54 mg/dL (<3.0 mmol/L)

Another commonly used but simple metric is the interquartile range (IQR) which means the same in this context as it does in statistics which is the range of middle half of the data set once all elements ordered from low to high (i.e., difference between the lower and upper quartile) [95].

5.6.2 Coefficient of variation

Additionally, both list contains a factor specifically named as "Glycemic variability". This is a broader term but in these cases, as it simply means coefficient of variation (CV). CV is the ratio of the standard deviation (SD) to the mean, often expressed as a percentage [96]. Thus higher CV means less well controlled GV. For

the sake of clarity I will use the term coefficient of variation for the metric and glycemic variability in its broader meaning.

5.6.3 MAGE (Mean Amplitude of Glycemic Excursions)

The slightly more complex mean amplitude of glycemic excursions (MAGE) metric was introduced by Service et al. in 1970 [58] and is still one of the most important index for glycemic variability assessment and treated as a key reference for blood glucose controlling at clinical practice [97]. 'Glycemic excursion' is a general term for phases where the blood glucose level is relatively high or low. For MAGE, the threshold limit for this is the standard deviation of a 24 hours long period. These excursions eventually reach a peak or nadir and the difference of the peak or nadir and the starting point of the excursions gives its amplitude ("height") and MAGE is the arithmetic mean of these amplitudes [97]. It is interesting to note that in their original paper, Service et al. used 14 patients only (3 non-diabetic, 3 stable diabetic and 8 unstable diabetic) with 48 hours of glucose measurements with 5 minutes frequency and used rank sum test to compare the calculated MAGE values between the above mentioned patient groups [58]. For a long time, MAGE had a very efficient practical advantage that clinicians and researchers could assess CGM results by visually inspecting the glucose profiles and use a "ruler and pencil" graphical approach to calculate MAGE. Obviously this kind of manual approach is time-consuming and error-prone [97]. It is also interesting to note that despite its relative simplicity, computerised MAGE calculator algorithms are still in development in the present [98].

5.6.4 CONGA (Continuous Overall Net Glycemic Action)

As it was mentioned in the Introduction (Chapter 2), in the past decades, CGMS sensors and hardware developed rapidly and became cheaper, more accurate and reliable, allowing observations of longer periods, producing much more data and the capability to assess longer term trends and variability, creating the need for methods to fully utilise the value of the provided information by more modern CGMS systems. In 2005 McDonell et al. presented a new metric called continuous overall net glycemic action (CONGA) to address this which became and still commonly used [97, 99, 100]. Its calculation is also relatively simple: to calculate CONGA(n), for each observation after the first n hours of observations, the difference between the current observation and the observation n hours before is calculated. CONGA(n) is defined as the standard deviation of these differences. The choice of the time interval

depends on the clinical or research question being addressed and of course on length of observations. This makes CONGA more flexible and enables to characterise within day and between day glycemic variability as well. Similarly to the other metrics, higher CONGA values indicate greater glycemic variability and increased glycemic excursions and lower CONGA values reflect more stable glycemic control. In their original paper, McDonell et al. used 72 hours long CGMS measurements of 10 adult healthy volunteer controls and 10 randomly selected childen's measurements who had type 1 diabetes to demonstrate the new metric. CONGA(1), CONGA(2) and CONGA(4) were calculated but no formal statistical comparison was conducted to compare the results between the groups [101].

5.7 The novel EVS model

As it was mentioned previously, even without the calibrational measurements of the run-in phase, the REPLACE-BG dataset contained 14.8 million CGMS measurements of 226 patients (median duration: 33 weeks). BM models were fitted to the dataset with the block maximum calculated with the previously used block size of 12 measurements (which meant hourly maxima of the twelve 5 minutes long periods). After calculating the hourly maxima, the secondary sample was still rather large with 1.23 million observations. The next step was to fit a block maxima model for each of the 226 patient's CGM record separately so all 3 parameters of the generalized extreme value distribution were estimated individually. This was a computationally intensive calculation, so the MTA – later ELKH – Cloud (https://-cloud.mta.hu/) was used, which is an infrastructural cloud computing service, originally developed by the Institute for Computer Science and Control of the Hungarian Academy of Sciences[1]. R statistical program package version 4.1.0 [81], ExtRemes 2.1 [82], gluvarpro 4.0 [84] and iglu 3.2.2 [102] packages were used.

The median and the range (in brackets) of all parameters for the 226 fitted models were the following: shape $\widehat{\xi}$: -0.077 (-0.432, 0.107); location $\widehat{\mu}$: 150.2 (94.6, 251.5) and scale $\widehat{\sigma}$: 52.3 (15.4, 100.3). The distribution of the parameters of the fitted models for each patient can be found in Figure 10. All three parameters had a symmetrical, bell shaped distribution with relatively long tails for the shape parameter ($\widehat{\xi} \lessapprox -0.2$) and location parameter ($\widehat{\mu} \gtrapprox 200$) mostly attributed to the same 4 outlying models which can be easily identified on the right side of the middle-

[1]On behalf of the KOMPLEXEPI project we are grateful for the possibility to use ELKH Cloud (see Héder et al. 2022; https://science-cloud.hu/) which helped us achieve the results published in this paper.

left and bottom left-graphs. These were patient 134 ($\xi = -0.31; \mu = 221; \sigma = 45.2$), patient 186 ($\xi = -0.43; \mu = 251; \sigma = 100$), patient 190 ($\xi = -0.29; \mu = 221; \sigma = 87$) and patient 239 ($\xi = -0.18; \mu = 204; \sigma = 69$). (It's worth to note here that the original REPLACE-BG IDs were used, and despite the total number of patients being 226, their ID's were not continuous and went from 2 to 293 with intermittent scarcity.)

It can be seen that the majority of the fitted distributions were Weibull distribution ($\xi < 0$) but some followed Frechet distribution ($\xi > 0$).

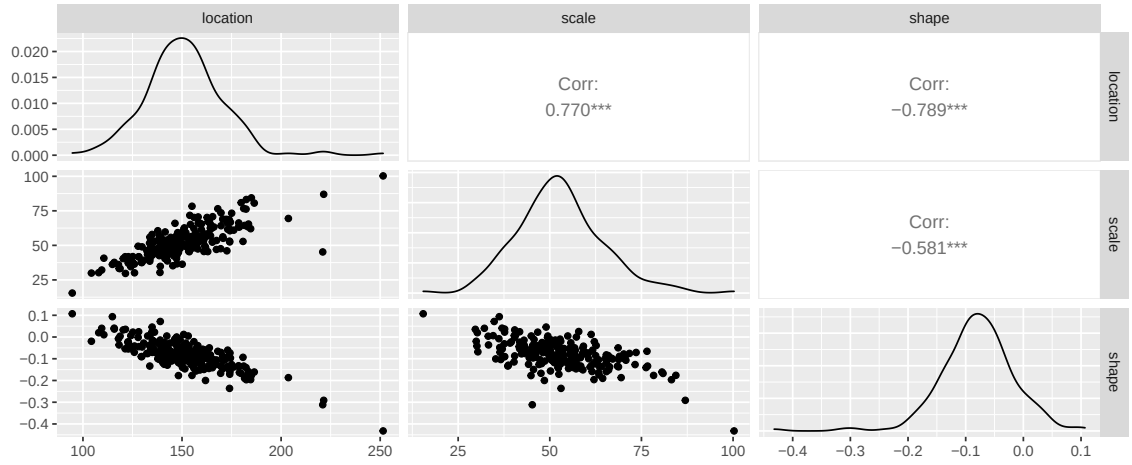

Figure 10: Matrix plot of model parameters of each patient's fitted model

Interestingly the parameters correlated with each other, moreover, some of them strongly. Their Pearson correlation coefficients were positive ($r = 0.770$) between the scale and location, a strong negative correlation was with $r = -0.789$ between shape and location and also negative, but moderate correlation between shape and scale with ($r = -0.581$) were observed. All of these correlations were statistically significant with $p < 0.001$.

5.8 Results

Using the BM EVS approach, 1 year return levels and probabilities for exceeding and expected time spent above 400 mg/dL and 600 mg/dL over a year were calculated similarly to the previous preliminary analysis using the BM approach [87]. Of the traditional metrics, the standardised ranges (from which the highest range, the "time above range (TAR) >250 mg/dL" was in the focus of interest), interquartile range (IQR) and coefficient of variation (CV) were calculated for the whole follow up period for each patient. For Mean Amplitude of Glycemic Excursions, the

mean of the patient's daily MAGE values were calculated, in order to summarise these in a single value for the whole follow up. For the Continuous Overall Net Glycemic Action (CONGA), the mean of the daily CONGA(24) values was calculated meaning that for each measurement after the first 24 hours, the difference between the current measurement and the measurement 24 hours before were calculated and the standard deviations of these for each day was obtained (referred to as CONGA for the rest of the analysis), similarly to MAGE. The summary of these results can be found in Table 2 and their comparison is visually demonstrated in Figure 15, displaying their distribution individually, comparing them pairwise using scatter plots and showing their pairwise Pearson correlation coefficients and also highlighting the 9 highest risk patients according to the estimated time above 600 mg/dl in the scatter plots. A similar plot with the comparison with the standardised ranges are presented in Figure 17.

Table 2: Summary results

Metric	**Mean**	**SD**	**Median**	**Min**	**Max**
1 year return level (mg/dL)	496	76	495	332	739
Hours above 600 mg/dL	0.331	1.026	0.007	0.000	7.797
Hours above 400 mg/dL	49.4	79.2	23.7	0.0	786.2
TAR >250 mg/dL (%)	9.68	7.41	8.00	0.03	52.7
IQR (mg/dL)	82.3	18.9	80.5	22.0	148.0
CV	0.374	0.047	0.372	0.206	0.489
MAGE	147.8	28.3	148.7	54.6	227.0
CONGA	38.4	6.78	39.0	17.9	62.8
TAR 181-250 mg/dL (%)	23.42	7.02	24.25	0.06	53.20
TIR 70-180 mg/dL (%)	63.12	12.22	63.16	17.57	96.91
TBR 54-69 mg/dL (%)	2.80	1.71	2.62	0.00	10.17
TBR <54 mg/dL (%)	0.96	0.90	0.76	0.00	6.5

5.8.1 1 year return levels

1 year return level means the maximum which occurs exactly once on average during the given time period which is a year in this case (higher values occur less frequently while lower values more frequently). These estimates, together with their 95% CI can are shown in Figure 11. The IDs of the 17 patients whose point estimate was above 600 mg/dl (the threshold for the above mentioned diabetic hyperosmolar syndrome, marked with red dashed line on the plot) were highlighted; they can be

considered to be at very high risk. The relatively narrow confidence intervals indicate reasonably precise estimates. Interestingly the vast majority of 1 year return level estimates were higher than the CGM's upper detection limit (400 mg/dl). The confidence intervals tend to be narrower for the lower 1 year return level estimates and wider for the higher estimated values, which is not unexpected as these estimates are further beyond the range of the actual observed data limited by the detection limit. The maximum of the estimates (739 mg/dl) is way beyond the detection limit but a clinically possible value [45].

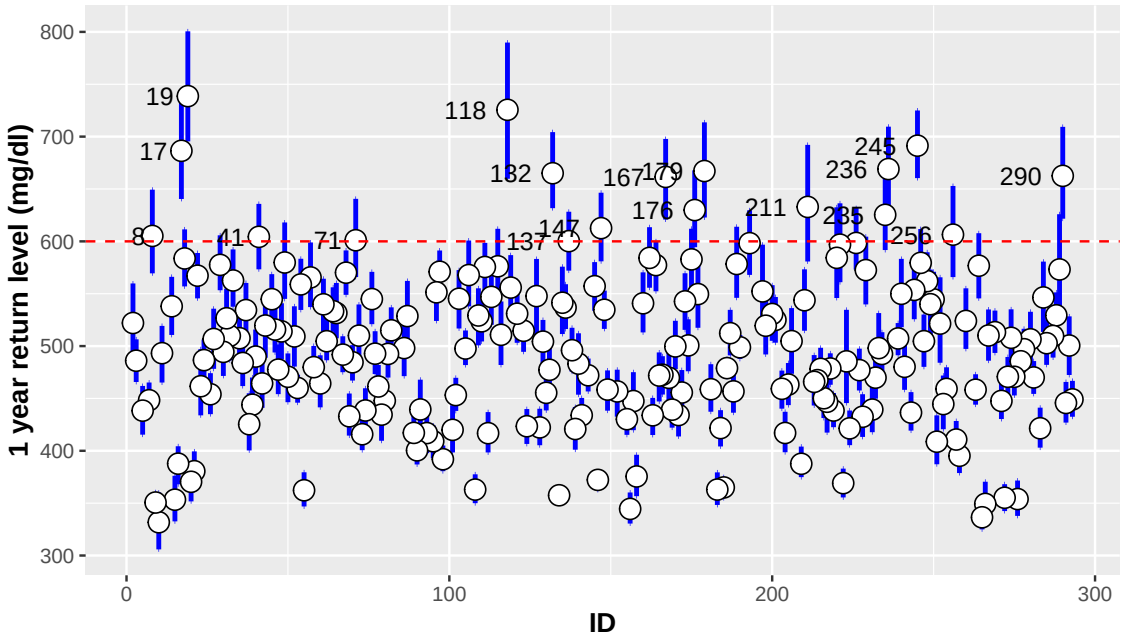

Figure 11: One year return level and its 95% CI for each patient. The ID's of the 17 patients whose point estimate was above 600 mg/dl are highlighted.

5.8.2 Time spent above 600 mg/dl

With the other, new EVS metric the investigated threshold 600 mg/dl, which is the previously highlighted level for the diabetic hyperosmolar syndrome. The mean was 0.331 hours (SD=1.026) and the median was 0.007 hours, while the maximum was 7.797 hours. As it can be seen in Figure 15 and Figure 12, the expected time spent above this over a year was zero for the vast majority of the patients and for many, it was just slightly above it. Therefore in general, the time spent above 600 mg/dl shown weak correlation with the other metrics with the highest correlation of $\rho = 0.643$ with the 1 year return level.

In Figure 12 the IDs of patients with more than 2 hours are highlighted. These

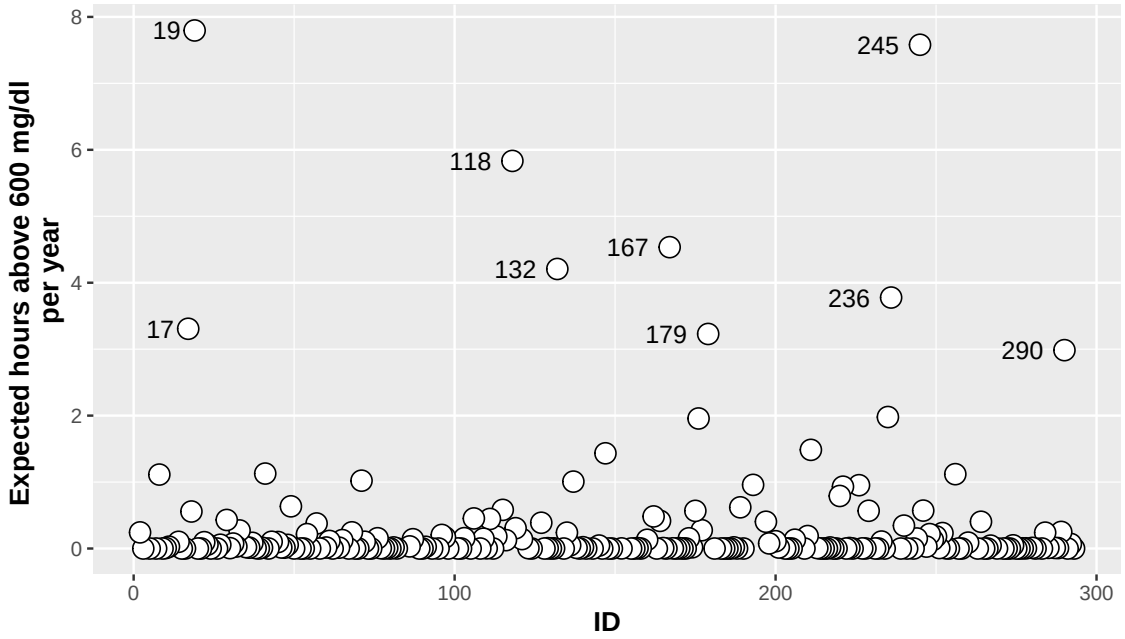

Figure 12: Expected hours (EVS) spent above 600 mg/dl in a year. IDs where this is above 2 hours were highlighted.

9 patients can be considered at very high risk as diabetic hyperosmolar syndrome leads to coma for 25-50% of the patients [33] thus spending hours above this level can lead to acute life-threatening condition. As it can be also seen in Figure 15, where these patients are highlighted with light blue colour, these patients got the highest results only in 1 year return level which explains the relatively strong correlation compared to the other metrics. With the rest of the metrics, they are roughly in the top third in CV but had only moderate scores in percentage time above range (TAR) >250 mg/dL, CONGA, IQR and MAGE. The measurements of these patients were further investigated and the histogram of their measurements are individually plotted in Figure 13 where their IDs and their estimated number of hours spent above 600 mg/dl can be also seen. Similarly to the same histograms for 400 mg/dl it can be seen that these patients' CGM measurements (or at least 8 of 9) where heavily affected and "trimmed" by the upper detection limit of the CGM sensor. The similar phenomenon was observed when more patients' CGM results were investigated in a similar manner of the patients with the highest estimated values. As these patients seem to be at the highest risk, this loss of relevant and valuable data is concerning.

5.8.3 Time spent above 400 mg/dl

Using the EVS model an estimation can be made for the probability of exceeding certain thresholds thus the average proportion of time spent above these thresholds. Given the threshold and the duration which we want to make an estimation, the expected time can be calculated. In order to produce clinically meaningful and interpretable results a year-long period was chosen to be consistent with the return level, with two thresholds: 400 and 600 mg/dl. This means that these results are extrapolated beyond the length of the follow-up and beyond the observed range which was limited by the detection limit.

The mean time spent above 400 mg/dl was 49.4 hours (SD=79.2) with the maximum of 786.2 hours and the median was 23.7 hours. The results of each patient are shown in Figure 14 highlighting those patients whose estimate was above 200 hours per year (N=10). As it can be seen in Figure 15 and in Figure 14 although the expected number of hours above 400 mg/dl is zero for many patient, the peak of its distribution is slightly above that.

The time spent above 400 mg/dl strongly correlated with TAR (>250) ($\rho = 0.851$), MAGE ($\rho = 0.722$) and with IQR ($\rho = 0.771$) and it had moderate correlation with return level ($\rho = 0.506$), CONGA ($\rho = 0.472$) and CV ($\rho = 0.444$). Interestingly, it only had quite low correlation with the time spent above 600 mg/dl ($\rho = 0.363$), mainly due to the high number of patients with zeros in the latter and due to the fact the outliers with the highest estimated time above 400 mg/dl such as patient 186 and 190 had very low values for the time above 600 mg/dl which makes sense and both of them were amongst the outliers of the parameters. But in general the results show that a patient can spend more time above 400 mg/dL but less above 600 mg/dL or the other way around compared to the others and these metrics are not interchangeable. All of the above mentioned correlations were statistically significant with $p < 0.001$.

Patients with the highest estimates were further investigated. The histograms of the CGM measurements of the top 12 patients with their IDs and their estimated hours per year above 400 mg/dl are presented in Figure 16. It can be seen that the measurements of these patients were heavily affected by the upper detection limit of the CGM sensor. Also worth to note that how did this extremely affected patient 186 (bottom left in the figure) who had the highest estimated number of hours above 400 mg/dl. The same can be observed in Figure 13 where similarly, the histograms of the CGM measurements were presented for patients with the highest estimates of the other EVS metric, the estimated time above 600 mg/dl.

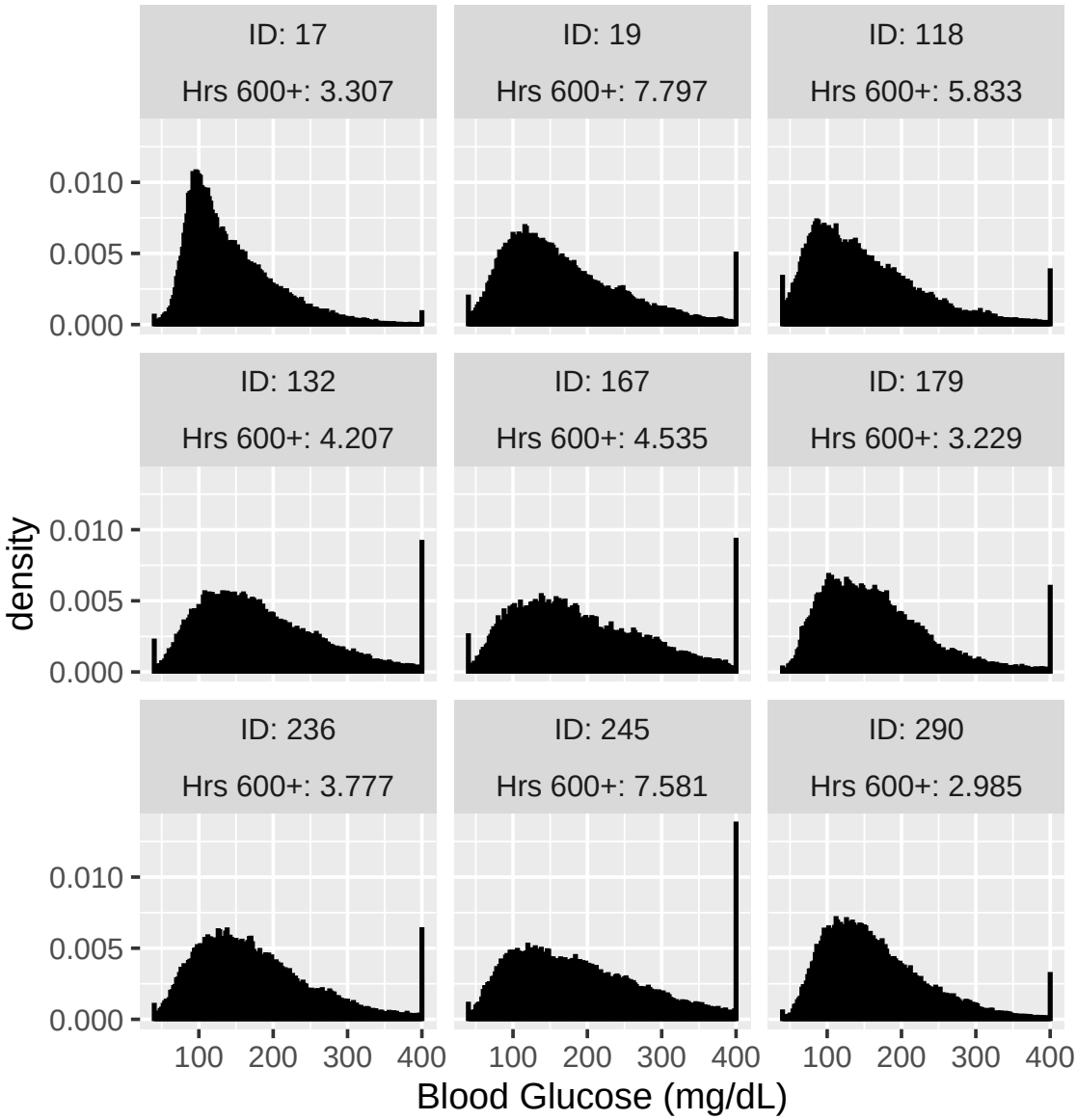

Figure 13: Histograms of the patients with the highest estimated time above 600 mg/dl per year

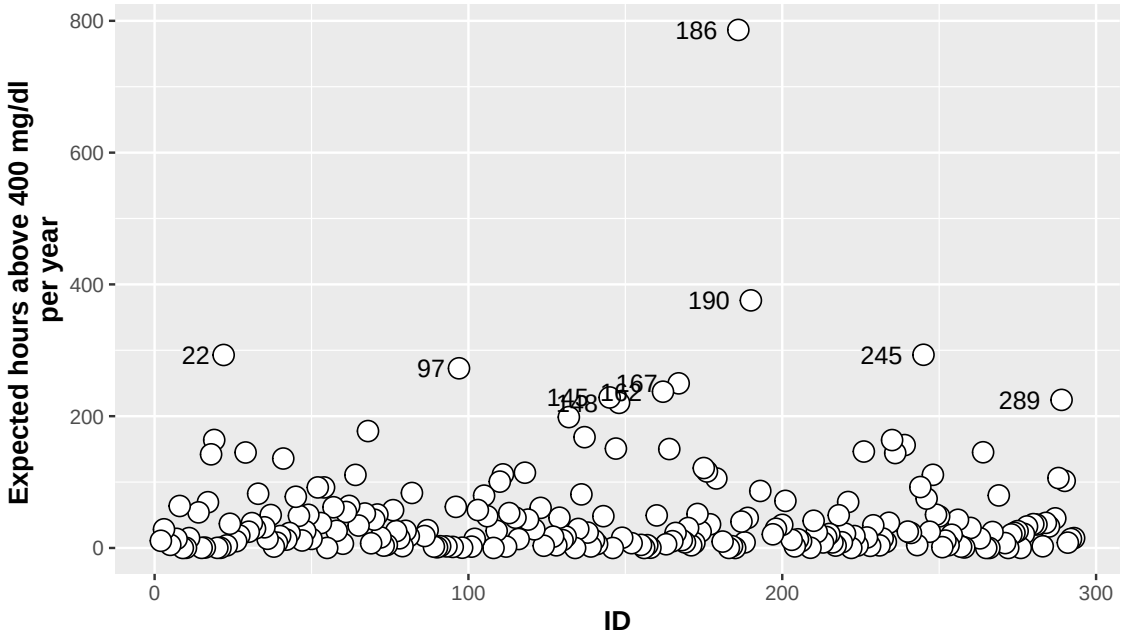

Figure 14: Expected hours (EVS) spent above 400 mg/dl in a year. IDs where this is above 200 hours were highlighted.

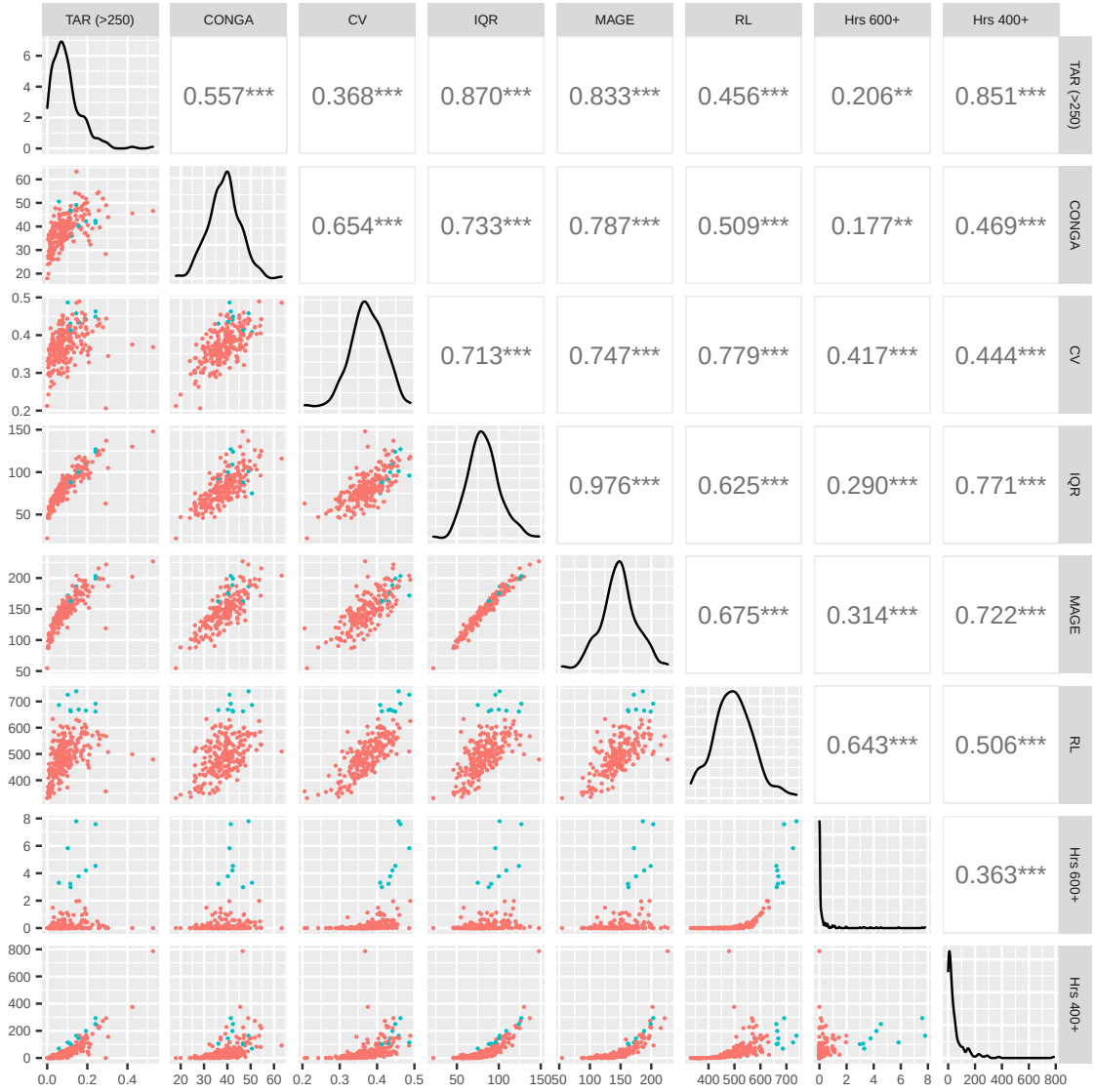

Figure 15: Pairwise scatterplots, distribution and linear correlation coefficients of the investigated metrics. Distribution of each metric can be found in the main diagonal, pairwise correlation coefficients in the upper right triangle and their pairwise scatterplots in the bottom left half. The 9 highest risk patients according to the estimated time above 600 mg/dl obtained with EVS are highlighted.

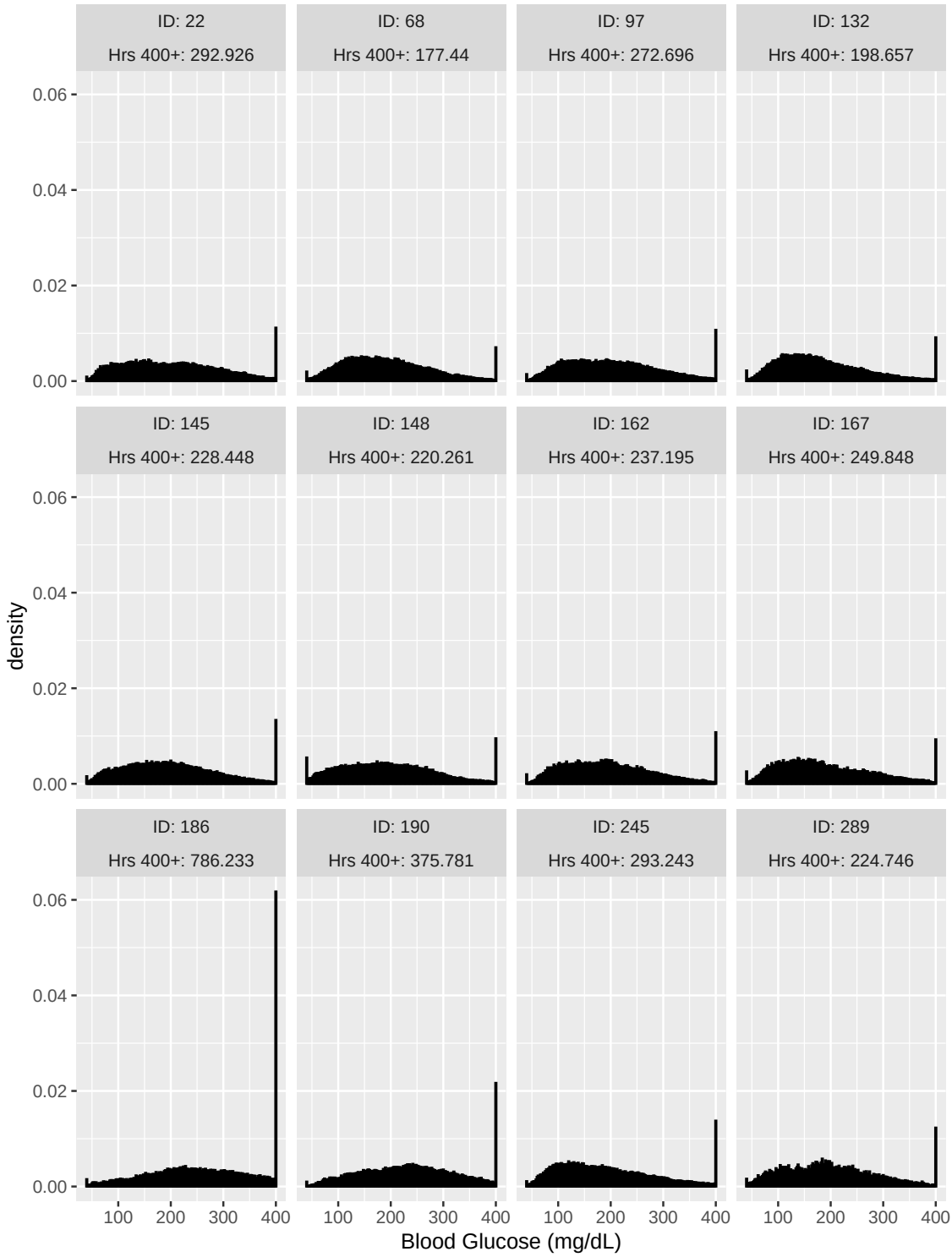

Figure 16: Histograms of the patients with the highest estimated time above 400 mg/dl per year

5.8.4 Standardised ranges

The standardised blood glucose ranges accepted as the International Consensus on Use of Continuous Glucose Monitoring [94] in 2017 and amended by Battelino et al. [26] in 2019 were calculated. Their main descriptive statistics are presented in Table 2 and their comparison with 1 year return level and the expected number of hours spent above 400 mg/dL can be found in Figure 17. In general, it can be said that this patient group on average spent almost 2/3 of the time in the target range TIR (70−180) and very little time below this. The rest of the time is roughly split in a 1/3 - 2/3 ratio between the top two ranges with little less than 10% in the highest range above 250 mg/dL on average. However, the patients show relatively large difference in time spent in these ranges. As it was expected, TAR (>250) strongly correlated ($\rho = 0.904$) with the time spent above 180 mg/dL as these ranges overlap. Additionally TAR (>250) strongly correlated ($\rho = 0.833$) with MAGE as well. What is more interesting is that TAR (>250) had a very strong negative correlation ($\rho = -0.923$) with TIR (70−180), which is considered as the target range for blood glucose levels, while it had moderate correlation with its neighbouring range TAR (181−250) $\rho = 0.614$). Comparing with the EVS metrics, TAR (>250) strongly correlated ($\rho = 0.851$) with the expected hours spent above 400 mg/dL. As it can be seen in the bottom left scatter plot, there were two outliers in both metrics strengthening the linear relationship. With the 1 year return level, TAR (>250) only shows a borderline moderate correlation ($\rho = 0.456$). In the scatter plot comparing these two it can be seen that there is some (linear) relationship between the two metric for the majority of the observations, but the highest values of each have only moderate levels of the other metric. All of these correlations were statistically significant with $p < 0.001$.

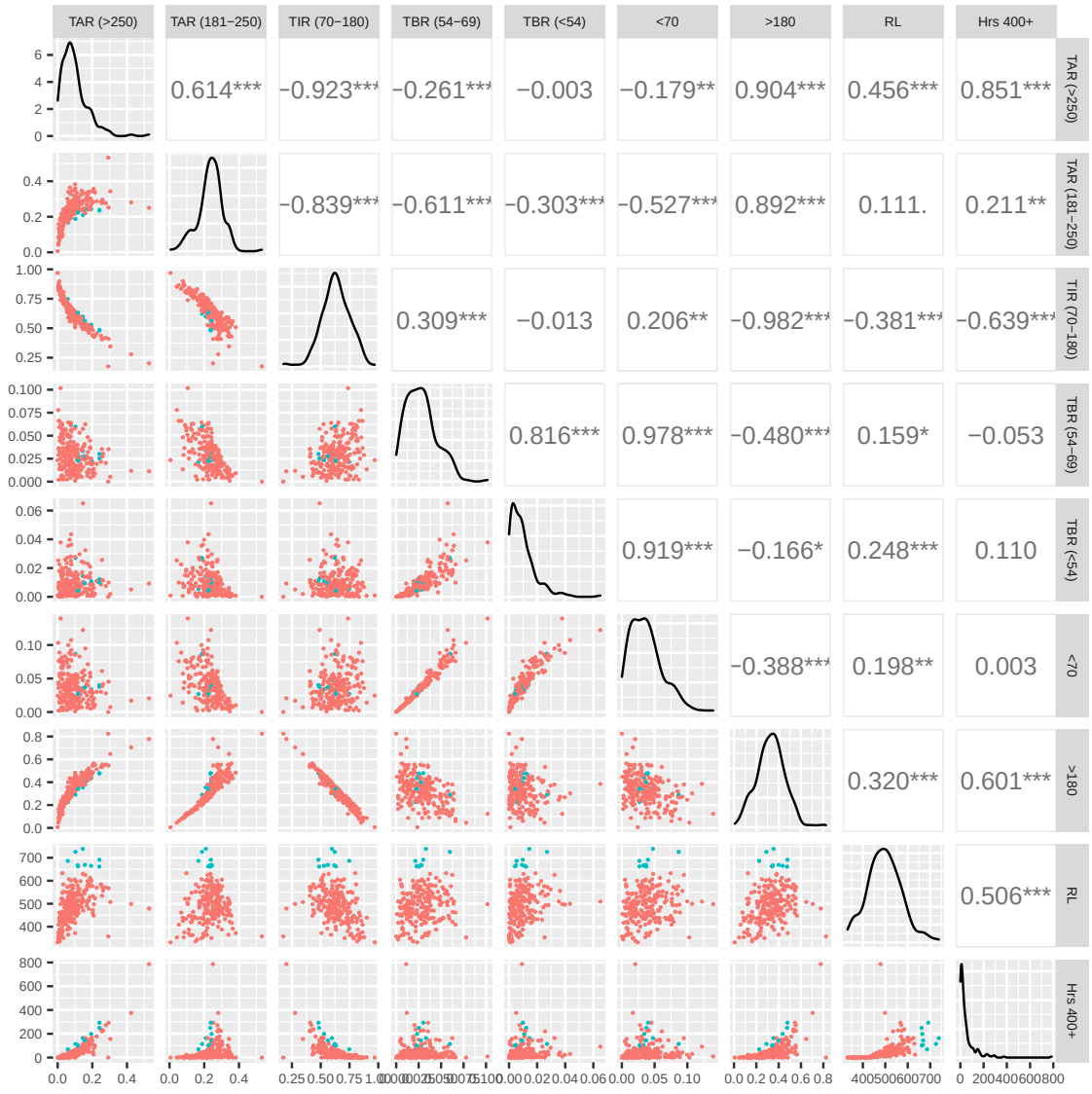

Figure 17: Pairwise scatterplots, distribution and linear correlation coefficients of the investigated metrics and percentage of time spent in standardized clinical ranges (in mg/dl). Distribution of each metric can be found in the main diagonal, pairwise correlation coefficients in the upper right triangle and their pairwise scatterplots in the bottom left half. The 9 highest risk patients according to the estimated time above 600 mg/dl obtained with EVS are highlighted.

5.8.5 Results of MAGE, CONGA and CV

The summary results of both MAGE and CONGA can be also found in Table 2. The investigation and the sensibility check of MAGE results let to a very interesting finding, namely the discovery of an important error in the used "gluvarpro" package which is presented in detail in Chapter 5.9. The mean MAGE was 147.8 (SD=28.3) very close to the median which was 148.7, indicating a symmetrical distribution which can be also seen in Figure 15, while the minimum was 54.6 and the maximum was 227.0. The mean and median of CONGA and CV were also very close. For the CONGA the mean was 38.4 (SD=6.78) and the median 39.0, while its minimum was 17.9 and the maximum 62.8. While for the CV the mean was 0.374 (SD=0.047) and the median 0.372, while its minimum was 0.206 and the maximum 0.489.

As it was expected the main two traditional metrics strongly correlated with each other. The Pearson correlation coefficient between the MAGE and CONGA was $\rho = 0.787$. The CV had also strong correlation with MAGE ($\rho = 0.747$) and moderate correlation with CONGA ($\rho = 0.654$). What was more interesting that a surprisingly strong correlation was found between the MAGE and IQR ($\rho = 0.976$), making them almost interchangeable. Their strong linear relationship can be also seen in the relevant sub-plot in the main diagonal of Figure 15. The correlations between these traditional metrics and the EVS metrics were investigated and presented above in the relevant chapters of the results of the EVS metrics.

5.9 Discovery of an error in gluvarpro 4.0 package

While I wrote this thesis I discovered an error in the gluvarpro 4.0 [84] package which had far-reaching consequences since this package is relatively widespread for the analysis of glycemic variability and the discovery of this error affected the primary results of at least two published clinical trials.

The scope of this error was the calculation of the MAGE (and the related MAGE+ and MAGE- scores) and was not discovered by the time my original publication appeared, as the focus was on the correlation between different metrics, not on the actual values of the traditional metrics. Careful analysis however later revealed that the MAGE values I presented are suspiciously low: in the original calculation the mean was 9.62 (SD=3.66) which is roughly 10 times smaller then what would be expected for such patient population and 2-3 smaller even compared to patients without diabetes. As their findings were summarised in the original MAGE paper: "MAGE was small for normals (range, 22 to 60 mg./100 ml.), larger for stable diabetics (67 to 82 mg./100 ml.), and largest for unstable diabetics (119

to 200 mg./100 ml.)" [58]. Additionally, the reference paper in the help file of gluvarpro's command for MAGE (called "magevp") had similar findings in the range of 80-90 [103], all of these in the same units (mg/dL or mg/100 ml). Looking at the values closely, I noticed that reason for such low mean is that for many days the daily MAGE was 0 which is only possible if exactly half of the measurements were precisely above the mean by the same amount as the other half were below it which is practically impossible. The rest, where it was not 0, seemed realistic then. Even running the command on the example dataset of the gluvarpro package itself resulted similarly low results with zeros for some of the daily MAGE value. In order to understand the reason of the error, it is important to note that by definition MAGE is the arithmetic mean of the amplitudes of glycemic excursions exceeding a certain treshold which is usually the standard deviation of the CGM values over a 24 hours long period. Instead of this, the code works as the following.

In line 75, it calculates the difference between the current and the previous measurements, this results in the change:

```
for (j in 1:(dim(aux)[1] - 1)) {
    diff <- c(diff, as.numeric(as.character(aux$glucose[j])) -
    as.numeric(as.character(aux$glucose[j + 1])))
    }
```

where "aux" is just a subset of the data containing the observations for the current day. Following this section it separates the changes into two groups: increases and decreases and then keeps only the ones that have a higher absolute value than the daily standard deviation:

```
diff.up <- diff[diff >= 0]
diff.up <- diff.up[abs(diff.up) > n * sd]
diff.down <- diff[diff < 0]
diff.down <- diff.down[abs(diff.down) > n * sd]
```

The problem is that it does not match with the definition and it is a very unlikely event as the frequency of the measurements are usually 5 minutes and it is very rare that in such a short time, the change of blood glucose level would be greater than the daily standard deviation. It is more likely that cases when the difference between two consecutive measurements is higher than the daily standard deviation arises due to CGMS stopping measuring for a while and thus the time between the two consecutive measurement is much longer than 5 minutes. Looking at the raw data I found examples of this when a gap in the measurements lead to high differences between two neighbouring measurement. Otherwise, and that

means for most of the time, the code above produces no observations thus the result in R would be a zero length vector. If that was the final result it would have been obvious to spot the error as MAGE should exist and be a finite calculable number, but the following piece of code replaces this with 0, masking the problem:

```
if (all(is.na(diff.down)) || length(diff.down) == 0) {
    diff.down <- 0
}
if (all(is.na(diff.up)) || length(diff.up) == 0) {
    diff.up <- 0
}
```

And in the end, the mean of these observations, which are in most cases 0, in the other cases there are only one or maybe two in this vector, gives us the false MAGE, MAGE+ and MAGE-:

```
x$lmage[position] <- round(mean(abs(diff.down),
                     na.rm = TRUE), digits = 2)
x$hmage[position] <- round(mean(abs(diff.up),
                     na.rm = TRUE), digits = 2)
x$mage[position] <- round(mean(c(x$hmage[position],
                     abs(x$lmage[position])),
                     na.rm = TRUE), digits = 2)
```

In summary, instead of peaks in the CGM built up over time through multiple measurements, this code used the change between the directly neighbouring two measurements which are normally 5 minutes apart, thus realistically the magnitude of change is very limited. These were then compared to the daily standard deviation which is also wrong.

After investigating this matter further, I discovered the likely source of this error. The answer lies in the paper referred in the help page of the "magegvp" command. This paper, published in the Cyprus Journal Of Medical Sciences in 2016 states the following:

"Standard deviation values were calculated using all the measurements for each patient. Each measurement value was subtracted from the previous one to calculate the difference (delta value). After absolute values of the delta were obtained, delta values smaller than the standard deviation were eliminated. The MAGE values were calculated by using the mean delta values that were greater than the standard deviation values" [103].

This is exactly the same procedure that the code above does, but is simply not the MAGE as defined by Service et al. in 1970. At this step I concluded that I reached the original source of the error and it was time to investigate its consequences.

5.9.1 The aftermath

After realising this error I looked for other studies that used gluvarpro 4.0 (and older versions) to calculate MAGE. I found that since its release it was used in at least three scientific papers (excluding the one made by my co-authors and I), all of them published in 2021 or 2022. In one of them it was only used to calculate other metrics of glycemic variability [104] but in the other two it was used to calculated MAGE thus their result were severely affected by this error.

The issue was raised to the author of to gluvarpro package, mentioning the other two studies that have possible been affected by the error. He confirmed the presence of the error and fixed it later on in the updated gluvarpro 6.0 and 7.0 versions. However, before this version was released, I already switched to use the "iglu" R package [102] to calculate MAGE values and the presented values of both MAGE and CONGA in this thesis were calculated with this package. For my previous paper, where gluvarpro was used, a corrigendum has been submitted. This did not affect the conclusions of the paper. The authors of the two paper of the trials have also been contacted.

5.10 Effects of artificially lowered detection limits

The hypothesis that traditional GV metrics are inherently limited as hyperglycaemia risk metrics because they are insensitive to high values, especially if there are only a few of them, have been indirectly strengthened as the vast majority of patients with the highest estimated EVS risk metrics had only moderate levels of the other metrics and all of them were heavily affected by the upper detection limit of the CGM sensor.

In order to investigate this behaviour further, an additional analysis was performed. Both the traditional variability metrics and the new EVS metrics were calculated after applying artificially lowered saturation points (or detection limits) by replacing all values above it with the value of the saturation point itself. Then the results of each metrics were compared to the value calculated in the main analysis, using the original dataset (with the 400 mg/dL upper detection limit of the CGM sensor). This means that the effect of saturation could be simulated through testing

on these synthetic datasets. The new values were plotted both using their actual value (absolute change) and their ratio to the value with the original, 400 mg/dL limit (relative change) on Figure 18.

As it was expected, the EVS metrics were in general more sensitive to this interference. In the plot of the relative changes it can be seen that the 1 year return level was roughly as sensitive as the traditional variability metrics, but the time spent above both 400 mg/dl and 600 mg/dl were much more sensitive to the lowered saturation levels. It can be seen that all patients' EVS result were affected at some point, while the results of using the traditional variability metrics remain the same regardless of lowering the saturation point even to 300 mg/dl for some patients. By its definition the TAR >250 did not change at all and it seems the IQR changed in only one case meaning the upper quartile of the CGM results of this patient was higher than the used saturation point. CONGA, CV and MAGE steadily decreased for a proportion of patients, while remain the same for the rest. A very interesting finding is that the MAGE increased in two cases when a lower saturation point was used. This is possible as it seems in these cases – in terms of MAGE at least – the lower saturation point had larger effect on the mean and SD than on the magnitude of peaks outside this range, thus resulting in a higher MAGE. This also shows a severe limitation of this metric used for risk assessment of hyperglycaemia as it results in a higher value in these cases, despite the patients actually having lower BG values and also lower variability.

Using regular statistical terminology this issue can be viewed as a missing data problem, as blood glucose levels beyond the upper detection limit cannot be measured but the CGM replaces these unknown measurements with a fixed value of 400 mg/dl. In a way it is similar to the "last observation carried forward" (or LOCF) method where the missing values of longitudinal or time-series data replaced by last observed (measured) value. It's a bit different because there is not necessarily an actual, exact measurement of 400 mg/dl at the 5 minutes interval when the CGM takes the measurement. It is also a "Missing Not at Random" (MNAR) problem which means the reason for being missing is related to the data itself (in this case too high to detect) as we know almost certainly that these values were above the upper detection limit. (Almost, because as it was shown in Chapter 5.5, there are cases when the confirmatory blood glucose measurement was below 400 mg/dl when the CGM recorded 400 mg/dl.)

It is also worth noting that being overly sensitive in this test could also mean that the metric is not ideal, depending on the purpose. It can be seen that the time above 600 mg/dl drops to zero relatively quickly if the saturation point falls to or

below 350 mg/dl. This could be interpreted as if there are no measurements above that level, it is very unlikely that any these patients would reach 600 mg/dl thus these patients would not be considered to be at high risk at all.

In summary, the results shown in Figure 18 confirmed that traditional metrics are very insensitive to saturation (i.e., extreme values) making them less likely to be an ideal metric for the risk assessment of hyperglycaemia, in contrast to the presented EVS metrics, especially the time spent above 400 and 600 mg/dl.

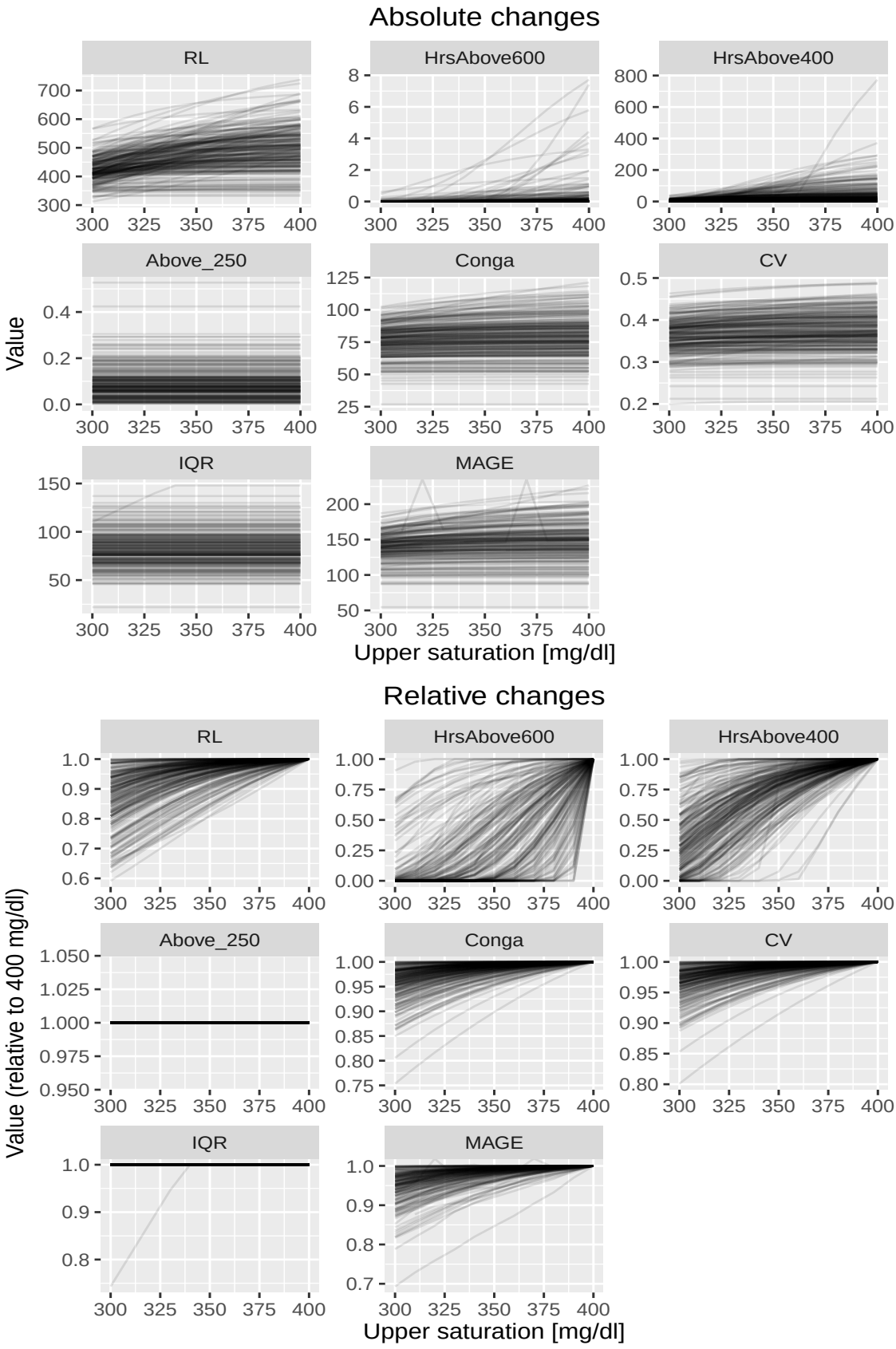

Figure 18: Results of the artificially lowered saturation points

5.11 Handling the dependence of the observations

An important attribute of the CGM data is that these observations are temporally dependent. By its clinical nature, blood glucose have limited ability to change in a short timeframe so the observations close in time are likely to be correlated. Using statistical terms, these observations are positively autocorrelated which means that there is a correlation between an observation of a time series and the values close to the given observation. This usually means a problem as it violates the assumption of some statistical methods but it does not cause much trouble with the BM approach (in contrast to the POT approach) [76, 105]. With the BM approach, this local dependence between the observations is less of a concern as in the secondary sample of the hourly block maximas we keep only 1 of every 12 observations and these are less correlated then the neighbouring observations which were normally taken at every 5 minutes. For the long-term dependence, a quite weak condition needs to be met: BM of the dependent data needs to have the same distribution as independent data would have. This is the $D(u_n)$ condition of Leadbetter [106]. It is true that the parameters will be different if the data are dependent, but as the parameters are estimated from the sample anyway, it does not cause any problem [76]. The drawback of the BM approach is the less efficient use of sample size as we "sacrificed" a large proportion of neighbouring, thus the most correlated observations. As in this case we had a very large sample of more than 14.8 million observations in total, this was less of a concern.

5.12 Non-stationary models

As it was briefly mentioned before, using extreme value analysis, it is possible to consider any or all of the three parameters describing the distribution of the extremes to be dependent on external variables of the subjects. This is similar to the backbone of regression modeling and is usually called non-stationary modelling within the framework of EVS. This of course means that the value of the estimated metrics – return levels probabilities, or time spent above certain levels – will be a function of the chosen variables thus the effect of these variables (such as time of the day or a clinical characteristic of the patient) on the outcomes can be investigated. Unfortunately as the available REPLACE-BG dataset went through a profound anonimisation process these was a very limited data available of the patients other than the CGM records. As their weight and height was available, Body Mass Index (BMI) was chosen to demonstrate the non-stationary modeling feature of EVS. This is a simple metric (weight in kilograms divided by the square of the height in meters)

widely used to assess obesity so it was an ideal choice to investigate its relationship with blood glucose extremes.

In the non-stationary modell outlined above, the effect of BMI on all parameters was statistically significant. Each unit of increase in BMI was associated with a change in the shape parameter of $\hat{\beta}_{\xi 1} = -0.0025(95\%CI : -0.0028, -0.0022)$, in the location parameter $\hat{\beta}_{\mu 1} = 0.76(95\%CI : 0.73, 0.79)$, while in the scale parameter $\hat{\beta}_{\sigma 1} = 0.400(95\%CI : 0.381, 0.418)$. In contrast to regression models, these coefficients cannot be interpreted directly. However, it is possible to plot the distribution of the hourly maxima for different BMI values, giving an interpretable illustration of the results; this is presented in Figure 19. In this graph the reference level of 20 kg/m^2, which is in the normal weight category according to the World Health Organization's categorisation, was compared with 40 kg/m^2 which is the cut-off point between grade 2 (obese) and grade 3 (or morbid) obesity [107, 108].

It can be seen in Figure 19 that the level of 40 kg/m^2 BMI (red line) was associated with a flatter distribution compared to the chosen reference level of 20 kg/m^2 (black line). This result is in contrast to relatively large previous studies which reported the opposite and found association between higher BMI and lower glycemic variability metrics [109, 110], however results of other studies with fewer patients [111, 112] contrasted them and were similar to our findings.

Additionally, we can see in Figure 19 a shift towards higher values and prolonged tail for the 40 kg/m^2 BMI level. It is important to emphasize that these are not the distribution of the actual (estimated) blood glucose values but the hourly maxima.

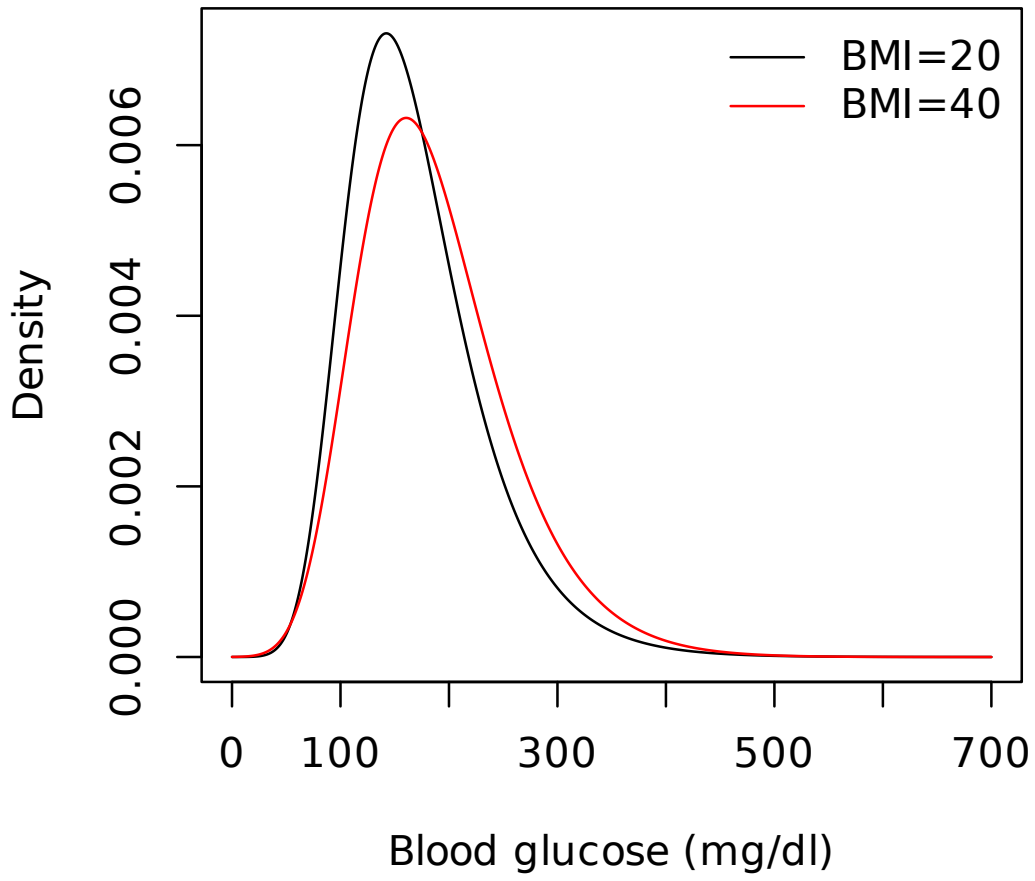

Figure 19: Distribution of the hourly maximum blood glucose for BMI: 20 vs 40 kg/m^2.

It is also important to note that despite BMI was added as a linear parameter to the model, its effect on the estimated values is non-linear because maxima depends non-linearly on the shape parameter.

5.13 Conclusion

The EVS methods were seldom used previously in the field of medicine or biology. Only a very few examples were found where EVS was used for such analysis; examples include the analysis of cholesterol levels [12] or the investigation of pneumonia and influenza deaths [13], both were published relatively recently. As

the previously presented early works on the applications of EVS for CGM analysis in Chapter 4 already indicated, the lack of sufficient data was a severe limitation for the application of EVS in the biomedical field and simulated data was not a proper solution for this matter. The utilization of the data of REPLACE-BG which contains more than 14.8 million real-life CGM measurements was a huge boost in the effort to develop and prove the applicability of the EVS methods for such analysis. The presented analysis and results in this chapter is the first where EVS was applied in diabetology, moreover on a particularly large dataset.

The validity of the CGM measurements were checked with a large number of available confirmatory blood glucose measurements with special attention to the upper detection limit which was proven to be an important limitation for the CGM sensor with measured values saturated at 400 mg/dl. The magnitude and distribution of measurement errors were assessed and confirmed that the findings of these investigations were based on valid measurements.

The results of the new EVS metrics were compared with some of the most important and widely used traditional metrics used to assess CGM measurements and to provide patient level information. These metrics rather characterise glycemic variability in some form and while glycaemic variability and risk of hyperglycaemia are likely correlated, the presented analyses also provided supporting evidence to the hypothesis that they are limited as hyperglycaemia risk metrics because they are quite insensitive to high values, especially if they are sparse. An analysis was conducted simulating the effect of lower detection limits – thus lower saturation points – for the CGM measurements and in general the new EVS metrics proven to be more sensitive to this effect compared to the traditional metrics.

Comparing the actual, patient level values, the results shown that in general there are not many patients who got high score according to the traditional metrics but low score with the EVS metrics, but patients with the highest values for EVS metrics (and thus the highest attributable risk) had only moderate values according to the regular metrics. This means that traditional metrics did not identified them as high-risk patients. The further examination of these cases shown that the CGM measurements of the vast majority of these patients often reached the upper detection limit of the CGM sensor. In these cases the recorded values had lower variability than the actual BG values would have due to this limitation.

Using the EVS models it is also possible to assess the impact of different clinical characteristics or treatments in a more precise and practical way using non-stationary models. Similarly to regression analysis, it's not just statistical significance but also the magnitude of this impact that can be estimated thus clinical

significance can be assessed as well. This makes this approach a suitable tool for the robust comparison of the importance of different patient characteristics and to compare the effectiveness of different type of treatments including not just drug therapies or lifestyle interventions but different algorithms of artificial pancreas systems for example. This was also presented through an example of the analysis of the effect of BMI on the hourly maxima of BG measurements. The results of this analysis shown that that higher BMI was associated with higher variability and higher BG levels. Similar results were published in some previous studies [111, 112] but these findings were in contrast with some others [109, 110]. The results presented here therefore also contribute to this debate.

It was successfully demonstrated that EVS enables the characterisation of CGM measurements focusing on the more relevant extremes in terms of hyperglycaemia risk; this was used to create relevant and clinically easily interpretable patient-level summary metrics. These more directly reflect the risk of hyperglycaemia than the traditional metrics which rather capture GV: they, for instance, give an estimate of spending a certain proportion of time above a BG level of interest within a time-frame, even if this threshold level is beyond the range of the observed data or way beyond the scope of the observation time.

An important limitation (and also a possible way of further development) of these results is that confirming whether the estimated patient level risks of hyperglycaemia were indeed better than those provided by the traditional metrics could only be validated with a longer follow up, where we could actually measure if the patient had hyperglycaemia phases and the duration of time spent above the EVS metric levels. This would, however, require CGM sensors that are capable to measure BG level in those relatively high ranges for a sufficient time and for a large number of patients. It would be even better if it was possible to confirm the predictive value of these metrics for the longer term clinical complications of type I diabetes (i.e., hard endpoints) but such analysis would require long follow up after a CGM measurement and a decent sample size.

5.14 Thesis group 2

Thesis 2.1

I applied the block maxima (BM) approach of EVS to a large sample of 226 patients from the REPLACE-BG clinical trial with CGM curves containing over 14.8 million observations. For the first time, EVS metrics were compared to widely used traditional metrics for patient level risk assessment of hyperglycaemia. In general, a relatively weak or moderate correlation was found between the EVS and the traditional metrics.

Related publications: [86, 87, 113].

Thesis 2.2

The patients with the highest risk according to the new EVS metrics had only moderate scores according to the traditional metrics. A further investigation of these measurements have shown that these were heavily affected by saturation caused by the detection limit of the CGM sensor. Subsequent analysis shown that EVS metrics were more sensitive to simulated decrease of these saturation levels.

Related publication: [113].

Thesis 2.3

Similarly to regression type analyses, coefficients can be added to EVS models as well to investigate their effect on the modelled outcome. I investigated the effect of body mass index (BMI) on blood glucose maxima. A statistically significant effect was found with higher BMI being associated with higher values of hourly maxima of blood glucose levels.

Related publications: [113, 105].

6 Regression analysis of the Recital trial

6.1 Introduction

In this chapter the statistical aspects of the analysis of the "Recital" clinical trial is presented focusing on the mixed effects regression models used for the analysis of its primary and some of its secondary outcomes. The background and the rationale behind the chosen methods, their advantage over other possible methods and the results are also presented. In order to understand these, a limited summary of the design of the trial and its clinical background is also provided as these play an important role in the statistical analysis and the choice of the methods used too. Such analysis is even more complex in a clinical trial setting, where the planned analyses and methods have to be pre-specified in detail with no or very with limited knowledge of the data at that time. This is needed to ensure the transparency and to prove that these choices where not driven by the results.

Personally I was responsible for the complete final analysis of this trial, development and writing up the statistical analysis plan (SAP) which can be found as a supplementary material of the paper of the final results and for regular reports of efficacy and safety of the trial to the Data monitoring and ethics committee (DMEC) and to the Trial steering committee (TSC) and contribution to the published paper.

6.2 Design and rationale of Recital

The Recital was a multi-centre, randomized, double blinded, controlled trial comparing rituximab (RTX) against intravenous cyclophosphamide (CYC) in Connective Tissue Disease (CTD) associated Interstitial Lung Disease (ILD). Altogether 101 patients were randomly allocated to receive one of the treatments and neither the participants nor the staff involved in the treatment of the patient knew which drug they received until the end of the trial (unless it was justified by an adverse event) [114]. Due to the unequal number and the different schedule of the admission of the infusions, it used a "double dummy" design meaning that both the treated (RTX) and control (CTD) arm patients received placebo as well in order to cover these differences [115]. Patients received these infusions at the first 7 visits in total until week 20 and there were two other follow-up visits at week 24 and 48.

ILD itself results in progressive breathlessness for the patients which often causes respiratory failure and death by inflammation and/or fibrosis causing thickening and distortion of the walls of the alveolars with consequent impairment of gas exchange. There are many possible causes of ILD but the most common with lung

involvement is the systemic autoimmune disease which is a dysregulation activation of immune system causing inflammation and tissue damages [116, 117].

CTD could be also a cause of death and disability and affects the working age population as well [118], however the developments in the treatments of CTD over the last years improved the prognosis of these patients dramatically [119]. Unfortunately, there has been very little improvement meanwhile in the therapy for CTD related ILD and this condition is still poorly understood and due to the lack of previous evidence, the choice of treatment was largely based on experts' opinion, meaning usually steroids and immunosuppressive drugs such as CYC (which was developed in the 1950s), but occasionally these are unable to control lung inflammation [120, 121]. RTX is a more sophisticated and more targeted alternative which causes mainly the depletion of a single type of white blood cells, called the B cells. Before the start of the Recital trial, RTX has been proven to effectively treat similar autoimmune diseases and in several cases other types of ILD as well [122, 123].

The primary outcome was the change in forced vital capacity (FVC) at 24 weeks which is an objective measurement of lung function being the amount of air that a person can exhale during a forced breath measured by a spirometer. According to the study's sample size calculation around 140ml (or 5%) change would be clinically meaningful. In order to detect such difference, with the assumptions made at the design phase (10% drop-out, 90% statistical power desired, 0.05 (two tailed) significance level) it would have needed 116 recruited participants which it failed to meet. The primary analysis was a modified intention to treat analysis meaning patients' data was analysed in respect to the groups as they were randomised into regardless of subsequent withdrawals or crossovers but only those were included in the analysis who received at least one dose of study drug.

The secondary outcomes included change in diffusing capacity for carbon monoxide (DLCO), oxygen saturation, 6 minute walk distance, different categorical changes in FVC, different health related and diseases specific quality of life scores, most of them investigating the difference between the value at 24 and 48 weeks to the baseline. The quality of life endpoints included the following patient reported outcomes: EQ-5D [124] (European Quality of Life Five-Dimension), KBILD [125] (King's Brief Interstitial Lung Disease), SGRQ [126] (St George's Respiratory Questionnaire). All of these measured on a 0–100 scale where higher values mean better quality of life for KBILD and EQ-5D while lower values mean better quality of life for SGRQ. A fourth quality of life assessment score was used as well which is not patient reported but reflect the physician's evaluation, called the physician's global disease assessment (GDA) [127] which is measured on a 10 cm visual analogue scale

and higher value represents worse status. Additionally safety, tolerability, mortality and healthcare utilisation and steroid usage were also assessed. Randomization was stratified by CTD diagnostic category meaning that within each of these categories 1:1 ratio between the two treatment arm will be maintained (in the background it means each of these categories have their "own" randomisation list) which is important because the underlying CTD diagnoses have different prognosis so if they are not balanced between the two arms that could lead to potential bias. More details on the study design can be found in the trial's protocol [128, 129].

6.3 Statistical considerations of Recital

In terms of data, the primary outcome and some of the secondary outcomes have a very similar structure: measured on a continuous scale and multiple times (between 3 and 9) and the difference between baseline and week 24 or 48 were the outcomes. This is a very important factor, as measurements (the observations) coming from the same person are not independent of each other. As the patients are in different condition their (mean) lung function and quality of life is different so that knowledge of the patient contains information about these measurements and each of these are likely correlated in time. This latter issue is called autocorrelation and it was briefly presented in chapter 5.11; but there, for the block maxima modelling, only a weak condition needed to be met. Independence of observations is a fundamental assumption of the basic regression models such as linear and logistic regression and for many other statistical methods as well [130]. Investigating the relationship between a baseline value and its subsequent change of a continuous variable is a common issue in both observational studies and in clinical trials and even with only two measurements the complexity of the available methods is much higher [131, 132]. In this case, there are more than two observations, and ideally it would be beneficial to utilise all the available information. Moreover, the visits should have occurred at the specified week post-randomisation ± 7 days and some of the available measurements are not equally spaced in time.

Apart from this, there was another layer of this issue because patients were recruited from multiple centres (hospitals) located in different regions of the United Kingdom and inherently, observations from the same centre are presumably not independent either. Because of this, outcome measurements of patients from the same centre could correlate caused by immeasurable, influential factors for example their income and lifestyle (affecting the quality of life outcomes even more) or the treatments and care they received before the trial and so on. Additionally, spirometers

can be calibrated differently causing a difference in FVC measurements between the sites due to measurement error assuming each hospital uses the same (set of) spirometers for at least a part of their measurements and the magnitude of measurement error, thus the variance due this could be different for each (type of) spirometer too. It is also important to note that while the number of measurements per patient supposed to be equal (with some missingness), the number of patients per each centre were expected to be very different. Considering the patient pools and the incidence of this disease it was expected that in worst case, some sites might recruit only one or two patients while the main site might be able to recruit the majority of the patients.

Additionally, some missingness was expected to occur in the measurements and it would be important to incorporate patients with incompete data too as the sample size was quite small and proportionally the loss of even a couple patients has larger effect compared to a larger trial, not to mention potential bias if the missignesses are non-random. Therefore, as a potential solution for this, multiple imputation was considered to be used depending on the amount of missing data and the pattern of missingness from very early on. This way the missing data is estimated using the available information and the estimation is repeated couple times (usually between 3-10) giving different possible values and creating different scenarios on which the analysis runs resulting different estimates each time, that can be combined in the end to obtain an average estimation. This way imputation also accounts from the variability arising from its uncertainty.

Overall, the data regarding the primary and most of the secondary outcomes had a similar structure: continuous, measured multiple times with some expected missingness and some not equally spaced in time. These repeated measurements are clustered at patient level with equal number of observations per patient but some missingness and the patients are also clustered (or nested) by the centres they were recruited from and these were expected to have different number of patients. Additionally it is known that the underlying CTD type is a strong predictor of the patients' prognosis and well-being thus directly or indirectly affects all of the outcomes of interest so the randomisation was stratified by this.

6.4 General guideline for statistical methods of clinical trials

In case of clinical trials, the hypotheses and methods are set in advance in the trial's protocol and its statistical analysis plan (SAP). This is important to ensure that the focus is on clinically relevant findings, and are not solely based on statistical

associations but also have some physiological explanation or evidence and is also important to avoid false positive findings, or biased conclusions due to "hunting" for the desired results. This also improves reproducibility, transparency, and validity as well [133]. The key difference compared to other fields is that this (ideally) happens before accessing the data which makes the choice of the right method exceptionally difficult as it is not certain that the collected data would meet the assumptions of the choice of the methods in question and often requires some speculation.

The protocol development is the first step of the clinical investigation as it is part of the research proposal sent to the competent national authorities and ethics committees to gain regulatory approval to run a clinical trial. Its mandatory contents and structure are set by the E6 section of The International Council for Harmonisation (ICH) guideline for Good Clinical Practice (GCP) [134]. This guideline is so fundamental for clinical trials that most of the national authorities requires all investigators and staff who are involved in the conduct, oversight or management of clinical trials (this includes the trial statisticians as well) to complete training in Good Clinical Practice (GCP) and refresh this training periodically (every 2 or 3 years) [135]. There are several other guidelines as well [136] but the most important that has to be mentioned is the SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials) which is the gold standard for clinical trials since its release [137].

For the same reasons mentioned above, the SAP should be written with no prior knowledge of the data but the timing is less restricted than that of the protocol's and is usually set by local Standard Operation Procedures. While the statistical principles for clinical trials are set by the E9 section of ICH GCP [134] there are no widely accepted guidelines for the contents of SAPs [138]. Of course, to some extent the methods can be changed during the trial but that has to be very well justified and documented so as not to violate transparency and trial integrity, and the changes have to be reviewed by the Data monitoring and ethics committee (DMEC) and Trial steering committee (TSC). So its advantageous if the methods set upfront are robust and flexible enough to be capable of dealing with some changes without the need of reconsidering the whole approach.

6.5 Choosing the appropriate method for the analyses

In the original version of protocol, in order to test the hypothesis if RTX was superior to CYC for these outcomes, analysis of covariance (ANCOVA) was planned to be used, including baseline FVC, the stratification factors for randomization

(CTD type) and treatment arm as covariates. ANCOVA would be capable to deal with repeated measurements but regards time as a factor so every subject must be observed at every fixed equally spaced time so it would not be ideal for the primary outcome analysis because of the ±7 days time window for the measurements and would be problematic neither to use for the secondary outcomes of the 48 week differences unless the 12 week measurements are discarded. But the main reason why it would be disadvantageous to use it is that it cannot deal with the further, centre level of clustering of the observations. Additionally, for this level it not too realistic to assume equal variance of the outcomes over all centre not just for the apparatus aided measurements (lung function tests) and neither for the quality of life outcomes.

6.5.1 Mixed effects regression models

In order to consider this third level, the originally proposed methods were changed to a three level mixed effects regression model. Unfortunately, the wording used in the literature in the various disciplines to describe repeated-measures, longitudinal or hierarchical data is inconsistent and varies by journals and topics, for example these models are also called hierarchical, random effects or multi-level regression models [139].

To make the matter worse for ANOVA-type analysis, an effect is called "fixed" if all possible levels of the factor required for inference are represented in the study, for example treatment arm or sex while an effect considered to be "random" if the levels of the factor represented in the study are just a "random" sample drawn from a larger set of levels such us hospitals in this case (even though they were not drawn entirely randomly). The same terminology have a different meaning in mixed effects regression modelling. All regression coefficients in the model are fixed and each can be also considered random depending on the level at which the variable is measured. The random part represents the variation of the regression coefficients between the subjects or groups considered at the higher levels of the hierarchy. To tackle these issues by clustering, it would be possible to include these factors as a covariates however these would increase the degrees of freedom by a lot and usually the differences between these levels are not in the focus of interest. To illustrate the concept of mixed effects models, let's take the simple linear regression equation first [140]:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

In case of a two level structure, instead of a new covariate, a second u_j residual (or error) term will be introduced:

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_j + e_{ij},$$

where $u_j \sim \mathcal{N}(0, \sigma_u^2)$ and $e_{ij} \sim \mathcal{N}(0, \sigma_e^2)$. We also assume that

$$\begin{aligned} \operatorname{cov}(u_j, e_{ij}) &= 0 \\ \operatorname{cov}(y_{i_1,j}, yi_{2,j} | x_{ij}) &= \sigma_u^2 \geq 0 \end{aligned}$$

Here, each i observation belong to group $1, \ldots, J$ and

- y_{ij} is the observed value of the dependent variable of observation i of group j
- x_{ij} is the value of the independent variable of observation i of group j
- β_0 is the fixed (mean) intercept
- β_1 is the change of the dependent variable for each unit increase of independent variable x
- u_j is the difference of group j from the overall mean intercept (called level 2 residual)
- σ_u^2 is the level 2 (between group) variance
- e_{ij} is the difference between the observed value of i and its predicted value (called level 1 residual)
- σ_e^2 is the level 1 (residual) variance
- $cov(u_j, e_{ij}) = 0$ means Level 1 and level 2 residuals are independent
- $cov(y_{i_1,j}, y_{i_2,j} | x_{ij}) = \sigma_u^2 \geq 0$ means if i_1 and i_2 are in the same j unit there is a positive covariance between their responses

This type of model often called variance components model as the residual variance is split into components corresponding to each level [140]. This is usually written as follows:

$$y_{ij} = \beta_{0ij} x_{0ij} + \beta_1 x_{ij},$$

where $\beta_{0ij} = \beta_0 + u_{0j} + e_{0ij}$ with $u_{0j} \sim \mathcal{N}(0, \sigma_u^2)$ and $e_{0ij} \sim \mathcal{N}(0, \sigma_e^2)$.

This way, a random parameter (u_0) has been added to the intercept (β_0) only, but it is possible to add a random parameter (u_1) to the intercept (β_1) as well, if the relationship of covariate x might vary across the groups. In that case the modified equation is the following:

$$y_{ij} = \beta_{0ij}x_{0ij} + \beta_{1ij}x_{ij},$$

with

$$\begin{aligned}\beta_{0ij} &= \beta_0 + u_{0j} + e_{0ij}\\ \beta_{1ij} &= \beta_1 + u_{1j}\\ x_{0ij} &= 1\end{aligned}$$

Here $\text{var}(u_{0j}) = \sigma_{u0}^2$ and $\text{var}(u_{1j}) = \sigma_{u1}^2$, and $\text{cov}(u_{0j}, u_{1j}) = \sigma_{u01}$ and $\text{var}(e_{0ij}) = \sigma_{e0}^2$.

Random slope was not used for the modelling of the Recital's outcomes, but a third level of clustering for the centres was used. This additional random component (v_{0k}) can be simply added to the intercept which will be therefore modified as:

$$\beta_{0ijk} = \beta_0 + u_{0j} + v_{0k} + e_{0ijk}.$$

Using this β_{0ijk} and the same conditions, the other covariates can be added same way as they would for a linear regression. The model was adjusted to CTD diagnosis, time was added as a covariate as the week when the measurement was taken and an interaction term was added to express the difference between the treatment arms at each week.

Overall, the mixed effects regression model used for all continuous, repeatedly measured outcomes in Recital can be put as:

$$y_{ijk} = \beta_{0ijk} + \beta_1(baseline(y_i)) + \beta_2(CTD_i) + \beta_3(week_{ij}) + \beta_4(treatment_i \cdot week_{ij})$$

Here:

- y_{ijk} is the observed value of the dependent variable for patient i at week j from centre k

- β_{0ijk} is the intercept (β_0) and the three residual terms as presented above
- $baseline(y_i)$ is the outcome of patient i at baseline, thus β_1 will express the association between the outcome at follow up visits and its baseline value
- CTD_i is the CTD diagnosis of patient i added as categorical variable, thus β_2 will express the mean difference in the outcome between each CTD category and a (selected) reference category
- $week_{ij}$ is just the week as a covariate at jth week so $week_{ij} = j$

Finally, $treatment_i \cdot week_{ij}$ is the interaction term that will answer the research question. Here, $treatment_i$ is the treatment arm for patient i and can have two values: for CYC group $treatment_i = 0$ and then $\beta_4(treatment_i \cdot week_{ij}) = 0$ and β_3 will express the mean weakly change for the outcome in the CYC arm. For the RTX arm $treatment_i = 1$ and then β_4 will show the mean weekly difference between RTX and CYC arms and this is exactly what the research questions asks, at specified (24 or 48) weeks.

6.6 Advantages and limitations

The main advantage of the model above is that it captures the three level, clustered structure of the data. Additionally, it utilises all available data as it is possible to fit it to patients with missing measurements and it also results an estimate at all timepoints even if the given measurement was missing, including weeks 24 and 48. It also uses measurements after the timepoint of interest for the estimates, in this case week 48 measures were used for the estimation of week 24 difference and for each outcome, the same model answers both research question for the shorter (week 24) and longer (week 48) term response.

Furthermore it can be easily extended for a subgroup analysis, for example to investigate how did each CTD group react the treatment by adding this variable as a third covariate to the interaction term which would then show different slopes for the effect of RTX for each CTD subgroup thus estimations for the difference in treatment effect at each timepoint. Naturally, as to any regression models, additional covariates can be added if its effect on the outcome would be put into the focus of interest. This could be especially useful if it is suspected to be a predictor of the outcome and the quantification of the association is needed or a difference has been observed between the two randomised arms in such variable, for example age or sex.

It provides confidence intervals for the estimated effects and can be used to predict the outcomes for patients outside the trial too.

Additionally, it also provides and estimate for the change of the outcome by time for the control group as well. Usually, this is not something a clinical trial investigates but in this case the results were not negligible. Arguably the control group alone has somewhat similar evidential value as a prospective observation study and it later turned there was a higher improvement in FVC in the CYC group than expected (and used for the sample size calculation) and the results of the other outcomes shown improvement as well which provided useful and additional evidence for the treatment options of CTD-ILD.

However, the main limitation of the model presented above is that it assumes a linear relationship between time and the outcome. This would also mean the week 48 measurement might have and overwhelming effect on the week 24 estimate. There are many possible solutions for this issue such as giving the measurements equal weight by adding the order of visits instead of weeks to the model, adding a non-linear term or terms to the model or splines (which means using different functions for different periods of the observation). The ability of modeling non-linear relationships between the dependent and independent variables are one of the most important advantages of using mixed effects regression models over ANOVA type of analysis, especially when missingness is present, even when both would be possible as the data structure is limited to only two levels [141]. The process of choosing the proper function for the covariates when non-linearity is present, especially in a clinical trial setting when this process should be outlined upfront and with limited knowledge of the data, has many further questions that are not the scope of the present investigation, but underlines the importance of using regression modelling instead of the more traditional ANOVA-type approach.

6.7 Results

Results of the primary outcome and the continuous secondary outcomes are summarised in Table 3. The estimated results of the treatment effect in the RTX group compared to CYC is presented at each timepoint of interest (with their 95% CI), together with the observed difference (and its SD) in the CYC group compared to baseline at the same timepoints. For the modelled differences a positive value means RTX performed better for all endpoints except GDA and SGRQ scores, for which a positive difference would favour CYC.

The baseline FVC was 2.23 litre (SD: 0.85) in the CYC and 2.25 litre (0.77)

Table 3: Observed values (CYC) and modelled differences (RTX) of Recital's primary and selected secondary outcomes

Outcome	**Observed difference in CYC compared to baseline (SD)**	**Modelled difference in RTX vs CYC (95% CI)**
FVC (mL)		
24 weeks	99 (329)	−40 (−153 to 74)
48 weeks	138 (440)	−58 (−178 to 62)
DLCO (mL/min/kPa)		
24 weeks	0.058 (0.706)	0.186 (−0.054 to 0.425)
48 weeks	0.131 (1.080)	0.117 (−0.137 to 0.372)
6 min walk distance (m)		
24 weeks	10.4 (78.6)	−0.72 (−24.76 to 23.32)
48 weeks	15.1 (82.8)	−22.46 (−48.43 to 3.51)
EQ-5D score		
24 weeks	3.5 (20.5)	3.06 (−3.05 to 9.18)
48 weeks	−1.2 (23.5)	4.77 (−1.73 to 11.27)
GDA score		
24 weeks	−2.9 (2.1)	−0.14 (−0.85 to 0.57)
48 weeks	−2.9 (2.5)	0.90 (0.11 to 1.68)
KBILD score		
24 weeks	9.4 (20.8)	0.40 (−5.73 to 6.52)
48 weeks	5.6 (25.6)	1.15 (−5.34 to 7.64)
SGRQ score		
24 weeks	−4.8 (19.6)	0.63 (−5.64 to 6.91)
48 weeks	−6.4 (24.3)	2.82 (−3.69 to 9.34)

in the RTX group so the observed average increase of 99 ml (329) at week 24 and 138 ml (440) in the CYC group means 4.4% and 6.2% respectively which is a lot better than the expected 1% decrease assumed for the sample size calculation. The estimated mean increase of FVC was lower in the RTX group at both timepoints, -40 ml (95% CI: -153, 74) at week 24 and -58 ml (95%CI: -178, 62) but the difference was not statistically significant as the 95% CI of the difference includes zero. DLCO increased as well, on average by 0.058 ml/min/kPa (0.706) at week 24 and by 0.131 ml/min/kPa (1.108) in the CYC arm and the estimated mean increase was even higher in the RTX group with 0.186 ml/min/kPa (95% CI: -0.054, 0.425) at week 24 and 0.117 ml/min/kPa (95% CI: -0.137, 0.372) but these estimated differences were not statistically significant either. The observed 6 minute walk test results were better in the CYC group at week 24 with 10.4 m (78.6) and 15.1 m (82.8) at week 48 on average and the estimated mean result in the RTX group was lower by a small margin at week 24 with -0.72 m (95% CI: -24.76, 23.32) but with quite large – but still not statistically significant – mean difference at week 48 with -22.46 m (95% CI: 48.43, 3.51).

In general, in the quality of life outcomes improvement was observed with rather small estimated differences between the two groups except for the EQ-5D where the observed mean change from baseline was 3.5 (20.5) at week 24 and -1.2 (23.5) at week 48 in the CYC arm and the estimated mean difference in the RTX arm was relatively high with 3.06 (95% CI: -3.05, 9.18) at week 24 and 4.77 (95% CI: -1.73, 11.27) at week 48. The observed mean difference compared to baseline in GDA score (where lower score means better quality of life) in CYC arm was -2.9 (2.1) at week 24 and -2.9 (2.5) at week 48 and the estimated mean difference compared to these was -0.14 (95% CI: -0.85, 0.57) at week 24 and RTX performed worse statistically significantly at week 48 with the change of 0.90 (95% CI: 0.11, 1.68). It worth to note that this was the only outcome when the SD of change was smaller than the mean change while in the other metrics SD was multiple of the mean change meaning the distribution of changes were rather uneven for both the physiological and quality of life metrics expect for the GDA. KBLID increased by 9.4 (20.8) at week 24 and 5.6 (25.6) in the CYC group and the mean estimated changes compared to these were 0.40 (95% CI: -5.73, 6.52) at week 24 and 1.15 (95% CI: -5.34, 7.64) at week 48 in the RTX arm. In the SGRQ score, where similarly to GDA, lower scores means better quality of life, the observed mean change in CYC group compared to baseline was -4.8 (19.6) at week 24 and -6.4 (24.3) at week 48 while the estimated difference compared to the CYC arm in the RTX arm was 0.63 (95% CI: -5.64, 6.91) at week 24 and 2.82 (95% CI: -3.69, 9,34) at week 48.

Note that not all comparisons of the study are listed here but only those continuous outcomes that were measured multiple times thus the above mentioned mixed effects regression models were suitable – and needed – to be used.

6.8 Further endpoints

Time to event type of endpoints were analysed using Cox regression [142, 143] (or proportional hazards regression). It is used to model the relationship between a set of independent variables and the time it takes for an event of interest to occur. The event of interest could be anything but Recital investigated mortality, progression-free survival where the event was a combination of death, transplant or decline in FVC greater than 10% from baseline and treatment failure where the event was defined as the need for transplant or rescue therapy after unblinding with either open-label CYC or RTX. The independent variables are assumed to affect the hazard rate, which is the conditional rate that an event will occur at a given time, given that it has not yet occurred. The hazard rate is modeled as a function of the independent variables. In this case the only independent variable was just the treatment, but similarly to other type of regressions, further covariates can be added. To measure and compare the relative risk of an event occurring at a given time, hazard ratio (HR) is used which is the ratio of the hazard rates for the two groups. A hazard ratio less than 1 indicates that the risk of the event occurring is lower in the group with the lower hazard rate, while a hazard ratio greater than 1 indicates that the risk of the event occurring is higher in the group with the higher hazard rate.

There was no statistically significant difference found between RTX group compared to CYC group as the estimated HR were 1.72 (95% CI: 0.31–9.56) for overall survival (2/48 (4%) died in CYC arm and 3/49 (6%) in RTX during the 48 weeks follow up), 1.11 (95% CI: 0.63–1.99) for progression-free survival and 1.25 (95% CI: 0.34–4.65) in time to treatment failure.

Furthermore steroid exposure was compared between the two arms as well, converted to hydrocortisone equivalents [144]. The average dose of each participant during the follow up was 13,291 mg (14,657) in the CYC group and 11,469 mg (10,041) in the RTX arm. This meant the daily dose per patient was 42.9 mg hydrocortisone per day in the CYC group and 37.6 mg hydrocortisone per day in the RTX group, which is a 12.3%, but statistically non-significant reduction (95% CI: –25.9 to 50.5). More adverse events were reported in the CYC group (646 events, 33 serious adverse events) than in the RTX group (445 events, 29 serious adverse

events).

6.9 Conclusion of Recital and mixed effects regression modelling

Through this real-life example of a multi-centered, randomised, controlled clinical trial, common challenges, problems, and possible approaches of statistical modelling of trials and the reasons why mixed effects regression modelling was a suitable and effective solution were presented in detail.

As in this trial there were continuous outcomes with multiple measurements taken for each patient during the follow up and these could not be treated as independent, basic regression models and statistical methods would not suitable for the analysis. A further twist, namely that these observations were also nested on a third level, by the patients' recruiting hospital (which might have even more important role as lung function outcomes were measured by spirometers), made the use of other statistical methods commonly used for repeated measurements unadvised for the analysis. Further advantages also utilised in this approach is that it enables the use of all available data, including data with missing observations, and the handling unequal group sizes which is inevitable in real-life clinical settings. It also allows adding further covariates to the model. This adds more flexibility and robustness to the analysis which is exceptionally useful in a clinical trial setting where ideally the details of the analysis are planned and pre-specified upfront, with limited or no knowledge of the actual data. Besides it's advantages, the mathematical background and difference of mixed effects models compared to linear regression was also presented through an illustration, which provided a detailed explanation on how it is able to model these relationships when the assumptions of the independence of observations are not met.

Through the use of these methods and the collected data it was possible to investigate if rituximab was superior to cyclophosphamide, the current first-line, standard treatment of interstitial lung disease associated with connective tissue disease. Recital was the first, large scale trial to investigate this research question while rituximab was often used off-label as rescue therapy with no evidence on the effectiveness of this therapy. Although the trial did not find rituximab superior over cyclophosphamide for neither the primary nor the secondary endpoint, improvements in lung function and disease related and general quality of life measures were established in both treatment groups and rituximab was associated with fewer adverse events and a reduction in corticosteroid exposure.

6.10 Thesis group 3

Thesis 3.1

I presented a mixed effects regression modelling strategy as a well-founded and more suitable solution than the possible alternatives to model a three-level, clustered, hierarchical data structure of a clinical trial. Through this, a reliable statistical analysis was conducted and published on a contemporary, multi-centered, randomised, controlled trial which was the first large scale study to assess the effectiveness of rituximab compared to cyclophosphamide for the treatment of interstitial lung disease associated with connective tissue disease.

Related publication: [129, 145].

7 Conclusion

This dissertation presented two applications of statistical modelling in biomedical setting.

The first was the development of a novel approach that focuses on the maximums of blood glucose measurements of patients with diabetes to characterize their risk using extreme value statistics instead of the classical metrics of diabetology. This branch of statistics had a very limited use in biomedical research before. Its theoretical background was presented, including the problem of statistical independence in this setting. Through some preliminary works that included the assessment of suitable data sources (at first using simulated data and then using CGM data of a single patient), the application of the two main EVS approaches – peak over threshold and block maxima – was assessed.

For the main work, block maxima approach was chosen as it is more suitable and easily interpretable tool for patient level risk assessment of hyperglycaemia, and it can directly estimate the time spent above certain, chosen, clinically important thresholds, even if these were never attained in the sample. Also, through the application of EVS, these were mathematically better founded and more sophisticated approach than the – mostly simple and ad hoc – classical metrics. The main analysis was conducted on a large dataset containing 14.8 million measurements of 226 patients of the REPLACE-BG trial. The results of the classical and the EVS metrics were compared and in general there were only a weak or moderate correlation between them, but more importantly, the patients identified as being of the highest risk were different. It was noticed that the detection limit of the sensors might have an important role as the measurements of the patients with the highest risk through the EVS metrics had been heavily affected by the upper detection limit of the sensor. The sensitivity of all metrics to this saturation due to the detection limit could be viewed as a non-random missing data problem. Its impact and relevance was assessed using simulated data with artificially lowered saturation points and it was found that EVS metrics are more sensitive to such saturation.

During this analysis, an important error was discovered in the widely used "gluvarpro" statistical package for CGM analysis which led to incorrect results in the calculation of one of the classical blood glucose variability metrics. This had a quite important impact as this was used in other studies to calculate this metric as one of their outcomes. The author of the package was contacted and the issue is now fixed (as of version 7.0 of gluvarpro).

Furthermore, the analysis was extended with the assessment of the relation-

ship between the body mass index and the hourly maximums of BG measurements with the use of non-stationary models. This is to some extant similar to regression analysis and provides a way to compare and estimate the magnitude of different patient characteristics or possible treatments. This analysis shown that higher BMI was associated with and higher BG maximums.

Naturally, this study has its limitations and opportunities for future development. Most importantly, in order to validate the results and draw conclusions on the performance of these metrics it would be important to see longer term CGM results that could prove the accuracy of EVS estimates by the actual time spent above the thresholds of which the estimates were given. This would require about one more year of CGM measurements, ideally without influential detection limits. Moreover, with decades long follow up and adequate sample size, it would be possible to see the predictive power of these metrics on the actual, long term complications of diabetes, but collection of such data is an exceedingly complicated undertaking.

The other biomedical application of statistical modelling presented in this document was the regression analysis of a clinical trial. Because of its repeated measures and nested setting through being multi-centered, it required the use of a three level, hierarchical, mixed effects regression model. Its background, the alternatives and the reasons behind this choice were presented. The key issue regarding this question was statistical independence which also appeared in the EVS analysis but in a very different context. It was the actual main analysis of a clinical trial, so amongst meeting the requirements of the related guidelines, most of the decisions regarding the analysis had to be set early on, with no or limited knowledge of the data, meaning excess difficulties compared to other applications of statistics. Thus the flexibility and robustness of reasonable regression models played an important role in this matter.

References

[1] Stephen J Roberts. "Extreme value statistics for novelty detection in biomedical data processing". In: *IEE Proceedings-Science, Measurement and Technology* 147.6 (2000), pp. 363–367.

[2] Michel K Ochi. "Principles of extreme value statistics and their application". In: *Paper of the Society of Naval Architects and Marine Engineers, SNAME, 1981* (1981).

[3] Erwan Le Roux, Guillaume Evin, Nicolas Eckert, Juliette Blanchet, and Samuel Morin. "Non-stationary extreme value analysis of ground snow loads in the French Alps: a comparison with building standards". In: *Natural Hazards and Earth System Sciences* 20.11 (2020), pp. 2961–2977.

[4] Jonathan Auerbach and Phyllis Wan. "Forecasting the urban skyline with extreme value theory". In: *International Journal of Forecasting* 36.3 (2020), pp. 814–828.

[5] Richard W Katz. "Statistics of extremes in climate change". In: *Climatic change* 100.1 (2010), pp. 71–76.

[6] Georgia Lazoglou and Christina Anagnostopoulou. "An overview of statistical methods for studying the extreme rainfalls in Mediterranean". In: *Multidisciplinary Digital Publishing Institute Proceedings* 1.5 (2017), p. 681.

[7] Erin Towler, Balaji Rajagopalan, Eric Gilleland, R Scott Summers, David Yates, and Richard W Katz. "Modeling hydrologic and water quality extremes in a changing climate: A statistical approach based on extreme value theory". In: *Water Resources Research* 46.11 (2010).

[8] Manfred Gilli et al. "An application of extreme value theory for measuring financial risk". In: *Computational Economics* 27.2 (2006), pp. 207–228.

[9] MB Adam and Jonathan Angus Tawn. "Modelling record times in sport with extreme value methods". In: *Malaysian Journal of Mathematical Sciences* 10.1 (2016), pp. 1–21.

[10] Serguei Y Novak. *Extreme value methods with applications to finance*. CRC Press, 2011.

[11] RR Kinnison. *Applied extreme-value statistics*. Tech. rep. Pacific Northwest Lab., Richland, WA (USA), 1983.

[12] P. Bermudez and Zilda Mendes. “Extreme Value Theory in Medical Sciences: Modeling Total High Cholesterol Levels”. In: *Journal of statistical theory and practice* 6 (Sept. 2012), pp. 468–491. DOI: 10.1080/15598608.2012.695673.

[13] Maud Thomas, Magali Lemaitre, Mark L Wilson, Cécile Viboud, Youri Yordanov, Hans Wackernagel, and Fabrice Carrat. “Applications of extreme value theory in public health”. In: *PloS one* 11.7 (2016), e0159312.

[14] SF Clarke and JR Foster. “A history of blood glucose meters and their role in self-monitoring of diabetes mellitus”. In: *British journal of biomedical science* 69.2 (2012), pp. 83–93.

[15] Bruce W Bode. “Clinical utility of the continuous glucose monitoring system”. In: *Diabetes Technology & Therapeutics* 2.1, Supplement 1 (2000), pp. 35–41.

[16] Jason K. Wang, David Ouyang, Jason Hom, Jeffrey Chi, and Jonathan H. Chen. “Characterizing electronic health record usage patterns of inpatient medicine residents using event log data”. In: *PLOS ONE* 14.2 (Feb. 2019), pp. 1–7. DOI: 10.1371/journal.pone.0205379. URL: https://doi.org/10.1371/journal.pone.0205379.

[17] Kurt George Matthew Mayer Alberti and Paul Z Zimmet. “Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation”. In: *Diabetic medicine* 15.7 (1998), pp. 539–553.

[18] Gerald M Reaven. “Role of insulin resistance in human disease (syndrome X): an expanded definition”. In: *Annual review of medicine* 44.1 (1993), pp. 121–131.

[19] Anjali D Deshpande, Marcie Harris-Hayes, and Mario Schootman. “Epidemiology of diabetes and diabetes-related complications”. In: *Physical therapy* 88.11 (2008), pp. 1254–1264.

[20] Gyula Soltesz, CC Patterson, G Dahlquist, and EURODIAB Study Group. “Worldwide childhood type 1 diabetes incidence–what can we learn from epidemiology?” In: *Pediatric diabetes* 8 (2007), pp. 6–14.

[21] Anastasia Katsarou, Soffia Gudbjörnsdottir, Araz Rawshani, Dana Dabelea, Ezio Bonifacio, Barbara J Anderson, Laura M Jacobsen, Desmond A Schatz, and Åke Lernmark. “Type 1 diabetes mellitus”. In: *Nature reviews Disease primers* 3.1 (2017), pp. 1–17.

[22] Nigel Unwin, Delice Gan, and David Whiting. “The IDF Diabetes Atlas: providing evidence, raising awareness and promoting action”. In: *Diabetes research and clinical practice* 87.1 (2010), pp. 2–3.

[23] Pouya Saeedi, Inga Petersohn, Paraskevi Salpea, Belma Malanda, Suvi Karuranga, Nigel Unwin, Stephen Colagiuri, Leonor Guariguata, Ayesha A Motala, Katherine Ogurtsova, et al. “Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas”. In: *Diabetes research and clinical practice* 157 (2019), p. 107843.

[24] Gregory A Roth, Degu Abate, Kalkidan Hassen Abate, Solomon M Abay, Cristiana Abbafati, Nooshin Abbasi, Hedayat Abbastabar, Foad Abd-Allah, Jemal Abdela, Ahmed Abdelalim, et al. “Global, regional, and national age-sex-specific mortality for 282 causes of death in 195 countries and territories, 1980–2017: a systematic analysis for the Global Burden of Disease Study 2017”. In: *The Lancet* 392.10159 (2018), pp. 1736–1788.

[25] Christian Bommer, Vera Sagalova, Esther Heesemann, Jennifer Manne-Goehler, Rifat Atun, Till Bärnighausen, Justine Davies, and Sebastian Vollmer. “Global economic burden of diabetes in adults: projections from 2015 to 2030”. In: *Diabetes care* 41.5 (2018), pp. 963–970.

[26] T Battelino, T Danne, RM Bergenstal, SA Amiel, R Beck, T Biester, E Bosi, BA Buckingham, WT Cefalu, KL Close, et al. “Doyle FJ 3rd, Garg S, Grunberger G, Heller S, Heinemann L, Hirsch IB, Hovorka R, Jia W, Kordonouri O, Kovatchev B, Kowalski A, Laffel L, Levine B, Mayorov A, Mathieu C, Murphy HR, Nimri R, Nørgaard K, Parkin CG, Renard E, Rodbard D, Saboo B, Schatz D, Stoner K, Urakami T, Weinzimer SA, Phillip M. Clinical Targets for Continuous Glucose Monitoring Data Interpretation: Recommendations From the International Consensus on Time in Range”. In: *Diabetes Care* 42.8 (2019), p. 1593.

[27] Rushad Patell, Denis Nigmatoulline, James Bena, Barbara Messinger-Rapport, MCecilia Lansang, et al. “Hyperglycemia and hypoglycemia in patients with diabetes in skilled nursing facilities”. In: *Endocrine Practice* 23.4 (2017), pp. 458–465.

[28] American Diabetes Association et al. “2. Classification and diagnosis of diabetes: standards of medical care in diabetes—2021”. In: *Diabetes care* 44.Supplement 1 (2021), S15–S33.

[29] Abbas E Kitabchi and Barry M Wall. “Diabetic ketoacidosis”. In: *Medical Clinics of North America* 79.1 (1995), pp. 9–37.

[30] Dyanne P Westerberg. “Diabetic ketoacidosis: evaluation and treatment”. In: *American family physician* 87.5 (2013), pp. 337–346.

[31] Abbas E Kitabchi, Guillermo E Umpierrez, John M Miles, and Joseph N Fisher. “Hyperglycemic crises in adult patients with diabetes”. In: *Diabetes care* 32.7 (2009), pp. 1335–1343.

[32] Stephen R Benoit, Yan Zhang, Linda S Geiss, Edward W Gregg, and Ann Albright. “Trends in diabetic ketoacidosis hospitalizations and in-hospital mortality—United States, 2000–2014”. In: *Morbidity and Mortality Weekly Report* 67.12 (2018), p. 362.

[33] Francisco J. Pasquel and Guillermo E. Umpierrez. “Hyperosmolar Hyperglycemic State: A Historic Review of the Clinical Presentation, Diagnosis, and Treatment”. In: *Diabetes Care* 37.11 (Nov. 2014), pp. 3124–3131. ISSN: 0149-5992. DOI: 10.2337/dc14-0984. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4207202/ (visited on 02/09/2020).

[34] Willa A Hsueh and Pamela W Anderson. “Hypertension, the endothelial cell, and the vascular complications of diabetes mellitus.” In: *Hypertension* 20.2 (1992), pp. 253–263.

[35] Hasan Temurtas, Nejat Yumusak, and Feyzullah Temurtas. “A comparative study on diabetes disease diagnosis using neural networks”. In: *Expert Systems with applications* 36.4 (2009), pp. 8610–8615.

[36] Julia Hippisley-Cox and Carol Coupland. “Diabetes treatments and risk of amputation, blindness, severe kidney failure, hyperglycaemia, and hypoglycaemia: open cohort study in primary care”. In: *bmj* 352 (2016).

[37] Jaimie D Steinmetz, Rupert RA Bourne, Paul Svitil Briant, Seth R Flaxman, Hugh RB Taylor, Jost B Jonas, Amir Aberhe Abdoli, Woldu Aberhe Abrha, Ahmed Abualhasan, Eman Girum Abu-Gharbieh, et al. “Causes of blindness and vision impairment in 2020 and trends over 30 years, and prevalence of avoidable blindness in relation to VISION 2020: the Right to Sight: an analysis for the Global Burden of Disease Study”. In: *The Lancet Global Health* 9.2 (2021), e144–e160.

[38] Kirsten L Johansen, Glenn M Chertow, David T Gilbertson, Charles A Herzog, Areef Ishani, Ajay K Israni, Elaine Ku, Shuling Li, Suying Li, Jiannong Liu, et al. "US Renal Data System 2021 Annual Data Report: Epidemiology of Kidney Disease in the United States". In: *American Journal of Kidney Diseases* 79.4 (2022), A8–A12.

[39] J Aaron Barnes, Mark A Eid, Mark A Creager, and Philip P Goodney. "Epidemiology and risk of amputation in patients with diabetes mellitus and peripheral artery disease". In: *Arteriosclerosis, thrombosis, and vascular biology* 40.8 (2020), pp. 1808–1817.

[40] Emerging Risk Factors Collaboration et al. "Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: a collaborative meta-analysis of 102 prospective studies". In: *The Lancet* 375.9733 (2010), pp. 2215–2222.

[41] S Ahtiluoto, T Polvikoski, M Peltonen, A Solomon, Jaakko Tuomilehto, B Winblad, R Sulkava, and M Kivipelto. "Diabetes, Alzheimer disease, and vascular dementia: a population-based neuropathologic study". In: *Neurology* 75.13 (2010), pp. 1195–1202.

[42] Celestino Sardu, Jessica Gambardella, Marco Bruno Morelli, Xujun Wang, Raffaele Marfella, and Gaetano Santulli. "Hypertension, thrombosis, kidney failure, and diabetes: is COVID-19 an endothelial disease? A comprehensive evaluation of clinical and basic evidence". In: *Journal of clinical medicine* 9.5 (2020), p. 1417.

[43] Viktor Dremin, Zbignevs Marcinkevics, Evgeny Zherebtsov, Alexey Popov, Andris Grabovskis, Hedviga Kronberga, Kristine Geldnere, Alexander Doronin, Igor Meglinski, and Alexander Bykov. "Skin complications of diabetes mellitus revealed by polarized hyperspectral imaging and machine learning". In: *IEEE Transactions on Medical Imaging* 40.4 (2021), pp. 1207–1216.

[44] E Andreoulakis, T Hyphantis, D Kandylis, and A Iacovides. "Depression in diabetes mellitus: a comprehensive review". In: *Hippokratia* 16.3 (2012), p. 205.

[45] Manoj Gopalakrishnan, Robin George Manappallil, Dipu Ramdas, and Jishnu Jayaraj. "The survival story of a diabetic ketoacidosis patient with blood sugar levels of 1985 mg/dL". In: *Asian Journal of Medical Sciences* 8.4 (2017), pp. 60–61.

[46] Joe F Jabre and Jeremy DP Bland. "Body mass index changes: an assessment of the effects of age and gender using the e-norms method". In: *BMC Medical Research Methodology* 21.1 (2021), pp. 1–7.

[47] *Highest blood sugar level Guinness World Records*. https://www.guinnessworldrecords.com/world-records/highest-blood-sugar-level. Accessed: 2022-08-10.

[48] Claudio Cobelli, Eric Renard, and Boris Kovatchev. "Artificial pancreas: past, present, future". In: *Diabetes* 60.11 (2011), pp. 2672–2682.

[49] Eray Kulcu, Janet A. Tamada, Gerard Reach, Russell O. Potts, and Matthew J. Lesho. "Physiological Differences Between Interstitial Glucose and Blood Glucose Measured in Human Subjects". In: *Diabetes Care* 26.8 (Aug. 2003), pp. 2405–2409. ISSN: 0149-5992. DOI: 10.2337/diacare.26.8.2405. eprint: https://diabetesjournals.org/care/article-pdf/26/8/2405/660964/dc0803002405.pdf. URL: https://doi.org/10.2337/diacare.26.8.2405.

[50] MJ Davies, Juan Jose Gagliardino, LJ Gray, K Khunti, V Mohan, and R Hughes. "Real-world factors affecting adherence to insulin therapy in patients with Type 1 or Type 2 diabetes mellitus: a systematic review". In: *Diabetic Medicine* 30.5 (2013), pp. 512–524.

[51] Jay S Skyler. "The economic burden of diabetes and the benefits of improved glycemic control: The potential role of a continuous glucose monitoring system". In: *Diabetes Technology & Therapeutics* 2.1, Supplement 1 (2000), pp. 7–12.

[52] Meryl Brod, Torsten Christensen, Trine L Thomsen, and Donald M Bushnell. "The impact of non-severe hypoglycemic events on work productivity and diabetes management". In: *Value in Health* 14.5 (2011), pp. 665–671.

[53] Yoshifumi Saisho, Chihiro Tanaka, Kumiko Tanaka, Rachel Roberts, Takayuki Abe, Masami Tanaka, Shu Meguro, Junichiro Irie, Toshihide Kawai, and Hiroshi Itoh. "Relationships among different glycemic variability indices obtained by continuous glucose monitoring". In: *primary care diabetes* 9.4 (2015), pp. 290–296.

[54] Thomas A Peyser, Andrew K Balo, Bruce A Buckingham, Irl B Hirsch, and Arturo Garcia. "Glycemic variability percentage: a novel method for assessing glycemic variability from continuous glucose monitor data". In: *Diabetes technology & therapeutics* 20.1 (2018), pp. 6–16.

[55] David Rodbard. “Hypo-and hyperglycemia in relation to the mean, standard deviation, coefficient of variation, and nature of the glucose distribution”. In: *Diabetes technology & therapeutics* 14.10 (2012), pp. 868–876.

[56] David Rodbard. “Quality of glycemic control: assessment using relationships between metrics for safety and efficacy”. In: *Diabetes technology & therapeutics* 23.10 (2021), pp. 692–704.

[57] Benjamin Jasha Van Enter and Elizabeth Von Hauff. “Challenges and perspectives in continuous glucose monitoring”. In: *Chemical Communications* 54.40 (2018), pp. 5032–5045.

[58] F John Service, George D Molnar, John W Rosevear, Eugene Ackerman, Lael C Gatewood, and William F Taylor. “Mean Amplitude of Glycemic Excursions, a Measure of Diabetic Instability”. In: *Diabetes* 19.9 (1970), pp. 644–655. ISSN: 0012-1797. DOI: 10.2337/diab.19.9.644. eprint: https://diabetes.diabetesjournals.org/content/19/9/644.full.pdf. URL: https://diabetes.diabetesjournals.org/content/19/9/644.

[59] C.M. McDonnell, S.M. Donath, S.I. Vidmar, G.A. Werther, and F.J. Cameron. “A Novel Approach to Continuous Glucose Analysis Utilizing Glycemic Variation”. In: *Diabetes Technology & Therapeutics* 7.2 (2005). PMID: 15857227, pp. 253–263. DOI: 10.1089/dia.2005.7.253. eprint: https://doi.org/10.1089/dia.2005.7.253. URL: https://doi.org/10.1089/dia.2005.7.253.

[60] Lalo Magni, Davide M Raimondo, Chiara Dalla Man, Marc Breton, Stephen Patek, Giuseppe De Nicolao, Claudio Cobelli, and Boris P Kovatchev. “Evaluating the efficacy of closed-loop glucose regulation via control-variability grid analysis”. In: *Journal of diabetes science and technology* 2.4 (2008), pp. 630–635.

[61] David Rodbard. “Evaluating quality of glycemic control: graphical displays of hypo-and hyperglycemia, time in target range, and mean glucose”. In: *Journal of diabetes science and technology* 9.1 (2014), pp. 56–62.

[62] Michelle Nguyen, Julia Han, Elias K Spanakis, Boris P Kovatchev, and David C Klonoff. “A review of continuous glucose monitoring-based composite metrics for glycemic control”. In: *Diabetes technology & therapeutics* 22.8 (2020), pp. 613–622.

[63] Daniel Cooley, Douglas Nychka, and Philippe Naveau. “Bayesian spatial modeling of extreme precipitation return levels”. In: *Journal of the American Statistical Association* 102.479 (2007), pp. 824–840.

[64] Lee Fawcett and David Walshaw. "Estimating return levels from serially dependent extremes". In: *Environmetrics* 23.3 (2012), pp. 272–283.

[65] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: https://www.R-project.org/.

[66] David Roodman. *EXTREME: Stata module to fit models used in univariate extreme value theory*. Statistical Software Components, Boston College Department of Economics. Jan. 2015. URL: https://ideas.repec.org/c/boc/bocode/s457953.html.

[67] Diethelm Wuertz, Tobias Setz, Yohan Chalabi, Maintainer Tobias Setz, and Suggests RUnit. "Package 'fExtremes'". In: (2009).

[68] Eric Gilleland and Richard W. Katz. "extRemes 2.0: An Extreme Value Analysis Package in R". In: *Journal of Statistical Software* 72.8 (2016), pp. 1–39. DOI: 10.18637/jss.v072.i08. URL: https://www.jstatsoft.org/index.php/jss/article/view/v072i08.

[69] R. A. Fisher and L. H. C. Tippett. "Limiting forms of the frequency distribution of the largest or smallest member of a sample". en. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2 (Apr. 1928), pp. 180–190. ISSN: 1469-8064, 0305-0041. DOI: 10.1017/S0305004100015681. (Visited on 02/09/2020).

[70] B. Gnedenko. "Sur La Distribution Limite Du Terme Maximum D'Une Serie Aleatoire". In: *Annals of Mathematics* 44.3 (1943), pp. 423–453. ISSN: 0003486X. URL: http://www.jstor.org/stable/1968974.

[71] Paul Embrechts, Claudia Klüppelberg, and Thomas Mikosch. *Modelling extremal events: for insurance and finance*. Vol. 33. Springer Science & Business Media, 2013.

[72] Laurens De Haan, Ana Ferreira, and Ana Ferreira. *Extreme value theory: an introduction*. Vol. 21. Springer, 2006.

[73] James Pickands III et al. "Statistical inference using extreme order statistics". In: *the Annals of Statistics* 3.1 (1975), pp. 119–131.

[74] Alexander McFarlane Mood. *Introduction to the Theory of Statistics*. McGraw-hill, 1950.

[75] M Ross Leadbetter. *Extremes and Local Dependence in Stationary Sequences.* Tech. rep. NORTH CAROLINA UNIV AT CHAPEL HILL DEPT OF STATISTICS, 1982.

[76] S. Coles, J. Bawa, Springer-Verlag (Berlin)., L. Trenner, and P. Dorazio. *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Springer, 2001. ISBN: 9781852334598.

[77] Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. "The UVA/PADOVA type 1 diabetes simulator: new features". In: *Journal of diabetes science and technology* 8.1 (2014), pp. 26–34.

[78] A. C. Davison and R. L. Smith. "Models for Exceedances Over High Thresholds". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 52.3 (1990), pp. 393–425. DOI: https://doi.org/10.1111/j.2517-6161.1990.tb01796.x.

[79] Max Rydman. *Application of the Peaks-Over-Threshold Method on Insurance Data*. 2018.

[80] Carl Scarrott and Anna MacDonald. "A review of extreme value threshold estimation and uncertainty quantification". In: *REVSTAT-Statistical journal* 10.1 (2012), pp. 33–60.

[81] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: https://www.R-project.org/.

[82] Eric Gilleland and Richard W. Katz. "extRemes 2.0: An Extreme Value Analysis Package in R". In: *Journal of Statistical Software* 72.8 (2016), pp. 1–39. DOI: 10.18637/jss.v072.i08.

[83] I. Pais, M. Hallschmid, K. Jauch-Chara, S. M. Schmid, K. M. Oltmanns, A. Peters, J. Born, and B. Schultes. "Mood and Cognitive Functions During Acute Euglycaemia and Mild Hyperglycaemia in Type 2 Diabetic Patients". en. In: *Experimental and Clinical Endocrinology & Diabetes* 115.01 (Jan. 2007), pp. 42–46. ISSN: 0947-7349, 1439-3646. DOI: 10.1055/s-2007-957348. URL: http://www.thieme-connect.de/DOI/DOI?10.1055/s-2007-957348 (visited on 02/09/2020).

[84] Sergio Contador. *gluvarpro: Glucose Variability Measures from Continuous Glucose Monitoring Data*. R package version 2.0. 2019. URL: https://CRAN.R-project.org/package=gluvarpro.

[85] Martin de Bock, Ethel Codner, Maria E. Craig, Tony Huynh, David M. Maahs, Farid H. Mahmud, Loredana Marcovecchio, and Linda A. DiMeglio. "ISPAD Clinical Practice Consensus Guidelines 2022: Glycemic targets and glucose monitoring for children, adolescents, and young people with diabetes". In: *Pediatric Diabetes* 23.8 (2022), pp. 1270–1276. DOI: https://doi.org/10.1111/pedi.13455. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/pedi.13455. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/pedi.13455.

[86] Mátyás Szigeti, Tamás Ferenci, and Levente Kovács. "The use of peak over threshold methods to characterise blood glucose curves". In: *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. 2020, pp. 000199–000204. DOI: 10.1109/SACI49304.2020.9118838.

[87] Mátyás Szigeti, Tamás Ferenci, and Levente Kovács. "The use of block maxima method of extreme value statistics to characterise blood glucose curves". In: *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*. 2020, pp. 433–438. DOI: 10.1109/SoSE50414.2020.9130427.

[88] T1D Exchange. *T1D Exchange*. 2021. URL: http://t1dexchange.org (visited on 07/10/2021).

[89] Grazia Aleppo, Katrina J Ruedy, Tonya D Riddlesworth, Davida F Kruger, Anne L Peters, Irl Hirsch, Richard M Bergenstal, Elena Toschi, Andrew J Ahmann, Viral N Shah, et al. "REPLACE-BG: a randomized trial comparing continuous glucose monitoring with and without routine blood glucose monitoring in adults with well-controlled type 1 diabetes". In: *Diabetes care* 40.4 (2017), pp. 538–545.

[90] Dexcom. *Dexcom G4 Platinum Continuous glucose monitoring system user's guide*. English. 2015.

[91] Sohail M Mulla, Ian A Scott, Cynthia A Jackevicius, John J You, and Gordon H Guyatt. "How to use a noninferiority trial: users' guides to the medical literature". In: *JAMA* 308.24 (2012), pp. 2605–2611.

[92] Jeremy Pettus, David A Price, and Steven V Edelman. "How patients with type 1 diabetes translate continuous glucose monitoring data into diabetes management decisions". In: *Endocrine Practice* 21.6 (2015), pp. 613–620.

[93] *Contour®next*. URL: https://diabetes.ascensia.com.au/products/contour-next-connected/#Sub-Menu-4.

[94] Thomas Danne, Revital Nimri, Tadej Battelino, Richard M Bergenstal, Kelly L Close, J Hans DeVries, Satish Garg, Lutz Heinemann, Irl Hirsch, Stephanie A Amiel, et al. "International consensus on use of continuous glucose monitoring". In: *Diabetes care* 40.12 (2017), pp. 1631–1640.

[95] Frederick Mosteller, John Wilder Tukey, et al. *Data analysis and regression: a second course in statistics*. Pearson, 1977.

[96] BS Everitt. "The cambridge dictionary of statistics cambridge university press". In: *Cambridge, UK Google Scholar* (1998).

[97] Xuefei Yu, Liangzhuo Lin, Jie Shen, Zhi Chen, Jun Jian, Bin Li, and Sherman Xuegang Xin. "Calculating the mean amplitude of glycemic excursions from continuous glucose data using an open-code programmable algorithm based on the integer nonlinear method". In: *Computational and mathematical methods in medicine* 2018 (2018).

[98] Nathaniel J Fernandes, Nhan Nguyen, Elizabeth Chun, Naresh M Punjabi, and Irina Gaynanova. "Open-Source Algorithm to Calculate Mean Amplitude of Glycemic Excursions Using Short and Long Moving Averages". In: *Journal of Diabetes Science and Technology* 16.2 (2022), pp. 576–577.

[99] Louis Monnier, Claude Colette, and David R Owens. "Glycemic variability: the third component of the dysglycemia in diabetes. Is it important? How to measure it?" In: *Journal of diabetes science and technology* 2.6 (2008), pp. 1094–1100.

[100] David Rodbard. "The challenges of measuring glycemic variability". In: *Journal of diabetes science and technology* 6.3 (2012), pp. 712–715.

[101] CM McDonnell, SM Donath, SI Vidmar, GA Werther, and FJ Cameron. "A novel approach to continuous glucose analysis utilizing glycemic variation". In: *Diabetes technology & therapeutics* 7.2 (2005), pp. 253–263.

[102] Steven Broll, Jacek Urbanek, David Buchanan, Elizabeth Chun, John Muschelli, Naresh Punjabi, and Irina Gaynanova. "Interpreting blood glucose data with R package iglu". In: *PloS One* 16.4 (2021), e0248560. DOI: 10.1371/journal.pone.0248560.

[103] Seniz Tuncan, Mehmet Uzunlulu, Ozge Caklili, HASAN MUTLU, and Aytekin Oğuz. "Evaluation of the Glycemic Fluctuation as Defined as the Mean Amplitude of Glycemic Excursion in Hospitalized Patients with Type 2 Diabetes". In: *CYPRUS JOURNAL OF MEDICAL SCIENCES* 1.3 (2016).

[104] Futoshi Ebara, Masayuki Domichi, Akiko Suganuma, and Naoki Sakane. "Comparison of Metformin and Alogliptin Fixed-Dose Tablets Once a Morning Versus Once an Evening Using Continuous Glucose Monitoring (AMPM Study): An Open-Label Randomized Cross-Over Trial". In: *Journal of Endocrinology and Metabolism* 11.1 (2021), pp. 8–13.

[105] Tamás Ferenci, Mátyás Szigeti, and Levente Kovács. "Using non-stationary extreme value analysis to characterize blood glucose curves". In: *2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE. 2022, pp. 000171–000176. DOI: 10.1109/SAMI54271.2022.9780743.

[106] Malcolm R Leadbetter, Georg Lindgren, and Holger Rootzén. *Extremes and related properties of random sequences and processes*. Springer Science & Business Media, 2012.

[107] Mabel Deurenberg-Yap and Paul Deurenberg. "Is a re-evaluation of WHO body mass index cut-off values needed? The case of Asians in Singapore". In: *Nutrition reviews* 61.suppl_5 (2003), S80–S87.

[108] World Health Organization et al. "Surveillance of chronic disease risk factors: country level data and comparable estimates". In: (2005).

[109] Francisco Gude, Pablo Díaz-Vidal, Cintia Rúa-Pérez, Manuela Alonso-Sampedro, Carmen Fernández-Merino, Jesús Rey-García, Carmen Cadarso-Suárez, Marcos Pazos-Couselo, José Manuel García-López, and Arturo Gonzalez-Quintela. "Glycemic variability and its association with demographics and lifestyles in a general adult population". In: *Journal of diabetes science and technology* 11.4 (2017), pp. 780–790.

[110] Jian Wang, Rengna Yan, Juan Wen, Xiaocen Kong, Huiqin Li, Peihua Zhou, Honghong Zhu, Xiaofei Su, and Jianhua Ma. "Association of lower body mass index with increased glycemic variability in patients with newly diagnosed type 2 diabetes: a cross-sectional study in China". In: *Oncotarget* 8.42 (2017), p. 73133.

[111] Sara J Salkind, Robert Huizenga, Stephanie J Fonda, M Susan Walker, and Robert A Vigersky. "Glycemic variability in nondiabetic morbidly obese persons: results of an observational study and review of the literature". In: *Journal of diabetes science and technology* 8.5 (2014), pp. 1042–1047.

[112] Silvio Buscemi, Loretta Cosentino, Giuseppe Rosafio, Manuela Morgana, Alessandro Mattina, Delia Sprini, Salvatore Verga, and Giovam Battista Rini. "Effects of hypocaloric diets with different glycemic indexes on endothelial function and glycemic variability in overweight and in obese adult patients at increased cardiovascular risk". In: *Clinical nutrition* 32.3 (2013), pp. 346–352.

[113] Mátyás Szigeti, Tamás Ferenci, and Levente Kovács. "The Use of Extreme Value Statistics to Characterize Blood Glucose Curves and Patient Level Risk Assessment of Patients With Type I Diabetes". In: *Journal of Diabetes Science and Technology* 17.2 (2023). PMID: 34814774, pp. 400–408. DOI: 10.1177/19322968211059547.

[114] Mette Thorlund Haahr and Asbjørn Hróbjartsson. "Who is blinded in randomized clinical trials? A study of 200 trials and a survey of authors". In: *Clinical Trials* 3.4 (2006), pp. 360–365.

[115] Ana Marušić and Stella Fatović Ferenčić. "Adoption of the double dummy trial design to reduce observer bias in testing treatments". In: *Journal of the Royal Society of Medicine* 106.5 (2013), pp. 196–198.

[116] Katerina M Antoniou, George A Margaritopoulos, Sara Tomassetti, Francesco Bonella, Ulrich Costabel, and Venerino Poletti. "Interstitial lung disease". In: *European Respiratory Review* 23.131 (2014), pp. 40–54.

[117] Marie Wahren-Herlenius and Thomas Dörner. "Immunopathogenic mechanisms of systemic autoimmune disease". In: *The Lancet* 382.9894 (2013), pp. 819–831.

[118] M Gaubitz. "Epidemiology of connective tissue disorders". In: *Rheumatology* 45.suppl_3 (2006), pp. iii3–iii4.

[119] Stephen C Mathai and Sonye K Danoff. "Management of interstitial lung disease associated with connective tissue disease". In: *Bmj* 352 (2016).

[120] Caterina Vacchi, Marco Sebastiani, Giulia Cassone, Stefania Cerri, Giovanni Della Casa, Carlo Salvarani, and Andreina Manfredi. "Therapeutic options for the treatment of interstitial lung disease related to connective tissue diseases. A narrative review". In: *Journal of Clinical Medicine* 9.2 (2020), p. 407.

[121] Adelle S Jee and Tamera J Corte. "Current and emerging drug therapies for connective tissue disease-interstitial lung disease (CTD-ILD)". In: *Drugs* 79.14 (2019), pp. 1511–1528.

[122] Katrina L Randall. "Rituximab in autoimmune diseases". In: *Australian prescriber* 39.4 (2016), p. 131.

[123] Yilin Wang and Liren Li. "Rituximab for connective tissue disease-associated interstitial lung disease: A systematic review and meta-analysis". In: *International Journal of Rheumatic Diseases* (2022).

[124] The EuroQol Group. "EuroQol-a new facility for the measurement of health-related quality of life". In: *Health policy* 16.3 (1990), pp. 199–208.

[125] Amit S Patel, Richard J Siegert, Katherine Brignall, Patrick Gordon, Sophia Steer, Sujal R Desai, Toby M Maher, Elisabetta A Renzoni, Athol U Wells, Irene J Higginson, et al. "The development and validation of the King's Brief Interstitial Lung Disease (K-BILD) health status questionnaire". In: *Thorax* 67.9 (2012), pp. 804–810.

[126] PW Jones, FH Quirk, and CM Baveystock. "The St George's Respiratory Questionnaire." In: *Respiratory medicine* 85 (1991), pp. 25–31.

[127] JS Smolen, FC Breedveld, MH Schiff, JR Kalden, P Emery, G Eberl, PL Van Riel, and P Tugwell. "A simplified disease activity index for rheumatoid arthritis for use in clinical practice". In: *Rheumatology* 42.2 (2003), pp. 244–257.

[128] Peter Saunders, Vicky Tsipouri, Gregory J Keir, Deborah Ashby, Marcus D Flather, Helen Parfrey, Daphne Babalis, Elisabetta A Renzoni, Christopher P Denton, Athol U Wells, et al. "Rituximab versus cyclophosphamide for the treatment of connective tissue disease-associated interstitial lung disease (RECITAL): study protocol for a randomised controlled trial". In: *Trials* 18.1 (2017), pp. 1–11.

[129] Toby M Maher, Veronica A Tudor, Peter Saunders, Michael A Gibbons, Sophie V Fletcher, Christopher P Denton, Rachel K Hoyles, Helen Parfrey, Elisabetta A Renzoni, Maria Kokosi, Matyas Szigeti, et al. "Rituximab versus intravenous cyclophosphamide in patients with connective tissue disease-associated interstitial lung disease in the UK (RECITAL): a double-blind, double-dummy, randomised, controlled, phase 2b trial". In: *The Lancet Respiratory Medicine* 11.1 (2023), pp. 45–54. ISSN: 2213-2600. DOI: `https://doi.org/10.1016/S2213-2600(22)00359-9`.

[130] Bart Meuleman, Geert Loosveldt, and Viktor Emonds. "Regression analysis: Assumptions and diagnostics". In: *The SAGE handbook of regression analysis and causal inference* (2015), pp. 83–110.

[131] Andrew J Vickers and Douglas G Altman. "Analysing controlled trials with baseline and follow up measurements". In: *Bmj* 323.7321 (2001), pp. 1123–1124.

[132] Arnaud Chiolero, Gilles Paradis, Benjamin Rich, and James A Hanley. "Assessing the relationship between the baseline value of a continuous variable and subsequent change over time". In: *Frontiers in public health* 1 (2013), p. 29.

[133] Ian Yuan, Alexis A Topjian, Charles D Kurth, Matthew P Kirschen, Christopher G Ward, Bingqing Zhang, and Janell L Mensinger. "Guide to the statistical analysis plan". In: *Pediatric Anesthesia* 29.3 (2019), pp. 237–242.

[134] ICH Harmonised Guideline. "Integrated addendum to ICH E6 (R1): guideline for good clinical practice E6 (R2)". In: *Current Step* 2 (2015), pp. 1–60.

[135] Thomas P Shanley, Nancy A Calvin-Naylor, Ruthvick Divecha, Michelle M Wartak, Karen Blackwell, Jonathan M Davis, Edward F Ellerbeck, Karl Kieburtz, Margaret J Koziel, Katherine Luzuriaga, et al. "Enhancing Clinical Research Professionals' Training and Qualifications (ECRPTQ): recommendations for Good Clinical Practice (GCP) training for investigators and study coordinators". In: *Journal of Clinical and Translational Science* 1.1 (2017), pp. 8–15.

[136] Jennifer M Tetzlaff, An-Wen Chan, Jessica Kitchen, Margaret Sampson, Andrea C Tricco, and David Moher. "Guidelines for randomized clinical trial protocol content: a systematic review". In: *Systematic reviews* 1.1 (2012), pp. 1–11.

[137] An-Wen Chan, Jennifer M Tetzlaff, Douglas G Altman, Andreas Laupacis, Peter C Gøtzsche, Karmela Krleža-Jerić, Asbjørn Hróbjartsson, Howard Mann, Kay Dickersin, Jesse A Berlin, et al. "SPIRIT 2013 statement: defining standard protocol items for clinical trials". In: *Annals of internal medicine* 158.3 (2013), pp. 200–207.

[138] Carrol Gamble, Ashma Krishan, Deborah Stocken, Steff Lewis, Edmund Juszczak, Caroline Doré, Paula R Williamson, Douglas G Altman, Alan Montgomery, Pilar Lim, et al. "Guidelines for the content of statistical analysis plans in clinical trials". In: *Jama* 318.23 (2017), pp. 2337–2343.

[139] R Diez. "A glossary for multilevel analysis". In: *Journal of epidemiology and community health* 56.8 (2002), p. 588.

[140] Jos WR Twisk. *Applied multilevel analysis: a practical guide for medical researchers*. Cambridge university press, 2006.

[141] Charlene Krueger and Lili Tian. "A comparison of the general linear mixed model and repeated measures ANOVA using a dataset with multiple missing data points". In: *Biological research for nursing* 6.2 (2004), pp. 151–157.

[142] David R Cox. "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[143] Dhananjay Kumar and Bengt Klefsjö. "Proportional hazards model: a review". In: *Reliability Engineering & System Safety* 44.2 (1994), pp. 177–188.

[144] F Buttgereit, JAP Da Silva, M Boers, GR Burmester, M Cutolo, J Jacobs, J Kirwan, L Köhler, PLCM van Riel, T Vischer, et al. "Standardised nomenclature for glucocorticoid dosages and glucocorticoid treatment regimens: current questions and tentative answers in rheumatology". In: *Annals of the rheumatic diseases* 61.8 (2002), pp. 718–722.

[145] Vicky Tsipouri, Peter Saunders, Greg J. Keir, Deborah Ashby, Sophie V. Fletcher, Michael Gibbons, Matyas Szigeti, Helen Parfrey, Elizabeth A. Renzoni, and Chris P. Denton. "Poster: Rituximab versus cyclophosphamide for the treatment of connective tissue disease associated interstitial lung disease (RECITAL): a randomised controlled trial". In: *4th International Clinical Trials Methodology Conference (ICTMC) and the 38th Annual Meeting of the Society for Clinical Trials*. Trials, May 2017, p. 130. DOI: 10.1186/s13063-017-1902-y.

Own publications related to the theses

[86] Mátyás Szigeti, Tamás Ferenci, and Levente Kovács. "The use of peak over threshold methods to characterise blood glucose curves". In: *2020 IEEE 14th International Symposium on Applied Computational Intelligence and Informatics (SACI)*. 2020, pp. 000199–000204. DOI: 10.1109/SACI49304.2020.9118838.

[87] Mátyás Szigeti, Tamás Ferenci, and Levente Kovács. "The use of block maxima method of extreme value statistics to characterise blood glucose curves". In: *2020 IEEE 15th International Conference of System of Systems Engineering (SoSE)*. 2020, pp. 433–438. DOI: 10.1109/SoSE50414.2020.9130427.

[105] Tamás Ferenci, Mátyás Szigeti, and Levente Kovács. "Using non-stationary extreme value analysis to characterize blood glucose curves". In: *2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. IEEE. 2022, pp. 000171–000176. DOI: 10.1109/SAMI54271.2022.9780743.

[113] Mátyás Szigeti, Tamás Ferenci, and Levente Kovács. "The Use of Extreme Value Statistics to Characterize Blood Glucose Curves and Patient Level Risk Assessment of Patients With Type I Diabetes". In: *Journal of Diabetes Science and Technology* 17.2 (2023). PMID: 34814774, pp. 400–408. DOI: 10.1177/19322968211059547.

[129] Toby M Maher, Veronica A Tudor, Peter Saunders, Michael A Gibbons, Sophie V Fletcher, Christopher P Denton, Rachel K Hoyles, Helen Parfrey, Elisabetta A Renzoni, Maria Kokosi, Matyas Szigeti, et al. "Rituximab versus intravenous cyclophosphamide in patients with connective tissue disease-associated interstitial lung disease in the UK (RECITAL): a double-blind, double-dummy, randomised, controlled, phase 2b trial". In: *The Lancet Respiratory Medicine* 11.1 (2023), pp. 45–54. ISSN: 2213-2600. DOI: https://doi.org/10.1016/S2213-2600(22)00359-9.

[145] Vicky Tsipouri, Peter Saunders, Greg J. Keir, Deborah Ashby, Sophie V. Fletcher, Michael Gibbons, Matyas Szigeti, Helen Parfrey, Elizabeth A. Renzoni, and Chris P. Denton. "Poster: Rituximab versus cyclophosphamide for the treatment of connective tissue disease associated interstitial lung disease (RECITAL): a randomised controlled trial". In: *4th International Clinical Trials Methodology Conference (ICTMC) and the 38th Annual Meeting of the*

Society for Clinical Trials. Trials, May 2017, p. 130. DOI: 10.1186/s13063-017-1902-y.

Own publications not related to the theses

[146] Manjit S Gohel, Jocelyn Mora, Matyas Szigeti, David M Epstein, Francine Heatley, Andrew Bradbury, Richard Bulbulia, Nicky Cullum, Isaac Nyamekye, Keith R Poskitt, et al. "Long-term clinical and cost-effectiveness of early endovenous ablation in venous ulceration: a randomized clinical trial". In: *JAMA surgery* 155.12 (2020), pp. 1113–1121. DOI: 10.1001/jamasurg.2020.3845.

[147] Ping-Tee Tan, Suzie Cro, Eleanor Van Vogt, Matyas Szigeti, and Victoria R Cornelius. "A review of the use of controlled multiple imputation in randomised controlled trials with missing outcome data". In: *BMC medical research methodology* 21.1 (2021), pp. 1–17. DOI: 10.1186/s12874-021-01261-6.

[148] Neil R Poulter, Christos Savopoulos, Aisha Anjum, Martha Apostolopoulou, Neil Chapman, Mary Cross, Emanuela Falaschetti, Spiros Fotiadis, Rebecca M James, Ilias Kanellos, Matyas Szigeti, et al. "Randomized crossover trial of the impact of morning or evening dosing of antihypertensive agents on 24-hour ambulatory blood pressure: the HARMONY trial". In: *Hypertension* 72.4 (2018), pp. 870–873. DOI: 10.1161/HYPERTENSIONAHA.118.11101. URL: https://www.ahajournals.org/doi/abs/10.1161/HYPERTENSIONAHA.118.11101.

[149] Mátyás Szigeti, Levente Kovács, and Tamás Ferenci. "Stability of relative and absolute metrics: empirical evidence from pulmonology". In: *2019 IEEE 17TH World Symposium on Applied Machine Intelligence and Informatics (SAMI 2019)*. 2019, pp. 235–238. DOI: 10.1109/SAMI.2019.8782769.

[150] Lívia Priyanka Elek, Matyas Szigeti, Berta Erdélyi-Hamza, Mátyás Szigeti, Konstantinos N Fountoulakis, and Xénia Gonda. "What you see is what you get? Association of belief in conspiracy theories and mental health during COVID-19." In: *Neuropsychopharmacologia Hungarica* 24 (2022), pp. 42–55. ISSN: 1419-8711. URL: https://europepmc.org/article/med/35451591.

[151] Lívia Priyanka Elek, Mátyás Szigeti, Berta Erdélyi-Hamza, Nikolett Beáta Vadon, Konstantinos N. Fountoulakis, Daria Smirnova, and Xénia Gonda. "Association of lifestyle changes during the pandemic are associated depression and its distinct symptom clusters – consideration for prevention and intervention". In: *Neuroscience Applied* (2023). ISSN: 2772-4085. DOI: 10.

1016/j.nsa.2022.100818. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9789424/.

[152] Nikolett Beata Vadon, Livia Priyanka Elek, Matyas Szigeti, Berta Erdelyi-Hamza, Daria Smirnova, Konstantinos N Fountoulakis, and Xenia Gonda. "Association between Lifestyle-and Circadian Rhythm-Related Changes, and Different Depression Symptom Clusters during COVID-19." In: *Psychiatria Danubina* 34.Suppl 8 (Sept. 2022), pp. 81–89. ISSN: 0353-5053. URL: https://europepmc.org/article/med/36170708.

[153] Péter Torzsa, László Kalabay, Dalma Dorottya Csatlós, Csenge Hargittay, Bernadett Márkus, András Mohos, Mátyás Szigeti, Tamás Ferenci, Verschoor Marjolein, Rozsnyai Zsofia, Gussekloo Jacobijn, K. E. Poortvliet Rosalinde, and Streit Sven. "A nagyon idős és esendő állapotú betegek antihipertenzív kezelési gyakorlata az alapellátásban". In: *Lege Artis Medicinae* 30 (2020), pp. 111–121. ISSN: 0866-4811. DOI: 10.33616/lam.30.011. URL: http://real-j.mtak.hu/14007/7/LAM_2020_03.pdf#page=30.

[154] Péter Torzsa, László Kalabay, Dalma Dorottya Csatlós, Csenge Hargittay, Bernadett Márkus, András Mohos, Mátyás Szigeti, Tamás Ferenci, Marjolein Verschoor, Zsófia Rozsnyai, Gussekloo Jacobijn, Rosalinde KE Poortvliet, and Sven Streit. "Antihipertenzív kezelés: A családorvosok inkább kezelik az esendő pácienseket". In: *Medical Tribune* 18 (2020), pp. 39–41. ISSN: 1589-1283.

[155] Tímea Vissi, Regina Szabó, Blanka Bágyi, Adél Göntér, Fanni Akkir, Mátyás Szigeti, Gabriella Erzsébet Papp, Éva Feketéné Szabó, and Anna Kelemen. *Cerebrális parézissel élő gyermekek számára készült diagnózis specifikus életminőség felmérő kérdőív (CPQOL) magyar nyelven történő alkalmazása.* 2018.

[156] Anita Zadori, Zsuzsanna Kis, Tibor Toth, Matyas Szigeti, Andras Temesvari, Geza Fontos, Noémi Nyolczas, and Peter Andreka. "Long-Term Efficacy and Safety of Left Atrial Appendage Closure Procedures". In: *International Heart Journal* 64.2 (2023), pp. 188–195. DOI: 10.1536/ihj.22-639.

List of Figures

List of Tables