Handling missing data, selection bias, and measurement error in observational studies

Choi, J.

Chapter 3

Comparing methods for measurement error detection in serial 24-hour hormonal data

Published in J Biol Rhythms. 2019 Aug;34(4): 347-363

Evie van der Spoel*, Jungyeon Choi*, Ferdinand Roelfsema, Saskia le Cessie , Diana van Heemst, Olaf M. Dekkers

*Contributed equally to this work

Abstract

Measurement errors commonly occur in 24-hour hormonal data and may affect the outcomes of such studies. Measurement errors often appear as outliers in such datasets; however, no well-established method is yet available for their automatic detection.

In this study, we aimed to compare the performances of different methods for outlier detection in hormonal serial data. Hormones (glucose, insulin, thyroid stimulating hormone (TSH), cortisol, and growth hormone (GH)) were measured in blood sampled every 10 minutes for 24 hours in 38 participants of the Leiden Longevity Study. Four methods for detecting outliers were compared: i) eyeballing, ii) Tukey's fences, iii) Stepwise approach, and iv) the Expectation-Maximization (EM) algorithm. Eyeballing detects outliers based on experts' knowledge, and Stepwise approach incorporates physiological knowledge with a statistical algorithm. Tukey's fences and the EM algorithm are data-driven methods, using interquartile range and a mathematical algorithm to identify underlying distribution, respectively. The performance of the methods was evaluated based on the number of outliers detected and the change in statistical outcomes after removing detected outliers. Eyeballing resulted in the lowest number of outliers detected (1.0% of all data points), followed by Tukey's fences (2.3%), Stepwise approach (2.7%), and the EM algorithm (11.0%). In all methods, the mean hormone levels did not materially change after removing outliers. However, their minima were affected by outlier removal. Although removing outliers affected the correlation between glucose and insulin on the individual level, when averaged over all participants, none of the four methods influenced the correlation.

Based on our results, the EM algorithm is not recommended given the high number of outliers detected, even where data points are physiologically plausible. Since Tukey's fences is not suitable for all types of data, and eyeballing is time-consuming, we recommend Stepwise approach for outlier detection which combines physiological knowledge and an automated process.

1. Introduction

Many physiological parameters such as hormones or metabolites exhibit rhythmicity. These rhythms are regulated by different systems. The most prominent rhythm is the circadian rhythm, which is induced by the biological clock located in the suprachiasmatic nucleus of the brain. The biological clock does not only synchronize molecular clocks in peripheral cells, but it also orchestrates many physiological functions, including blood pressure, core body temperature, and hormone secretion. An example of a hormone that exhibits strong circadian rhythmicity is cortisol. The sleep-wake cycle is another form of rhythm, and although similar to the circadian rhythm, it has other effects on hormone secretion than the biological clock. The secretion of growth hormone, for example, is more strongly influenced by sleep than by clock time. External cues, including food intake and physical activity, also can influence hormone secretion, such as the secretion of insulin (Oike *et al*., 2014).

Hormones and metabolites are measured for different purposes; e.g., in clinical settings to make a diagnosis or to evaluate the effect of treatment and in research settings to investigate how these parameters change upon interventions or differ between groups. Different cues can elicit changes in hormone secretion, amongst which circadian time, nutrient availability and food intake, physical activity, and sleep. Circulating concentrations of many hormones change over time, because these hormones are secreted in a pulsatile fashion and have a relatively short half-life (Spiga *et al*., 2015). Therefore, to obtain reliable hormonal time series data, hormones need to be measured in blood that is sampled frequently. For some hormones, such as insulin, the preferred sampling frequency is 2 minutes because of its short half-life (Porksen *et al*., 1997). Other hormones, including thyroid stimulating hormone (TSH), can be measured every 20 minutes to obtain reliable profiles (Odell et al., 1967; Grossmann et al., 1997). To take into account practical possibilities, half-lives, costs, and ethics, most studies investigating hormone secretion are performed with a sampling frequency of every 10 minutes during 24 hours, as reviewed by Veldhuis *et al.* and Roelfsema *et al.* (Veldhuis *et al*., 2016; Roelfsema *et al*., 2017).

When measuring hormones frequently over time, measurement errors are likely to occur. Measurement errors can be caused by pre-analytical experimental variation of various sources, including sample dilution (possibly because of flushing the intravenous line with heparinized saline), or the presence of a blood clot in the sample. Measurement error can influence the outcomes of studies with serial hormonal data. Therefore, it is important to identify measurement errors. Measurement errors are likely to be outliers (Grubbs, 1969), which deviate largely from the overall trend of the data. The challenge is that there is no clear-cut distinction between measurement errors and true biological variation. The starting point to detect measurement errors, however, is by identifying outliers.

No well-established method is yet available to automatically detect measurement errors. Therefore, we aimed to compare four methods to detect outliers likely due to measurement errors in 24-hour hormonal data: eyeballing (relying on experts' opinions), Tukey's fences (identifying outliers based on inter-quartile ranges), Stepwise approach (identifying outliers based on standard deviations), and the Expectation Maximization (EM) algorithm (using a mathematical algorithm based on disentangling the two different distributions of outliers and non-outliers). Furthermore, we studied the influence of removing the detected outliers on the assessment of statistical features of 24-hour hormonal data such as mean, minimum, maximum, and cross-correlation.

For this study, we used data on the pituitary hormones growth hormone (GH), adrenocorticotropic hormone (ACTH) and TSH, the adrenal hormone cortisol, as well as data on the metabolic signals insulin, and glucose, which were all measured during 24 hours every 10 minutes in serum from 38 participants of the Switchbox Leiden Study (Jansen *et al.*, 2015).

2. Methods

2.1. Data collection

Study population

The Leiden Longevity Study comprises 421 families with at least two long-lived Caucasian siblings fulfilling the age criteria (men ≥89 years and women ≥91 years) without selection on health or demographics (Westendorp *et al.*, 2009). In the current study, the Switchbox Leiden Study, we included 20 offspring of long-lived families from the Leiden Longevity Study together with 18 partners of the offspring as environmental and age-matched controls. The primary aim of the Switchbox Leiden Study was to compare the levels and dynamics of hormones and metabolites and their interplay between offspring of long-lived families and controls. In- and exclusion criteria were described previously in detail (Jansen *et al.*, 2015). Participants were middle-aged (52–76 years) and had a stable body mass index (BMI) between 18 and 34 kg/m^2. The Switchbox Leiden Study was approved by the Medical Ethical Committee of the Leiden University Medical Centre and was performed according to the Helsinki declaration. All participants gave written informed consent for participation.

24-hour blood sampling

The 24-hour blood sampling procedure started with placing a catheter in a vein of the forearm of the non-dominant hand, and blood withdrawal started around 9:00h (Akintola *et al.*, 2015). Samples of 2 ml serum and 1.2 ml EDTA plasma were withdrawn every 10 min. To prevent blood clotting, heparinized saline (0.9% NaCl)

was continuously infused via an infusion pump at a rate of 20 ml per hour. Before each blood withdrawal, 5 ml of saline/heparin mixed with blood was collected (without disconnecting the syringe from the blood withdrawal system) to prevent contamination of heparin/saline in the blood samples. After blood withdrawal, this 5 ml was flushed back into the subject to reduce the total amount of blood that would be withdrawn. Participants received standardized feeding consisting of 600 kcal Nutridrink (Nutricia Advanced Medical Nutrition Zoetermeer, The Netherlands) at three fixed times during the day. Participants were not allowed to sleep during the day, and except for lavatory use, no physical activity was allowed during the study period. Lights were switched off for approximately 9 hours (circa between 23:00h to 08:00h) to allow the participants to sleep.

Assays

All laboratory assays were performed with fully automated equipment and diagnostics from Roche Diagnostics (Almere, The Netherlands) at the Department of Clinical Chemistry and Laboratory Medicine of the Leiden University Medical Centre in The Netherlands.

Thyroid-stimulating hormone (TSH), cortisol, insulin, and glucose were measured in the same serum tube. Growth hormone (GH) was also measured in the same serum tube but after one additional freeze/thaw cycle. TSH and cortisol were measured by ElectroChemoLuminescence ImmunoAssay (ECLIA) using a Modular E170 Immunoanalyzer from Roche (Roche Diagnostics, Almere, The Netherlands). For TSH, the overall interassay coefficients of variation (CV) ranged in our study between 1.41–4.16%, and the overall CV of cortisol ranged between 2.4–5.1%. Human GH with a molecular mass of 22 kDa and insulin were measured using an IMMULITE® 2000 Xpi Immunoassay system (Siemens Healthcare diagnostics). The interassay CV of GH ranged between 5.4% at 5.43 mU L^{-1} and 7.2% at 25.0 mU $L^{-1,}$ and the overall CV of insulin ranged between 3.19–7.69%. Glucose was measured using Hitachi Modular P800 from Roche Diagnostics (Almere, the Netherlands), and the overall interassay CV of glucose ranged between 0.90–7.44%. If a measurement was below the detection limit, half of the lower detection limit was taken as a result.

Although ACTH was also measured, we did not take along these data in our mathematical models because this hormone was measured in EDTA plasma, so in another tube than the other hormones. However, we used ACTH data for the eyeballing, because they were instrumental for inspecting physiologically abnormal points in the cortisol data.

2.2. Physiological considerations

Since hormones are secreted in a pulsatile manner, a sudden increase is more likely to occur than a sudden decrease. Also, glucose < 2.8 mmol/L does not occur in healthy persons without an accompanying strong stress response (cortisol and GH pulses).

ACTH stimulates the secretion of cortisol. Therefore, cortisol should show a pulse following an (extreme) increase in ACTH. If an outlier is caused by sample dilution, then all hormones measured in that sample should be lower than expected. These physiological considerations could be taken into account in measurement error detection.

2.3. Methods of detecting outliers

In the following section, we will discuss four methods for outlier detection: i) eyeballing, ii) Tukey's fences, iii) Stepwise approach, and iv) the EM algorithm. The procedures of these methods are visualized in Figure 1.

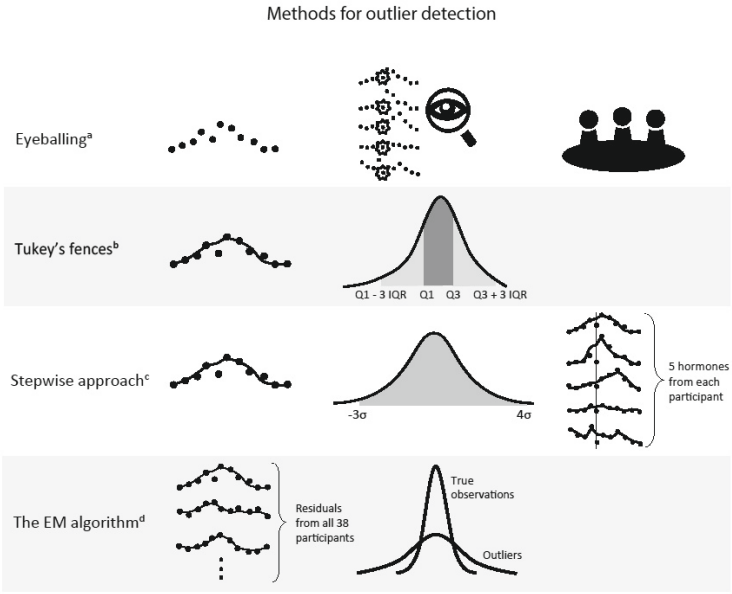


Figure 1. (a) Eyeballing detects outliers without fitting smooth curves. By visual inspection, individual experts detect outliers by taking into account that some hormones were measured in the same sample. Afterward, a consensus meeting is held, and the experts discuss all data points with conflicting detection results. (b) Tukey's fences starts with fitting a moving average curve to per-person per-hormone data and taking residuals of all data points. Then the interquartile range (IQR = Q3–Q1) of the residuals is calculated. The data points lying outside the range between Q1 − 31QR and Q3 + 3IQR are detected as outliers. (c) The stepwise approach fits the moving average curve to per-person per-hormone data, and standardized residuals of all data points are calculated (step 1). The data points lying outside the range between −3 and 4 standard deviations are detected as outliers (step 2). Then, the residuals of 5 hormones measured at the same time points are summed. When the sum of the residuals is smaller than −8, the data points are detected as outliers (step 3). Afterward, steps 1 and 3 are repeated (step 4). (d) The expectation-maximization (EM) algorithm first fits a smoothing curve to per-person per-hormone data, and the residuals are calculated. Then, all the residuals of a hormone from all 38 participants are put in the EM algorithm. The algorithm then identifies two distinguishable distributions and yields the probability of each data point being an outlier.

Eyeballing

Eyeballing was based on a visual inspection of a graphical display of individual hormone profiles from all 38 patients. This was performed by four reviewers with expert knowledge in endocrinology (EvdS, FR, OMD, and DvH). Hard copies of the 24-hour trajectories of all hormones measured per participant were provided. Three reviewers examined all 38 participants' hormone profiles, and one reviewer checked half of the participants. Information about which hormones were measured in the same tube was given verbally. Reviewers were also explicitly told that dilution of the sample may have led to measurement errors in all hormones from the same tube. After reviewing the data separately, a consensus meeting was held to reach an agreement on data points which only one (out of three or four) or two out of four reviewers had marked as an outlier.

Tukey's fences

For this algorithmic approach of outlier detection, we made the following assumptions: i) A hormone trajectory of a person follows a smooth general trend over 24 hours while measurement errors may deviate clearly from the trend, and ii) Hormone levels cannot abruptly decrease within 10 minutes. If a measurement is vastly distant from the adjacent measurements before and after, that measurement is likely to be a measurement error. Thus, by fitting a smooth curve to the data points and measuring the distance between the curve and each measurement, the algorithm can detect outliers expected to be measurement errors.

Tukey's fences is a non-parametric method developed to detect observations out of the normal range by using interquartile ranges (Tukey, 1977), and it is often used for detecting outliers in various fields (Muraleedharan *et al.*, 2016; Pham and Eggleston, 2016; Luo *et al.*, 2018; O'Brien *et al.*, 2018). Before performing Tukey's fences, the normality of the data was checked before fitting the curve. The distributions of insulin and GH data were highly skewed. Therefore, these data were log-transformed prior to applying the algorithm. Afterward, Tukey's fences was implemented using the following two steps:

I. Hormone data were smoothened over time by fitting moving average curves for every hormone per-person separately. Moving average is a method commonly applied for smoothing time series data (Montgomery *et al.*, 2015). The moving average with window size n (with n an odd number) at a certain time point is the average of the current, the -½(n-1) previous, and ½(n-1) subsequent measurements in time. In our analyses, moving averages were calculated using a window of five points. Residuals were calculated for all data points. We defined a residual as the vertical distance between an original data point and a fitted moving average curve.

II. between the first quartile and the third quartile ($Q_1 - Q_3$), and the median (Q_2) were identified. The ranges between Q_2 $k(Q_3 Q_1)$ and $Q_2 + k(Q_3 - Q_1)$ are referred to as fences. The data points that are below the lower fence or above the higher fence are identified as outliers. The value k determines the width of the fences. The larger the value of k, the lower the number of outliers that will be detected. In our analyses, we set k=3, which according to the literature, implies that the data point is "far out" (Tukey, 1977). To use the method as it was originally suggested and commonly applied, we did not adjust the value of k=3 (Horn et al., 1988; Hung and Yang, 2006; Kimenai et al., 2016).

Stepwise approach

Stepwise approach is an automatic detection process based on an algorithm that incorporates physiological knowledge and statistical methods comprising three steps as described below. We aim to detect potential outliers within a 24-hour hormone trajectory in several steps. As in Tukey's fences, the insulin and GH data were log-transformed.

I. Step 1: Fitting smoothed curves
Likewise to Tukey's fences, a moving average curve is fitted to each participant's 24-hour hormone data using a window of 5 points. By computing the distance between each data point and the fitted curve, residuals are acquired. The residuals are standardized to have a mean of 0 and a standard deviation of 1.

II. Step 2: Detecting outliers within a 24-hour hormone trajectory
Data points with standardized residuals smaller than -3 or larger than 4 are detected as outliers. The cut-off of 3 standard deviations is a commonly applied empirical rule for detecting outliers in normally distributed data. However, asymmetrical cut-offs are chosen to be more liberal for the upper boundary, as hormones are secreted in a pulsatile fashion which makes rapid increases in hormone levels biologically more plausible than rapid decreases since clearance of the hormone will occur slower. Note that this cut-off boundary is wider than the width of Tukey's fences with k=3. Furthermore, data points where glucose < 2.8 mmol/L were detected as outliers as discussed under *Physiological considerations*.

III. Step 3: The standardized residuals of all hormones measured in the same serum tube are added up for each participant. If the sum of the standardized residuals is lower than -8, all data points measured in that tube are detected as outliers. This means that the residuals of the five hormones are, on average, below the 5th percentile of standard normal distribution (1.64 standard deviation). This step allows detecting measurement errors due to the dilution of the samples. The underlying assumption is that when samples were diluted, levels of the hormones measured in the same sample are likely to all be lower at the same time point. In this step, we aim to detect these types of measurement errors which occur across the hormones.

IV. Step 4: Repeat step 1 and step 3
After all outliers detected so far are removed, a new moving average curve is fitted and step 1 and 3 are repeated once. If already detected outliers are removed, the newly fitted curves will be flatter than the fitted curve from the original data, which will allow detecting potential outliers that were missed in the previous steps.

The EM algorithm

Another approach is to estimate the probability for a data point to reflect measurement error, rather than using a dichotomous division. This starts with assuming two distinguishable data distributions: true measurement variation and background noise due to measurement errors. Based on this assumption, we expect the residuals of the true measurements to be normally distributed with standard deviations close to 0, while those of the erroneous measurements would be normally distributed with a larger standard deviation. The expectation maximization (EM) algorithm is a method that can be used to identify these two distinguishable distributions. The algorithm estimates model parameters when data is incomplete or when the model depends on a latent variable; a variable that is not directly observed but can be inferred by other observed variables (Dempster *et al*., 1977), and the method was suggested for detecting outliers (Aitkin and Wilson, 1980). The EM algorithm was applied in R version 3.5.1, using the normalmixEM function of the package mixtools (Benaglia *et al.,* 2009). In our situation, the latent variable of interest would be whether a data point is a true measurement or a measurement error. Further technical details about the EM algorithm can be found in Supplementary Material, Appendix 1.

The EM algorithm has the advantage that detected outliers do not have to be removed. Instead, the probabilities can later be used as weights for estimating outcomes, such as mean hormone levels or cross-correlations.

The outlier detection method using the EM algorithm followed the steps below. Again, insulin and GH data were log-transformed.

I. As in Tukey's fences and Stepwise approach, a moving average curve per 24-hour hormone profile for each participant was fitted. Afterward, residuals were calculated and standardized for each data point.
II. The EM algorithm was applied for each hormone with residuals of all participants together taken into account in one model.

2.4. Comparing methods on statistical outcomes

Since we do not know with certainty which data points reflect measurement errors, it is not possible to ascertain which of the four methods performed best. Therefore, we compared the number of outliers detected which were counted *per time point* and in *total data points*. In addition, the overlap in detected outliers between the four methods

was visually presented with Venn diagrams (Larsson, 2018). We chose these parameters since these descriptive statistics give a transparent description of the data and will give an insight into how removing outliers have an impact on general measures.

Furthermore, we analyzed statistical outcomes of 24-hour hormonal data before and after removing the outliers as detected by the four different methods. In this way, we could investigate whether removing outliers influenced the statistical outcome and how different methods may do so differently. Therefore, the 24-hour means, median, minima, and maxima of the five hormones were assessed, which provides a transparent description of the data and insights into how removing outliers impacts general measures. Another relevant analysis is the cross-correlation between two hormones. Cross-correlation estimates the temporal relationship between two hormonal concentrations. It is a common analysis performed with data from two simultaneously measured hormonal time series (Vis *et al.,* 2014). Therefore, it could be of interest for researchers to know to which extent measurement error would affect the estimates, especially since this method might be sensitive to the presence of outliers that co-occur in different time series data, for example, due to the dilution of a sample. Two relevant outcome measures are the strongest correlation coefficient (the maximal correlation) and the correlation coefficient at lag time 0. For the purpose of this paper, we performed cross-correlation on concentrations of glucose and insulin, which are expected to display strong cross-correlation (Feneberg *et al.*, 1999). When estimating the mean and cross-correlations after outlier removal by the EM algorithm, the weighted mean and weighed correlation are calculated, with the weight equal to the probability of each data point being an outlier. All statistical analyses were performed using the software program R, version 3.5.1.

3. Results

For each of the 38 participants, blood samples were collected at 144 time points over 24 hours, with five hormones being measured in the same serum tube. After discarding missing data, the total number of data points was 21,467. We counted detected outliers *per time point* and in *total data points*. If counted *per time point*, at least one outlier was detected in a time point among all hormones assayed in serum (i.e., glucose, insulin, TSH, cortisol, and growth hormone). In the case of a complete series, a single participant has 144 time points for each hormone. If counted in *total data points*, every data point is counted individually. In the case of a complete dataset, one participant has in total 720 data points, that is, 144 time points times five hormones.

3.1. Number of detected outliers

Table 1 summarizes the mean percentage of outliers detected per time point and in total data points. The results are averaged across 38 participants. Since the EM algorithm yields continuous probability as its outcome, we defined a data point in which its probability of being an outlier is higher than 0.9 as an outlier. For the percentage of detected outliers, we observed some differences between the four methods. Eyeballing resulted in the smallest percentage of detected outliers both per time point (mean=1.7%) as well as for total data points (1.0%), followed by Stepwise approach (per time points: 5.1%, total data points: 2.7%). Tukey's fences yielded more outliers per time point (9.3%) but a similar percentage in total data points (2.3%). The EM algorithm method yielded the largest percentage of outliers (per time points: 40.3%, total data points: 11.0%).

In Figure 2, the numbers of detected outliers for each hormone averaged over all participants are presented. The EM algorithm and Tukey's fences both detected more outliers in cortisol and GH compared to other hormones. Eyeballing and Stepwise approaches detected a similar number of outliers across the different hormones.

Table 1. The percentage of time points with at least one detected outlier among the hormones measured, and the percentage of total data points detected as outliers among the same set of hormones. The mean and standard deviation of the 38 participants are given.

	mean (sd); n=38	
	Time points detected to contain an outlier (%)	**Total data points detected to be outliers (%)**
Eyeballing	1.7 (2.1)	1.0 (1.4)
Tukey's fences	9.3 (5.6)	2.3 (1.4)
Stepwise approach	5.1 (1.5)	2.7 (1.5)
EM algorithm*	40.3 (7.7)	11.0 (2.8)

*For the EM algorithm results, the measurement points where the probability of being an outlier > 0.9 was counted.

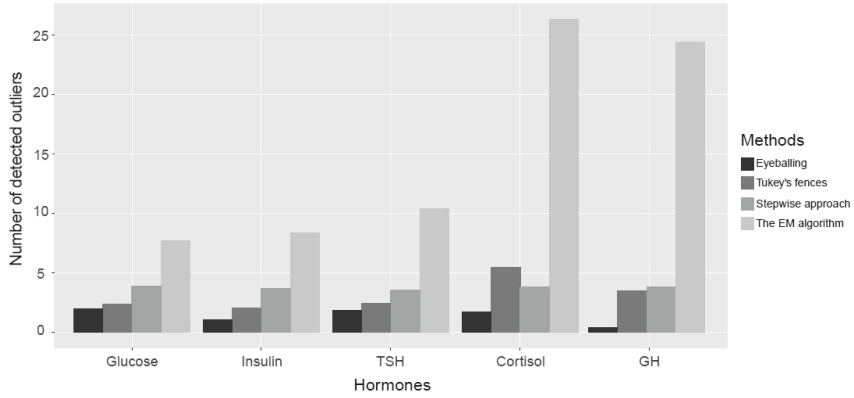


Figure 2. The mean number of data points detected per hormone per method across all participants.

3.2. Overlap in detected outliers

Figure 3 displays Venn diagrams presenting the number of outliers detected by eyeballing, Stepwise approach, and Tukey's fences and their overlap. We did not include the results of the EM algorithm in the Venn diagrams for two reasons i) the EM algorithm detected an implausibly large number of outliers (per time point=1,590 and in total data points =2,728), and (ii) three sets of data is the maximum to draw a proportional Venn diagram in two-dimensional space. Figure 3a presents the number of outliers per time point, and Figure 3b presents that of the total data points. In Figure 3a, most of the outliers detected by eyeballing were also detected by the other two methods, while the overlap is larger with Stepwise approach. In Figure 3b, the overlap between eyeballing and Stepwise approach is again larger than the overlap between eyeballing and Tukey's fences. Here, Stepwise approach and Tukey's fences detected a similar number of outliers. However, the overlap is relatively small, which indicates that they are detecting different data points. Eyeballing detected 47 total data points, which were not detected by Stepwise approach or Tukey's fences. Among outliers per time point detected by eyeballing, Stepwise approach, and Tukey's fences, 95.8% overlapped with the outliers detected by the EM algorithm (data not shown). Additionally, 70.1% of the total data points detected by the three methods overlapped with the outliers detected by the EM algorithm (data not shown).

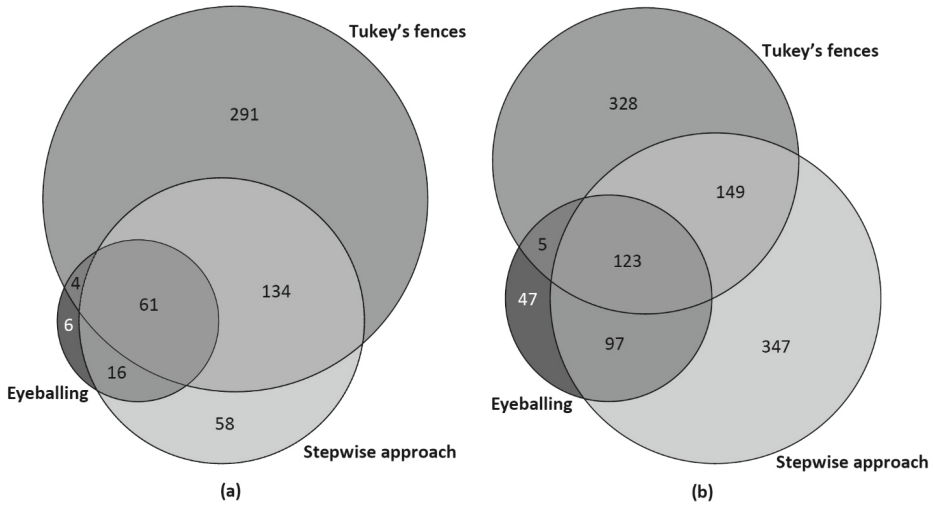


Figure 3. Venn diagrams visualizing the number of measurement errors detected by each method (eyeballing, Stepwise approach, and Tukey's fences) and their overlap counted in total time points (a) and in all data points (b). The overlap with the EM algorithm is not presented here for the reasons mentioned in the results section.

3.3. Representative 24-hour hormone figures presented with detected outliers

Figures 4a-d display the detected outliers in glucose, insulin, TSH, cortisol, and GH for eyeballing, Tukey's fences, Stepwise approach, and the EM algorithm, respectively in one representative participant. By eyeballing (Figure 4a), four data points are detected as outliers in glucose, TSH, and cortisol, and these four outliers are all in the same time points. Of these four time points, outliers in insulin were detected in three time points and GH in one time point. Tukey's fences (Figure 4b) detected the same outliers for glucose, insulin, TSH, and cortisol but detected several more than eyeballing. In both TSH and cortisol between time points 110 to 130, several points that are biologically unlikely to be measurement errors were detected. No outliers were detected in GH. Stepwise approach (Figure 4c) identified the same outliers as eyeballing. However, several extra points were detected as well. Here in several time points (42nd, 76th, and 114th), outliers were detected in all hormones, which is a result of Step 3 of the Stepwise approach. The EM algorithm (Figure 4d, note that the points are only marked if the probability of being an outlier is higher than 0.9) resulting in many detected outliers in the pulses that are unlikely to be outliers. Remarkably in GH, data points close to detection limits were detected as outliers.

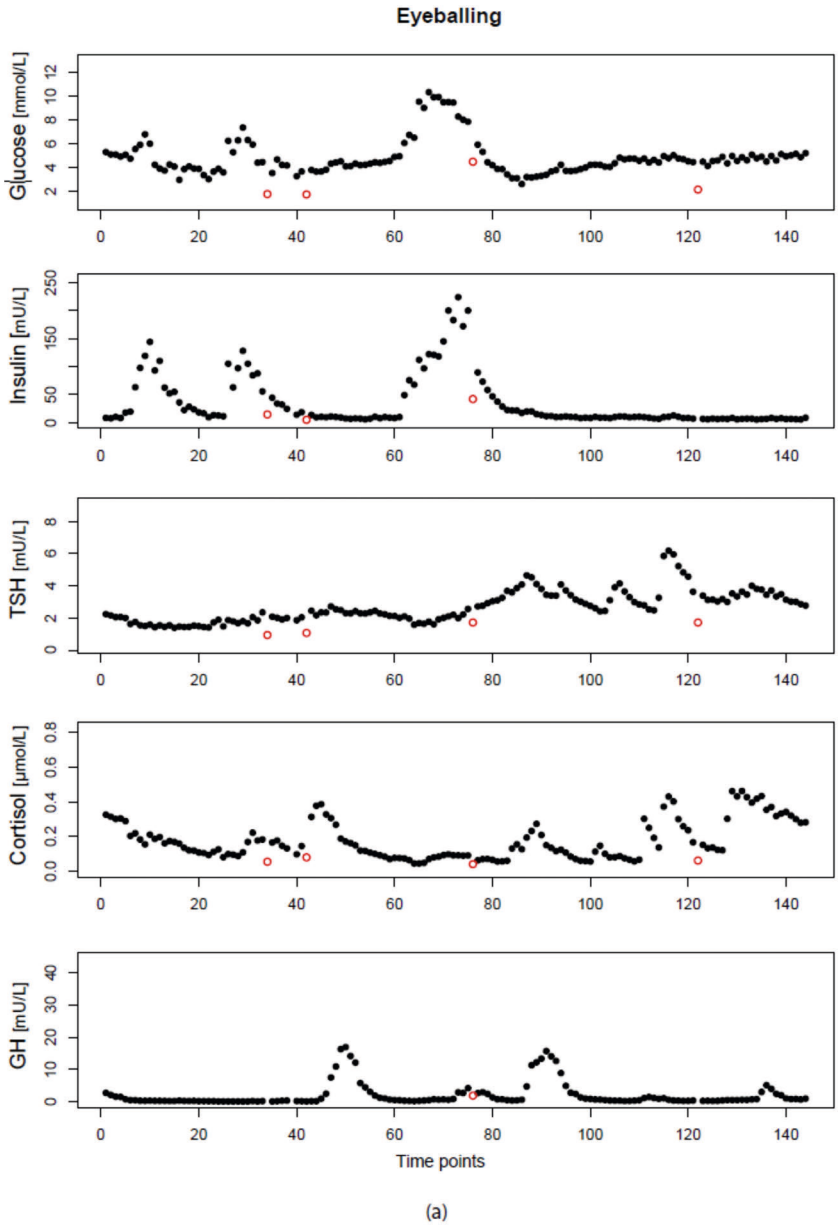


Figure 4a. The results of outlier detection by eyeballing in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate detected outliers.

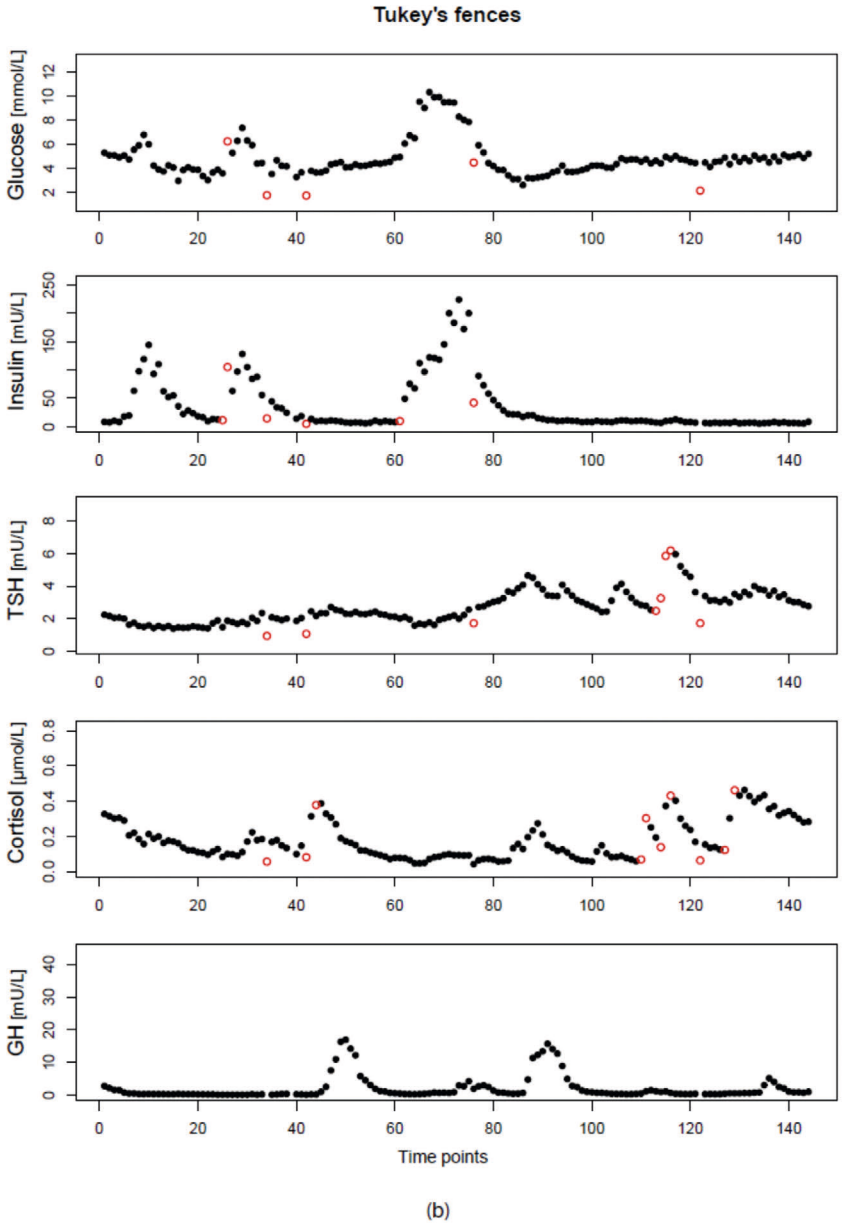


Figure 4b. The results of outlier detection by Tukey's fences in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate detected outliers.

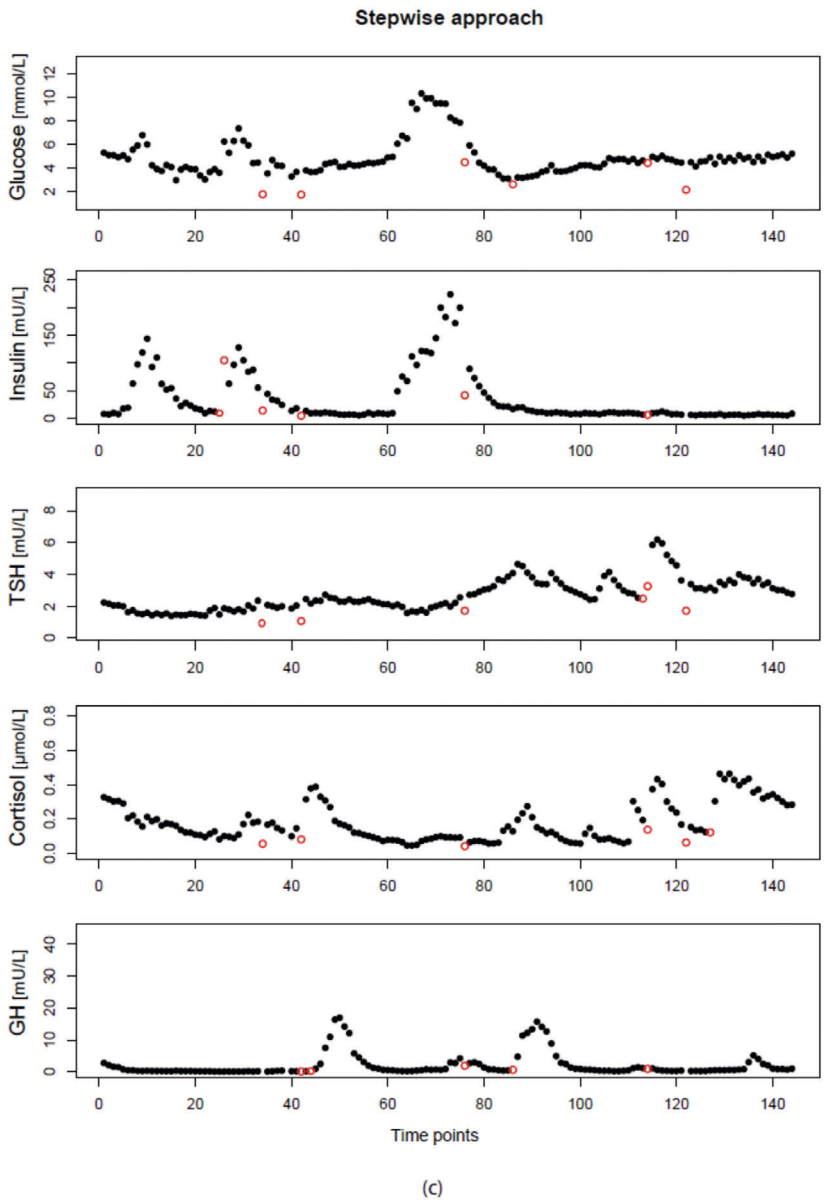


Figure 4c. The results of outlier detection by Stepwise approach in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate detected outliers.

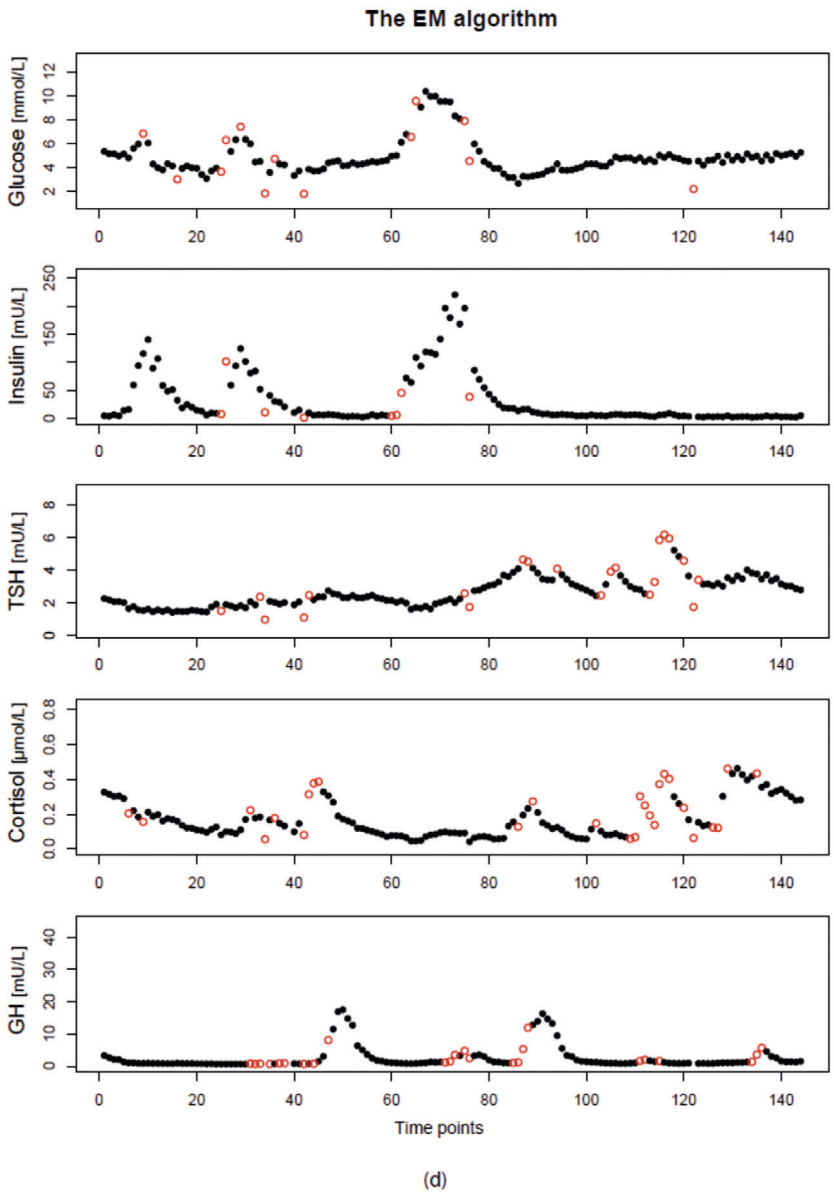


Figure 4d. The results of outlier detection by the EM algorithm in glucose, insulin, TSH, cortisol, and growth hormone of participant 19. Red hollow data points indicate the probability of the data point being an outlier is higher than 0.9.

3.4. Effects of removing outliers on statistical outcomes

Descriptive statistics: 24-hour mean, median, minimum, and maximum

The mean, median, minimum, and maximum values for every hormone were calculated over time before and after removing outliers detected by the four methods. This is shown in Table 2. Mean and median values did not change substantially after outlier removal. Minimum values changed for glucose and TSH after removing outliers by all four methods, while in insulin, the value did not change much after eyeballing. The EM algorithm had the largest influence on maximum values in all hormones.

Cross-correlation of glucose and insulin

In Table 3 cross-correlations between glucose and insulin are presented before and after removing outliers. Overall, removing outliers did not have a major influence on the cross-correlation of glucose and insulin, and on the lag time at the maximum cross-correlation. Figure 5 shows the individual changes in correlation at lag time 0. In Figure 5, we observe large differences between participants. Especially the first participant shows a big change in correlation after removing outliers by all methods. Overall, the changes after eyeballing, Tukey's fences, and Stepwise approach were mostly small, and the changes were not in one direction dominantly. However, after removing outliers detected by the EM algorithm, cross-correlation decreased in most cases.

Table 2. Mean, median, minimum, and maximum values for glucose, insulin, TSH, cortisol, and growth hormone in 24 hours, before (raw data) and after outlier removal (Eyeballing, Tukey's fences, Stepwise approach, and the EM algorithm). The mean and standard deviation of the 38 participants are given.

	mean (sd); n=38											
	Glucose [mmol/L]				Insulin [mU/L]				TSH [mU/L]			
	Mean	**Median**	**Min**	**Max**	**Mean**	**Median**	**Min**	**Max**	**Mean**	**Median**	**Min**	**Max**
Raw data	5.09 (.36)	4.80 (.39)	2.76 (.70)	9.51 (1.52)	19.90 (10.11)	9.66 (5.51)	2.76 (2.41)	91.61 (54.41)	2.02 (1.05)	1.92 (1.01)	1.01 (.61)	3.57 (1.89)
Eyeballing	5.11 (.36)	4.81 (.39)	3.16 (.53)	9.48 (1.47)	19.96 (10.14)	9.66 (5.52)	2.80 (2.46)	91.61 (54.41)	2.03 (1.05)	1.93 (1.02)	1.21 (.67)	3.57 (1.89)
Tukey's fences	5.07 (.37)	4.80 (.39)	3.04 (.62)	9.21 (1.42)	19.96 (10.16)	9.69 (5.52)	3.39 (2.39)	91.34 (54.69)	2.02 (1.04)	1.93 (1.02)	1.19 (.63)	3.49 (1.82)
Stepwise approach	5.12 (.36)	4.80 (.39)	3.29 (.41)	9.40 (1.48)	20.47 (10.43)	10.21 (5.86)	3.54 (2.31)	91.03 (54.26)	2.02 (1.05)	1.92 (1.01)	1.21 (.64)	3.50 (1.81)
EM algorithm*	5.00 (.37)	4.77 (.40)	3.14 (.58)	9.08 (1.47)	20.05 (10.35)	10.16 (5.97)	3.74 (2.44)	87.65 (49.73)	1.98 (1.01)	1.90 (1.00)	1.24 (.73)	3.33 (1.64)
	Cortisol [µmol/L]				GH [mU/L]							
	Mean	**Median**	**Min**	**Max**	**Mean**	**Median**	**Min**	**Max**				
Raw data	.21 (.05)	.18 (.05)	.05 (.03)	.57 (.09)	2.49 (1.51)	.95 (.94)	.16 (.22)	20.63 (10.31)				
Eyeballing	.21 (.05)	.18 (.05)	.05 (.03)	.57 (.09)	2.48 (1.58)	.95 (.94)	.16 (.22)	20.63 (10.31)				
Tukey's fences	.20 (.05)	.18 (.05)	.05 (.03)	.55 (.09)	2.47 (1.55)	.95 (.94)	.17 (.22)	20.27 (10.67)				
Stepwise approach	.21 (.05)	.18 (.05)	.05 (.03)	.56 (.09)	2.51 (1.54)	.96 (.95)	.17 (.22)	20.27 (10.59)				
EM algorithm*	.18 (.04)	.16 (.05)	.05 (.03)	.50 (.08)	2.24 (1.48)	.94 (1.02)	.18 (.22)	18.90 (11.13)				

* For the EM algorithm results, weighted mean and standard deviation is used.

Table 3. Cross correlations between glucose and insulin. Mean and standard deviation across 38 participants.

	mean (sd); n=38		
	Correlation at lag time 0	**Maximum cross corr.**	**Lag time at maximum cross corr. (min)**
Raw data	0.74 (0.12)	0.74 (0.12)	-4.7 (7.3)
Eyeballing	0.74 (0.11)	0.75 (0.12)	-5.3 (7.6)
Tukey's fences	0.73 (0.14)	0.74 (0.14)	-6.3 (8.2)
Stepwise approach	0.74 (0.12)	0.75 (0.12)	-5.0 (8.0)
EM algorithm*	0.71 (0.12)	0.73 (0.17)	-9.5 (9.8)

*For the EM algorithm results, weighted correlation is used.

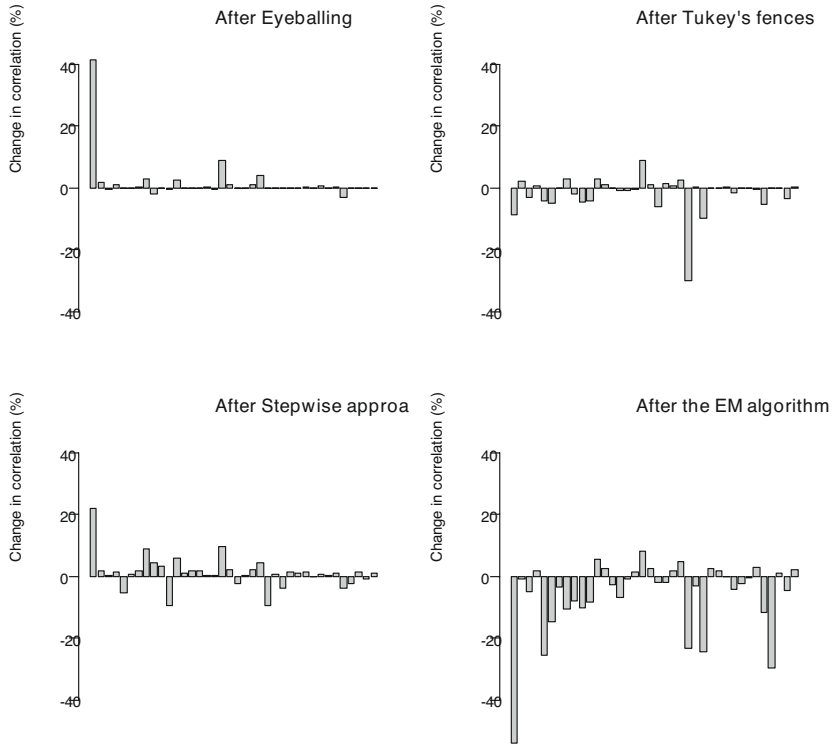


Figure 5. Change in correlation at lag time 0 (%) after removal of measurement errors detected by the four methods; eyeballing, Tukey's fences, Stepwise approach, and the EM algorithm. Each bar represents an individual participant.

4. Discussion

In this study, we aimed to evaluate and compare different methods to detect outliers in 24-hour hormonal data since no specific methods were routinely available for this purpose. We assumed that measurement errors would deviate largely from the physiological curves of hormones. By identifying outliers in the data, therefore, we expected to detect likely measurement errors. The main outcomes of this study were that human-judgement (eyeballing) defined fewer data points as an outlier than the other three automatic approaches. Among the automatic approaches, the data-driven methods (Tukey's fences and the EM algorithm) were prone to detect more outliers likely to be true measurements than the method involving subject-specific knowledge (Stepwise approach). The mean, minima, and maxima of the hormones did not change much after removing outliers. However, the minima of glucose and TSH did change, and the EM algorithm had a large influence on maximum values in all hormones. The effect of removing outliers on the correlation between glucose and insulin can be large within an individual but had no major impact on a group level.

A relatively low number of outliers were detected by eyeballing. This may be an advantage of this method, as only truly deviating points will be discarded in the analysis. Another advantage of eyeballing is that the data points detected as outliers are based on physiological arguments and are not data-driven. This allows eyeballing to detect (i) a sequence of data points that was physiologically implausible to display the same pattern in several hormones, and (ii) outliers at the beginning or end of a time series. These types of outliers cannot be detected by fitting smoothing curves, which explains the 47 data points that were exclusively detected by eyeballing, and not by Stepwise approach or Tukey's fences. However, a disadvantage of eyeballing is that it is time-consuming and depends on individual reviewers' background knowledge and subjective decision. If the number of reviewers is large enough and a consensus meeting is held, the precision may increase. However, the amount of time to reach a unanimous decision would take longer. Also, eyeballing is a one-off process that cannot be generalized to other settings.

Although Tukey's fences are advocated as a non-parametric approach, the method did not perform well in our case when applied with moving median curves instead of moving average curves. Especially when the hormone profile is mostly flat with sudden pulses, such as GH, Tukey's fences with moving median curves detected a biologically implausible number of outliers (54.6% of the total data points). Therefore, when using Tukey's fences to detect outliers, we suggest researchers to be aware of the type of their data and smoothing methods.

We introduced Stepwise approach as a new method to detect measurement errors in 24-hour hormonal data. The advantage of Stepwise approach is that by using the

standardized residuals, it facilitates the detecting of measurement errors caused by dilution, which may not have been identified by only looking into individual hormones. Additionally, it is expected to be a more objective method than eyeballing, as it explicitly incorporates the information from multiple hormones and applies the same cut-off values of standard deviations to every hormone. Furthermore, it is less time-consuming than eyeballing and can relatively easily be applied to different hormonal datasets. Compared to Tukey's fences, Stepwise approach has more flexibility to incorporate physiological knowledge, such as adopting asymmetrical cut-off or removing glucose measurements lower than 2.8 mmol/L. However, the performance of the method may depend on parameters such as a time window for moving average, or cut-off points of standard deviations. These parameters still require decisions and need to be chosen with care; the decisions should also be clearly reported. Another disadvantage of Stepwise approach, which also applies to eyeballing and Tukey's fences, is that we discard data according to a dichotomous division. Whether a data point is an outlier or not is often dependent on the degree of belief instead of a clear dichotomous distinction. Furthermore, this dichotomous distinction reduces the statistical power in further analyses.

The strength of the EM algorithm is that, instead of the dichotomous distinctions, it gives probabilities of each point being an outlier. Therefore, we acquire extra information which can be incorporated into further analysis, such as for probability weighting. Additionally, the EM algorithm requires less prior knowledge compared to the previously discussed methods. However, a critical disadvantage of the EM algorithm is that we cannot ensure whether the two identified distributions are actually distinguishing outliers and non-outliers. In our dataset, the detected points were often not plausible to be detected as outliers from a physiological perspective.

It is worth to mention the performances of Tukey's fences, Stepwise approach, and the EM algorithm depend on which smoothing technique is applied. Moving average, which was used in the study, does not require extensive modeling and can capture local fluctuations of hormone concentration. However, it may smooth out the transient increase of hormone concentration and lead to detect true measurements as outliers. Stepwise approach takes this shortcoming of moving the average into account by setting different cut-off values for positive and negative residuals. There are more advanced model-based smoothing techniques, such as deconvolution analysis, which takes the underlying dynamics of hormone secretions into account (Brown et al., 2001; Faghih et al., 2014). These methods were not considered in this study as our aim was to compare outlier detection methods that could be easily adopted by applied researchers in a pre-analysis phase.

To test the efficacy of the outlier detection methods, we simulated 24-hour hormonal data and measurement errors as comparable as possible to real data. The advantage of

the simulation study is that we know which data points are true measurement errors. We compared the performance of Stepwise approach, Tukey's fences, and the EM algorithm. The simulation description and the results are attached as an appendix (see Supplementary Material, Appendix 2). The EM algorithm resulted in a high percentage of true measurements wrongly detected as errors, especially when a simulated hormone has a higher variation during the day than during the night. Most methods yielded relatively low percentages of true error detected. This could be due to the fact that some simulated errors are close to fitted curves, while the methods we are comparing are based on detecting errors deviating from the curves. For detecting dilution errors, Stepwise approach performed better than other methods. This is because Stepwise approach could detect dilution errors that were not deviating much from the curves by taking the sum of the residuals from all hormones.

In this study, the effect of removing outliers on the cross-correlation between glucose and insulin had no major impact on a group level. Note that these results may not be generalized to other statistical outcomes, such as deconvolution analysis and approximate entropy analysis, which are also common analyses for 24-hour hormonal data. Furthermore, glucose and insulin are strongly cross-correlated; however, when two hormones are less strongly correlated, the impact of removing outliers may be higher.

5. Conclusions

Based on our results, we generally recommend methods that incorporate physiological knowledge over data-driven methods. The EM algorithm is not recommended for outlier detection in 24-hour hormonal data since the method seems to falsely distinguish true biological variations due to circadian factors, such as meal response or day-night differences, as outliers. Tukey's fences, the other data-driven method, is not recommended in 24-hour hormonal data. Since no statistical assumptions have to be made and fewer data points will be removed, eyeballing could be a good method for detecting outliers. However, since it is time-consuming (depending on the number of participants studied), it might not always be practical. The strengths and limitations of each method are presented in Table 4.

Table 4. Methods for detecting measurement errors

	Eyeballing	Tukey's fences	Stepwise approach	The EM algorithm
Underlying assumptions	• Researchers' expert knowledge is reliable.	• From how much standard deviations (or interquartile range) away from a smoothing curve is considered to be an outlier should be decided by researchers.		• Normal distributions
Efficiency and generalizability of the method	• Relatively time-consuming process. • Different experts' knowledge is required for different types of data.	• Although it needs several adjustments for different types of time series (e.g., parameters for smoothing curves), the processes can be easily applied to different settings.		
Limitations	• Explicit knowledge and clear physiological reasoning behind the detection process. • Disagreement between experts may happen.	• The method is highly affected by smoothing techniques and type of data, especially when the hormone levels are mostly constant over time.	• Measurement error within a hormone and within a sampling method (serum) can both be detected	• Yields a probability • Need a large sample to be able to distinguish between the distributions

In conclusion, we recommend Stepwise approach for detecting outliers in serial 24-hour hormonal data since this method combines both physiological knowledge and an automated process. However, decisions such as which cut-offs of standard deviation should be applied or which hormones can be used together in the method should be supported by solid physiological knowledge. Stepwise approach is especially suitable for data of several hormone measurements from the same tube and when dilution is a possible cause of measurement errors. In this case, the outlier detection process can improve by taking along a reference measurement together with the hormonal measurements, whose concentration is stable over the day, such as creatinine or urea.

Although the methods showed different performances in outlier detection, this had little impact on the statistical outcomes. Overall, 24-hour means and cross-correlations did not materially change, but on an individual basis, correlations might change. The influence of outliers may depend on the study's sample size and outcome of interest. We recommend researchers be aware of the potential influence of measurement errors in their study and consciously decide which method to choose for outlier detection and whether it is necessary to remove outliers at all.

References

1. Aitkin M, and Wilson GT (1980) Mixture Models, Outliers, and the EM Algorithm. Technometrics 22:325-331.
2. Akintola AA, Jansen SW, Wilde RB, Hultzer G, Rodenburg R, and van Heemst D (2015) A simple and versatile method for frequent 24 h blood sample collection in healthy older adults. MethodsX 2:33-38.
3. Benaglia T, Chauveau D, Hunter DR, and Young D (2009). mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29. URL http://www.jstatsoft.org/v32/i06/.
4. Brown EN, Meehan PM, and Dempster AP (2001) A stochastic differential equation model of diurnal cortisol patterns. American Journal of Physiology-Endocrinology and Metabolism 280:E450-E461.
5. Dempster AP, Laird NM, and Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological):1-38.
6. Faghih RT, Dahleh MA, Adler GK, Klerman EB, and Brown EN (2014) Deconvolution of Serum Cortisol Levels by Using Compressed Sensing. PLOS ONE 9:e85204.
7. Feneberg R, Sparber M, Veldhuis JD, Mehls O, Ritz E, and Schaefer F (1999) Synchronous fluctuations of blood insulin and lactate concentrations in humans. J Clin Endocrinol Metab 84:220-227.
8. Grossmann M, Wong R, Szkudlinski MW, and Weintraub BD (1997) Human thyroid-stimulating hormone (hTSH) subunit gene fusion produces hTSH with increased stability and serum half-life and compensates for mutagenesis-induced defects in subunit association. J Biol Chem 272:21312-21316.
9. Grubbs FE (1969) Procedures for Detecting Outlying Observations in Samples. Technometrics 11:1-21.
10. Horn PS, Britton PW, and Lewis DF (1988) On The Prediction of a Single Future Observation from a Possibly Noisy Sample. Journal of the Royal Statistical Society Series D (The Statistician) 37:165-172.
11. Hung W-L, and Yang M-S (2006) An omission approach for detecting outliers in fuzzy regression models. Fuzzy Sets and Systems 157:3109-3122.
12. Jansen SW, Akintola AA, Roelfsema F, van der Spoel E, Cobbaert CM, Ballieux BE, Egri P, Kvarta-Papp Z, Gereben B, Fekete C, Slagboom PE, van der Grond J, Demeneix BA, Pijl H, Westendorp RG, and van Heemst D (2015) Human longevity is characterised by high thyroid stimulating hormone secretion without altered energy metabolism. Sci Rep 5:11525.
13. Kimenai DM, Henry RM, van der Kallen CJ, Dagnelie PC, Schram MT, Stehouwer CD, van Suijlen JD, Niens M, Bekers O, Sep SJ, Schaper NC, van Dieijen-Visser MP, and Meex SJ (2016) Direct comparison of clinical decision limits for cardiac troponin T and I. Heart 102:610-616.
14. Larsson J (2018) eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses. R package version 4.1.0. In.
15. Luo J, Frisken S, Machado I, Zhang M, Pieper S, Golland P, Toews M, Unadkat P, Sedghi A, Zhou H, Mehrtash A, Preiswerk F, Cheng C-C, Golby A, Sugiyama M, and Wells WM (2018) Using the variogram for vector outlier screening: application to feature-based image registration. International Journal of Computer Assisted Radiology and Surgery.

16. Montgomery DC, Jennings CL, and Kulahci M (2015) Introduction to time series analysis and forecasting. John Wiley & Sons.

17. Muraleedharan G, Lucas C, and Guedes Soares C (2016) Regression quantile models for estimating trends in extreme significant wave heights. Ocean Engineering 118:204-215.

18. O'Brien JD, Kahn RM, Zenko Z, Fernandez JR, and Ariely D (2018) Naïve models of dietary splurges: Beliefs about caloric compensation and weight change following non-habitual overconsumption. Appetite 128:321-332.

19. Odell WD, Utiger RD, Wilber JF, and Condliffe PG (1967) Estimation of the secretion rate of thyrotropin in man. J Clin Invest 46:953-959.

20. Oike H, Oishi K, and Kobori M (2014) Nutrients, Clock Genes, and Chrononutrition. Curr Nutr Rep 3:204-212.

21. Pham NM, and Eggleston K (2016) Prevalence and determinants of diabetes and prediabetes among Vietnamese adults. Diabetes Research and Clinical Practice 113:116-124.

22. Porksen N, Nyholm B, Veldhuis JD, Butler PC, and Schmitz O (1997) In humans at least 75% of insulin secretion arises from punctuated insulin secretory bursts. Am J Physiol 273:E908-914.

23. Roelfsema F, Boelen A, Kalsbeek A, and Fliers E (2017) Regulatory aspects of the human hypothalamus-pituitary-thyroid axis. Best Pract Res Clin Endocrinol Metab 31:487-503.

24. Spiga F, Walker JJ, Gupta R, Terry JR, and Lightman SL (2015) 60 YEARS OF NEUROENDOCRINOLOGY: Glucocorticoid dynamics: insights from mathematical, experimental and clinical studies. J Endocrinol 226:T55-66.

25. Tukey JW (1977) Exploratory data analysis. Reading, Mass.

26. Veldhuis J, Yang R, Roelfsema F, and Takahashi P (2016) Proinflammatory Cytokine Infusion Attenuates LH's Feedforward on Testosterone Secretion: Modulation by Age. J Clin Endocrinol Metab 101:539-549.

27. Vis DJ, Westerhuis JA, Hoefsloot HC, Roelfsema F, van der Greef J, Hendriks MM, and Smilde AK (2014) Network identification of hormonal regulation. PLoS One 9:e96284.

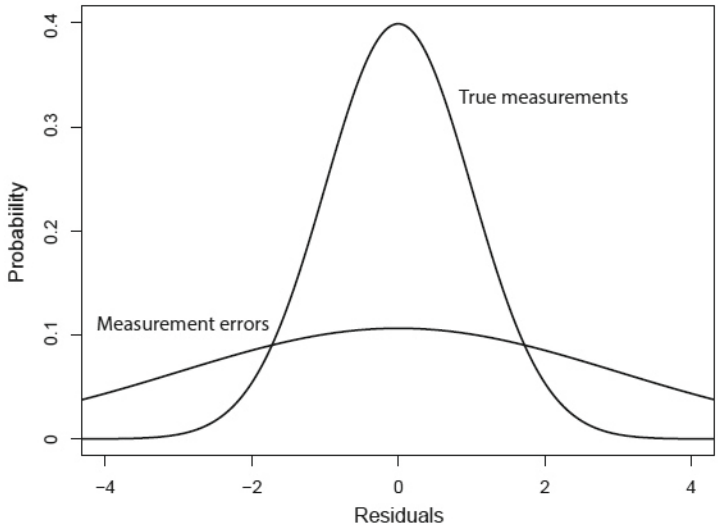
28. Westendorp RG, van Heemst D, Rozing MP, Frolich M, Mooijaart SP, Blauw GJ, Beekman M, Heijmans BT, de Craen AJ, Slagboom PE, and Leiden Longevity Study G (2009) Nonagenarian siblings and their offspring display lower risk of mortality and morbidity than sporadic onagenarians: The Leiden Longevity Study. J Am Geriatr Soc 57:1634-1637.

Appendix 1

Details on the EM algorithm to detect outliers

For each of the hormones separately, the EM algorithm was applied to the residuals of all subjects simultaneously, where the residual of the ith measurement of subject j was calculated as $R_{ij}= Y_{ij}- \hat{Y}_{ij}$, with $Y_{ij,}$ the observed measurement and $\hat{Y}_{ij}$, the moving average smoothed estimate.

We assumed that there were two types of measurements: true measurements and erroneous measurements. We expected that the residuals of the true measurements had standard deviations close to 0, while erroneous measurements had a much larger standard deviation.



The (unobserved) indicator variable Z denotes whether a measurement is an error, with Z_{ij}=1 if the ith measurement of subject j is an error and Z_{ij} =0 if it is a true measurement. The proportion of erroneous measurements $Pr(Z_{ij}=1)$ is denoted by π_e . We assumed that residuals R of true measurements were normally distributed with mean 0 and standard deviation σ_1 while the residuals of the erroneous measurements were normally distributed with mean 0 and standard deviation σ_2 , with $\sigma_2 >> \sigma_1$. The proportion of erroneous parameters π_e and the standard deviations σ_1 and σ_2, can be estimated using maximum likelihood. The complete likelihood of the data is

$$L(\sigma_1, \sigma_2; R, Z) = \prod_{ij} f(R_{ij};\ \sigma_1)^{(1-Z_{ij})}\ f(R_{ij};\ \sigma_2)^{Z_{ij}},$$

with $f(R_{ij}; \sigma_i)$, the normal density with mean 0 and standard deviation σ_i . Because the Z_{ij} are unobserved, the EM algorithm is applied with following EM steps:

E step: given current estimates p_e s_1 and s_2 for π_e σ_1 and σ_2 , the expected probability of being an error is estimated using Bayes formula:

$$\Pr(Z_{ij}=1|\ R_{ij}) = \frac{p_e f(R_{ij};\, s_2)}{(1-p_e) f(R_{ij};\, s_1)+p_e f(R_{ij};\, s_2)} \quad (1)$$

M step: the likelihood function where the Z_{ij} are replaced by the expected probabilities that Z_{ij} is 1, is maximized.

The EM steps are repeated until convergence. The final estimates p_e, s_1, and s_2 are filled in in equation (1). This yields for each measurement an estimated probability of being an error measurement.

The EM algorithm was applied in R version 3.5.1, using the normalmixEM function of the package mixtools.

Reference

Benaglia T, Chauveau D, Hunter DR, Young D (2009). mixtools: An R Package for Analyzing Finite Mixture Models. Journal of Statistical Software, 32(6), 1-29. URL http://www.jstatsoft.org/v32/i06/.

Appendix 2

Detecting outliers in 24-hour hormonal data: a simulation study

1. Data generation

We simulate measurements for five hormones; glucose, insulin, thyroid stimulating hormone (TSH), cortisol, and growth hormone (GH), according to their physiological characteristics and the laboratory setting where our sample was drawn. This setting was reproduced in simulation as described below:

- 24 hours with measurements every 10 minutes, in total 144 measurements per hormone and person.
- Three meals at time 0, 18, 54.
- Night from time 84 to time 138.

For each hormone, we generated measurements. The mean hormone value at time t, Y(t) consisted of a constant baseline level and one or more peaks using an absorption/elimination model. A peak starting at t_s has the form:

$$Y(t) = C_0 + C_1 (t>t_s)(e^{-\lambda_e t} - e^{-\lambda_a t}),$$

where C_0 determines the minimum hormone values over time, C_1 the peak value and λ_a and λ_e the rate of absorption and elimination of the hormone in the blood. The latter is directly related to the half-life of the hormone by λ_e=ln(2)/half-life. Random between and within-person variation was added to the generated mean values. The specific minimum, location and duration of peaks, and the random intra/inter-person variation were based on the observed patterns in our data. Specific features of each hormone are:

- Glucose: Three clear peaks after meals, where the third one is slightly higher than others. At night, the hormone level is stable and low, and the variation is smaller. Physiologically, glucose levels cannot be below 2.8 mmol/L.
- Insulin: Three clear peaks after each meal, and the hormone is highly correlated with glucose (corr.=0.75). At night, the hormone level is stable and low, and the variation is smaller.
- TSH: One prominent peak, where the hormone builds up in the evening from 6 pm (t=54) with the highest levels at 11 pm (t=84), with large variation.
- Cortisol: Peaks at the end of the night.
- GH : Sharp peaks and the number of peaks varies from 0 to 20 across the individuals.

Inter-person variation is generated by varying the highest concentration reached during peaks, following a normal distribution (specific parameters are provided in the table

below). For TSH, cortisol, and GH, the location of the peaks also varies across people. In this way, we generated 24-hour hormonal data for 38 simulated subjects. Table A1 shows the specific parameters used for simulating the 24-hour hormonal data of 38 individuals.

In each individual, for each hormone, we generated measurement errors at 14 time points. To generate random measurement errors in each hormone at seven randomly selected time points (5% out of 144 points), we replaced the true measurement with an error measurement drawn from a uniform distribution with a wide range (-10 x intra-person SD to 15 x intra-person SD). Furthermore, we generated related dilution errors at seven time points which were the same across all hormones for one individual. The dilution errors were generated by dividing the original measurement by 2.

Table A1. Parameters for generating 24-hour glucose, insulin, TSH, cortisol and GH data

	Glucose [mmol/L]	Insulin [mU/L]	TSH [mU/L]	Cortisol [μmol/L]	GH [mU/L]
Starting value (C_0)	3.8	6.6	1	0.05	1
Number and location of peaks	3 peaks, increase starts at mealtimes	3 peaks, increase starts at mealtimes	One wide peak, increase starts between t=45 and 65	3 peaks, Increase starts between (i) t=75 and 100, (ii) between t=100 and 124, and (iii) between 124 and 140	0 to 20 peaks, increase starts from t=0 and 143
Half-life	35 min	35 min	120 min	50 min	10 min
Intra-person variation (SD)	Day 0.50, Night 0.25	Day 6.5 Night 3.2	0.17	0.03	0.27
Mean and SD of peaks: first peak (i), second peak (ii), third peak (iii), with inter person Sd	(i) 4 (0.5), (ii) 4 (0.5), (iii) 7 (0.7)	90 (5)	2.5 (0.5)	(i) 0.3 (0.1), (ii) 0.4 (0.1), (iii) 0.5 (0.1)	15 (1)
Remarks	Values <2.8 are set to 2.8	Values <2.8 are set to 2.8	Values <1 are set to 1	Values <0.05 are set to 0.05	Values <0.2 are set to 0.2
Absorption/ elimination rate	$\lambda_a = 1.1\,\lambda_e$	$\lambda_a = 1.1\,\lambda_e$	$\lambda_a = 1.1\,\lambda_e$	$\lambda_a = 2\,\lambda_e$	$\lambda_a = 1.1\,\lambda_e$
Comments		Log transformation			Log transformation

2. Simulation results

Figure A1 shows simulated 24-hour hormonal data for glucose, insulin, TSH, cortisol, and GH of the first two generated individuals are shown. The hormone-specific measurement errors are indicated by a red dot. The dilution errors are indicated by a green dot.

Figure A2 displays how many points are indicated as measurement errors by each method averaged across the 38 simulated subjects. The EM algorithm indicated the highest number of measurement errors, followed by the stepwise approach. Especially for the hormones where the intra-person variation was larger during the day than during the night (glucose and insulin), the EM algorithm indicated high numbers of measurement errors.

Table A2 shows what percentage of true errors (random errors and dilution errors) were detected by each method and how many non-errors were identified as errors by each method. When it comes to detecting a true error, the EM algorithm performed best. However, the EM algorithm also indicated the most non-errors as measurement errors. Especially for insulin, the number of true measurements falsely indicated as errors was extremely high. This is explained by the fact that the intra-person variation in insulin differed between day and night, and the insulin residuals were not normally distributed without log transformation. The percentage of non-error detected as measurement error was much lower in Stepwise approach and Tukey's fences than in the EM algorithm. Stepwise approach is to be preferred when detecting dilution errors.

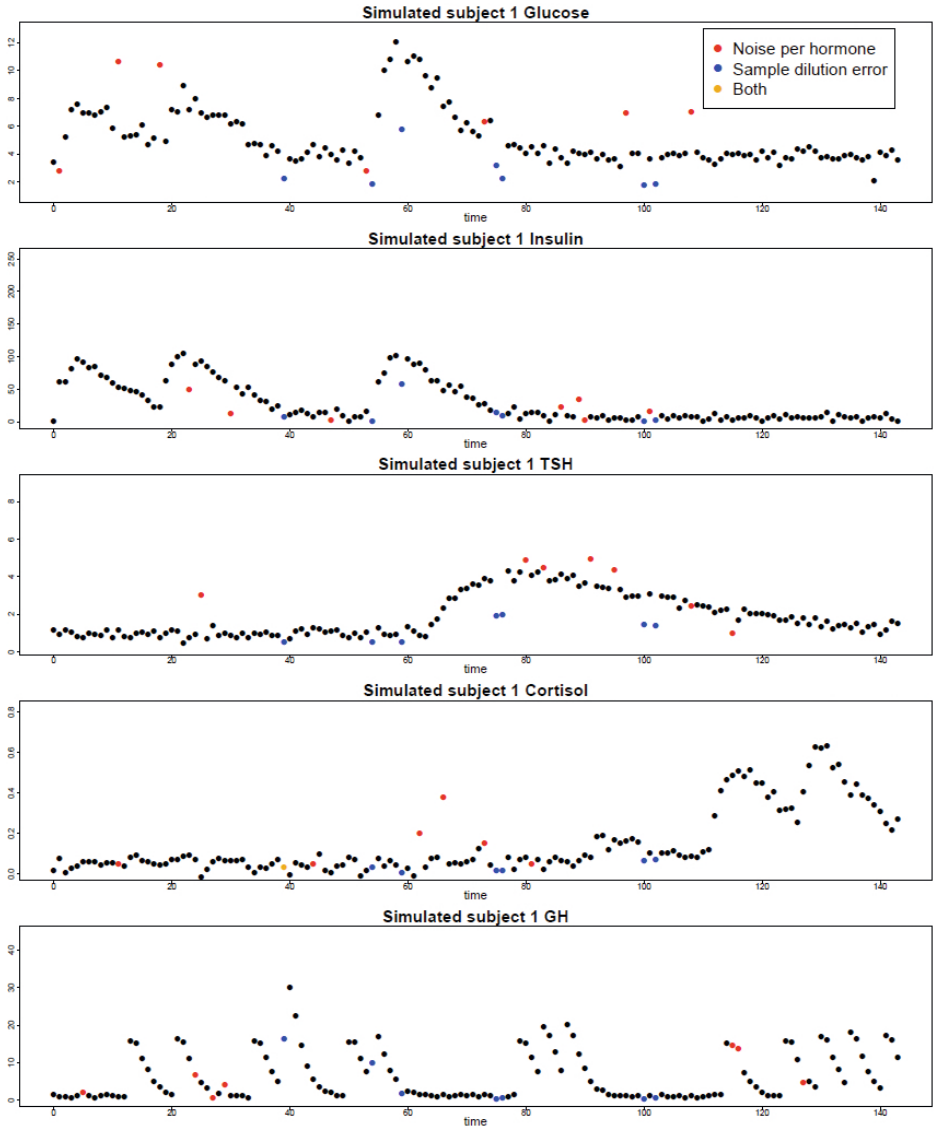


Figure A1. Simulated 24-hour glucose, insulin, TSH, cortisol, and GH data of the first two generated individuals

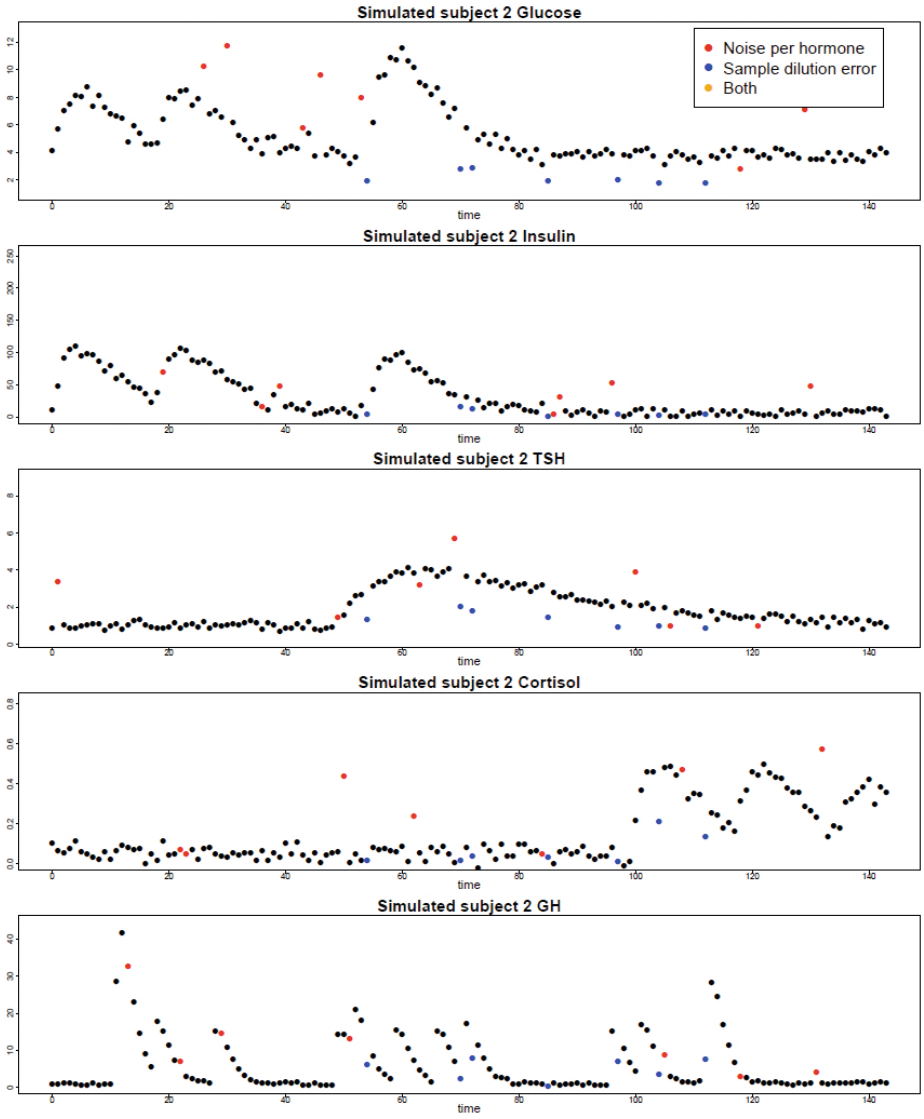


Figure A1 (cont'd)

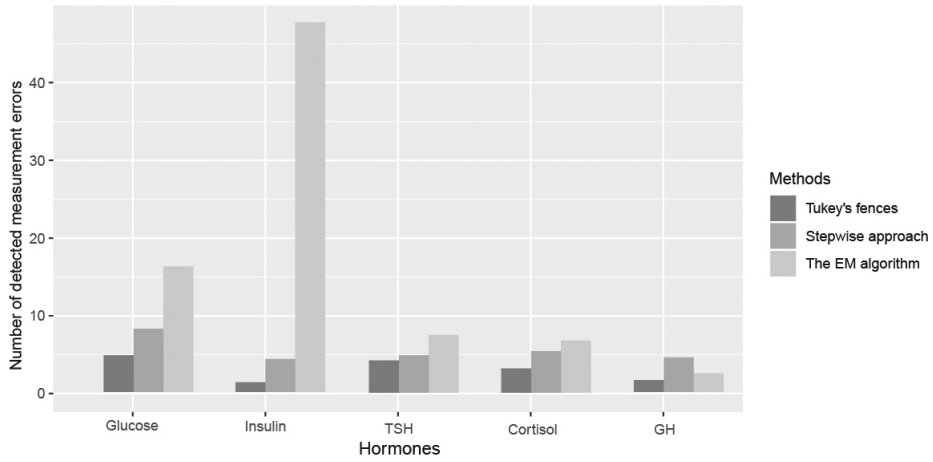


Figure A2. Simulated 24-hour glucose, insulin, TSH, cortisol, and GH data of the first two generated individuals

Table A2. Percentage of true errors detected and true measurement wrongly indicated as an error by each method stratified by random error and dilution error.

		Random error		Dilution error	
		True errors detected (%)	True measurements wrongly indicated as error (%)	True errors detected (%)	True measurements wrongly indicated as error (%)
Stepwise approach	Glucose	22.18	4.80	92.86	1.19
	Insulin	4.51	2.86	49.25	0.58
	TSH	18.42	2.59	53.76	0.79
	Cortisol	24.06	2.67	49.62	1.36
	GH	8.27	2.82	49.25	0.73
	mean	15.49	3.15	58.95	0.93
Tukey's fences	Glucose	34.96	1.67	29.32	1.96
	Insulin	7.89	0.50	4.14	0.69
	TSH	31.58	1.42	27.82	1.61
	Cortisol	31.95	0.65	7.52	1.90
	GH	7.89	0.71	2.63	0.98
	mean	22.86	0.99	14.29	1.43

Table A2. Percentage of true errors detected and true measurement wrongly indicated as an error by each method stratified by random error and dilution error. *(continued)*

		Random error		Dilution error	
		True errors detected (%)	True measurements wrongly indicated as error (%)	True errors detected (%)	True measurements wrongly indicated as error (%)
The EM algorithm	Glucose	60.53	8.70	90.98	7.15
	Insulin	74.81	30.89	77.82	30.73
	TSH	49.25	2.92	46.24	3.07
	Cortisol	43.98	2.65	18.80	3.94
	GH	8.65	1.31	4.14	1.54
	mean	47.44	9.29	47.59	9.29