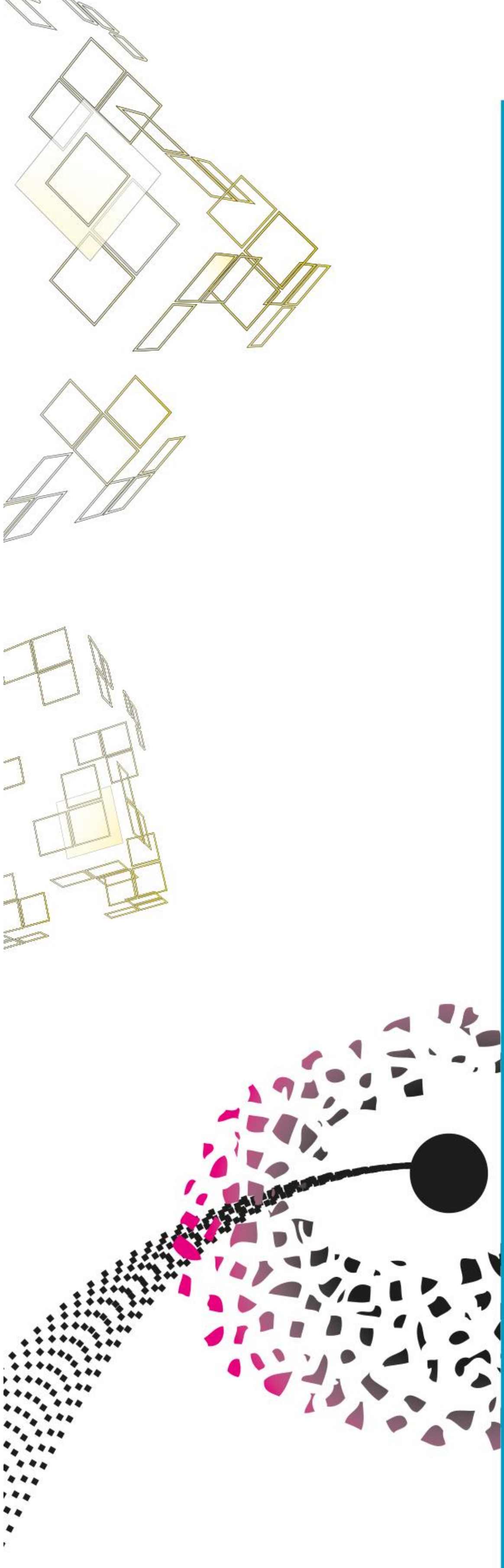


VALIDATION OF VITAL SIGN MONITORING DEVICES

MSC THESIS

12-05-2023

P. van ’t Oosster, BSc

Graduate Intern Technical Medicine
Track Medical Sensing and Simulation
University of Twente
University Medical Centre Utrecht
Division Vital Functions
Department of Anaesthesiology

Supervisors

Chair:	prof. dr. D.W. Donker
Medical supervisor:	dr. T.H. Kappen
Technical supervisor:	dr. Y. Wang
Daily supervisor:	dr. M.J.M. Breteler
Process supervisor:	drs. R.M. Krol
External member:	R.S.P. Warnaar, MSc

UMC Utrecht

UNIVERSITY OF TWENTE.

TABLE OF CONTENT

ABBREVIATIONS

AI	Artificial Intelligence
ANOVA	Analysis of variance
BP	Blood pressure
Cov	Covariance
CPC	Checkpoint Cardio
DBP	Diastolic blood pressure
ECG	Electrocardiogram
EWS	Early warning score
HR	Heart rate
IEEE	Institute of Electrical and Electronics Engineers
LoA	Limits of agreement
MAD	Mead absolute difference
MAP	Mean arterial pressure
MDD	Medical Device Directive
MDR	Medical Device Regulation
MS	Mean square
NIBP	Non-invasive blood pressure
PPG	Photoplethysmography
PTT	Pulse transit time
RR	Respiratory rate
RRT	Rapid response team
S2	Second heart tone
SBP	Systolic blood pressure
SpO2	Oxygen saturation
SD	Standard deviation
UMC Utrecht	University Medical Centre Utrecht

ABSTRACT

Introduction

The vital sign devices market is flooded with devices that are not adequately validated, leading to potential inaccuracies in the readings and posing a significant risk to patient safety. The Limits of Agreement (LoA) analysis is the preferred methodology but can be complicated for manufacturers and researchers, as statistical expertise and programming skills are required.

Methods

We developed an open-source *ValidSense* toolbox, with correct statistical methods to assess the agreement between two devices using a Python package supplemented with a user-friendly graphical user interface. In addition, we developed a longitudinal analysis to assess the agreement over time. Moreover, we performed a validation study of a wearable continuous cuff-based BP device using the *IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2].

Results

The toolbox includes four existing LoA analyses to allow for the correction of multiple measurements per subject (clustering) or non-constant agreement over the measurement range. These four LoA analyses are correctly implemented in the toolbox. A simulation study showed that a newly developed longitudinal analysis allows for detecting non-constant agreement over time (such as a sensor or patient drift). Validation of the BP device fails the IEEE standard for the SBP measurements but passes for the DBP measurements. LoA analysis revealed a bias (95% LoA) of 0.7 (-4.8 to 6.2) mmHg for static SBP measurements and 2.8 (-10.2 to 15.8) mmHg for induced SBP measurements.

Discussion

The ValidSense toolbox guides the user through the four LoA analyses and newly developed longitudinal analysis. The toolbox is easily accessible and allows for reliable LoA analysis without requiring high-level statical knowledge of programming skills. Further research is needed to improve the longitudinal analysis and show the benefits of the longitudinal analysis in a real-world setting. The validation study of the BP device fails the IEEE standard for the SBP measurements but passes for the DBP measurements. Further improvement of the SBP algorithm is needed for reliable measurements in clinical usage.

1 INTRODUCTION

Vital signs are essential indicators of physiological decline and often precede adverse events in hospital wards [3,4]. Late recognition of patient deterioration is associated with several side effects, such as (I) unplanned admission to the intensive care unit (ICU), (II) avoidable cardiopulmonary arrests, (III) increased morbidity and mortality, (IV) extended length of stays in the hospital, and (V) increased hospitalisation costs [5–8]. Accurate measurements are crucial for ensuring the patient’s timely and correct diagnosis and treatment [9].

However, the market is flooded with vital signs devices that are not adequately validated, leading to potential inaccuracies in the readings and posing a significant risk to patient safety [10–12]. Several reasons can be mentioned for the inadequate validation, such as (I) limited regulatory oversight, (II) implementation issues with the current regulation and (III) methodological problems in the analysis, as further discussed in section 2.2.3. The authors believe the best current statistical method to assess the agreement between two quantitative methods or devices measuring the same quantity is the limits of agreement (LoA) analysis, initially introduced in 1986 by Bland and Altman [13]. The authors opt to use the term LoA analysis rather than Bland-Altman analysis, as other methodologies developed by different researchers [14,15] also exist. The LoA analysis's strength is its ability to assess the accuracy and precision of two devices. The Bland-Altman plot [13] visualises the difference and mean between paired measurements, along with lines indicating precision and accuracy (bias and 95% LoA lines).

However, for manufacturers and researchers, the LoA analysis can be complicated, as statistical expertise and programming skills are required. There are different statistical methods to perform the LoA analysis, such as adjusting for multiple measurements in one subject or for the relationship between the difference and mean of the paired measurements. The preferred method is not always obvious [10]. Additionally, of the vital sign devices currently available, it is unknown if they remain accurate over time, with the potential risk of misjudgement of a patient’s condition. Currently, there is no toolbox or library accessible in Matlab, R or Python that can easily perform the LoA analysis, perform multiple variants of the LoA analysis, or can assess the accuracy over time.

This thesis aims to develop an open-source toolbox with the correct statistical methods to assess the agreement between two devices *first time right*. We developed an open-source Python package named *ValidSense.py* that enables the validation of sensors. We also built a user-friendly graphical user interface (GUI) that users can utilise to perform LoA analysis, which may enhance the quality of vital sign monitoring devices. Moreover, we developed a new solution to assess the agreement over time through a longitudinal analysis. Additionally, we validated blood pressure (BP) measurements in a wireless continuous monitoring device.

Validating a vital sign monitoring device: An imaginary company's case

VitalWatch Technologies, a fictional startup, has created a new vital sign monitoring device that can be worn as a watch. The device can monitor multiple vital signs using advanced sensors and algorithms. However, the company has no resources to set up a comprehensive validation study and lacks statistical expertise. Therefore, *VitalWatch Technologies* can not demonstrate the validity of its new device.

The CEO of *VitalWatch Technologies* is aware of the emerging, new *ValidSense* toolbox. The toolbox is exactly what the company needs to assess the reliability of their new vital signs monitoring device the first time right, without the need for thorough statistical expertise or programming skills. The company can now be guided through several steps in the graphical user interface. In addition, all the statistical assumptions are listed and easily tested. In less than fifteen minutes, the CEO performed the analysis on his own. He has the opportunity to show Bland-Altman plots to visualise the accuracy and precision of the device.

The toolbox even offers a solution to assess agreement over time, enabling the manufacturer to identify the until-then-unknown sensor drift. Based on this knowledge, the CEO decided to investigate the root cause of the sensor drift to improve their product. He wants the best device to enhance the monitoring of patients. The *ValidSense* toolbox now allows *VitalWatch Technologies* to validate their device and initiate further improvements. The company can demonstrate the validity of their device to hospitals using the *ValidSense* toolbox.

2 BACKGROUND

In the introduction, the authors stated that verifying vital sign devices is inadequate. The reasons for the inadequate validation of vital sign devices are outlined in section 2.2. Before these reasons are discussed, the rationale for using the LoA analysis as the preferred method is explained in section 2.1. Moreover, the potential of a longitudinal analysis is discussed in section 2.3, and the variants of the LoA analysis are explained in section 2.4.

2.1 THE RATIONALE FOR USING THE LIMITS OF AGREEMENT ANALYSIS AS THE PREFERRED METHOD

In medical applications, demonstrating the validity of a new measurement device for quantifying variables is essential to establish its reliability and reproducibility [16]. The LoA analysis, introduced by Bland and Altman in 1986, known as the classic LoA analysis [13], is an old but still widely used statistical method for assessing the agreement between two measurement methods [16,17]. In medical research, it is often used to compare the reliability of a new device to a reference device [17]. In a Bland-Altman plot [13], each data point represents a pair of measurements, with the horizontal axis representing the average of the two measurements and the vertical axis representing the difference between the two measurements. The Bland-Altman plot also includes a line indicating the bias (accuracy) between the two measurements and lines indicating the upper and lower LoA (precision), which define the range within which 95% of the differences between the two measurements are expected to fall. Accuracy refers to the proximity of measurements to the actual value, while precision represents the variability in repeated measurements [18]. Figure 1 shows the relationship between the accuracy and bias and the precision and 95% LoA. The LoA analysis computes the agreement intervals but does not evaluate the acceptability of these boundaries, which should be determined based on clinical considerations [17]. If the two devices show sufficient agreement, they can be used interchangeably [17].

The authors believe the LoA analysis is the best current available statistical method for assessing the agreement between two quantitative methods or devices. Correlation and regression studies are frequently proposed [19–23]. However, correlation examines the magnitude and significance of the relationship between two variables, and regression predicts the best relationship between two variables by quantifying the goodness of fit [17]. These two methods assess the relationship's strength, not the agreement's quantification. A high correlation does not automatically imply a good agreement between two variables [13,17]. In other words, the correlation and regression methods evaluate the standard error rather than the standard deviation of the variables. To summarise, the appropriate approach to evaluate the agreement between two variables is to consider their differences using the LoA analysis. More information about the LoA analysis can be found in Appendix A.

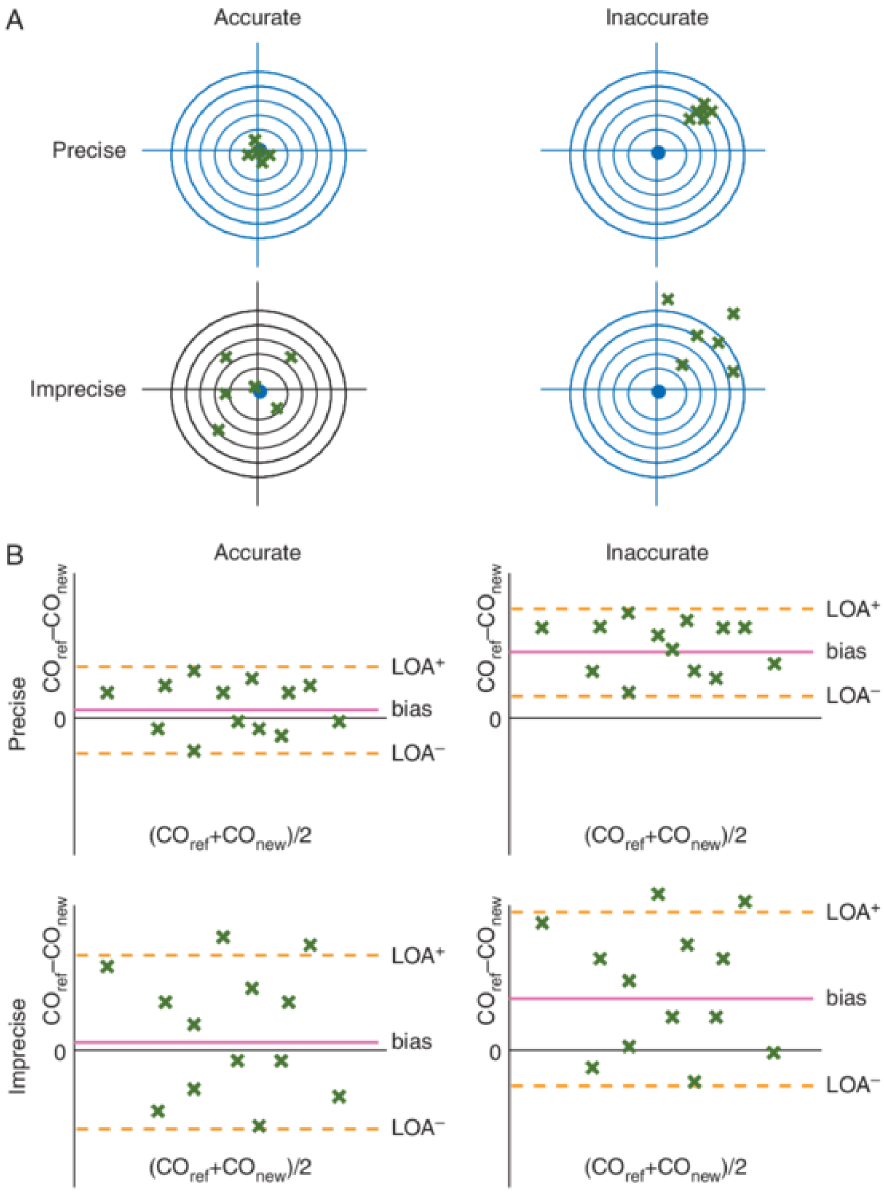


Figure 1. Bias and the limits of agreement representing the accuracy and precision between two devices. A) Accurate measurements are close to the true value, irrespective of the spread of the measurements. In contrast, precise measurements are close to each other, irrespective of their deviation from the true value. B) In Bland-Altman plots, accurate cardiac output monitors show a bias (solid line) close to the zero line. In contrast, precise monitors show limits of agreement close to the bias (dotted lines). Figure with permission derived from [24]. COnew, cardiac output of the experimental technique; COref, reference cardiac output; LOA+, upper limit of agreement; LOA-, lower limit of agreement.

2.2 Reasons for inadequate validation of vital sign devices

Vital signs devices available on the market are often not adequately validated, leading to potential inaccuracies in the readings [10–12]. Inaccurate vital sign devices are widely available for sale and used by clinicians and the general public [10]. Several studies showed that vital signs monitoring systems often lack evidence about the device's accuracy [25–27]. The *Medical Device Assessment Ltd* [28] showed that of the 4100 cuff-based BP measuring devices available on the market, less than 20% had published evidence on the accuracy performance [10]. Clinicians or consumers of these devices are probably unaware of inaccurate or unknown accuracy. This unawareness can have serious implications, as incorrect diagnosis and treatment decisions can be made, and an opportunity is lost to perform the best-practise clinical care. The following sections discuss why the market is flooded with devices that are not adequately validated (summarised in Table 1).

Table 1. Summary of problems and consequences leading to an inadequate validation of vital sign monitoring devices.

Explained in section	Problem	Consequence
2.2.1	Validation testing was not mandatory until May 2021 under the MDR.	Manufacturers of vital sign devices could receive CE certification without proving an acceptable accuracy performance.
2.2.2.1	Validation testing is not mandatory to be performed by independent parties under the new MDR.	Internal testing could lead to potentially questionable company expertise and conflict of interest.
2.2.2.2	The new MDR does not require manufacturers to follow any particular protocol for validating vital sign devices.	Variable protocols can be used to assess and report on the accuracy of vital sign devices.
2.2.2.3	The validation setting does not represent the real-world setting.	Validation outcomes may be overly optimistic.
2.2.2.4	Statistical assumptions in validation studies are violated.	Validation studies present too optimistic results.
2.2.3.1	Multiple measurements in one subject result in clustering.	The within-subject variation in the LoA analysis is neglected.
2.2.3.2	The agreement over the measurement range is non-constant.	The bias and 95% LoA do not represent the accuracy and precision between measurements.
2.2.3.3	The current literature lacks a methodological solution to correct for multiple effects, such as clustering and non-constant agreement over the measurements.	The validity of the LoA analysis is reduced when correction for only one effect takes place and the other effect is ignored.
2.2.3.4	The agreement over time is non-constant.	Methodological solutions of non-constant agreement in the LoA analysis are missing.

MRD, Medical Device Regulation; CE, Conformité Européene; LoA, Limits of Agreement.

2.2.1 Limited regulatory oversight under European law

The European Medicines Agency (EMA) is the regulatory body in the European Union responsible for the safe and effective use of vital sign devices, according to their intended use. The EMA is responsible for setting and enforcing the regulatory protocols and requirements to ensure that medical devices meet the abovementioned objectives and are safe for patients [29].

However, many medical device firms have obtained a CE certification, proving that their device meets European requirements and be used in a clinical setting. Medical device companies could receive CE certification until May 2021 under the European Medical Device Directive (MDD) [30], only needing to demonstrate that the device can measure the intended vital signs. The central focus of the MDD was on patient safety, such as showing that the device is not liable for causing an electric shock rather than the performance of accurate measurements [31]. Manufacturers must show that the device achieves its intended performance. The 'intended purpose' statement in the MDD is ambiguous and could be interpreted as 'how the device is used' [32]. This loophole results in the CE certification not assuring the clinical performance of vital sign monitoring in patients at risk of clinical deterioration [33].

With the new Medical Device Regulation (MDR) [34] taking effect in May 2021, clinical performance needs to be proven, resulting in greater attention to performance than has previously been the case under the MDD [31]. Before CE certification can be granted, a clinical evaluation is required. Article 61 of the MDR states, "*The manufacturer shall specify and justify the level of clinical evidence necessary to demonstrate conformity with the relevant general safety and performance requirements. That level of clinical evidence shall be appropriate in view of the characteristics of the device and its intended purpose.*". Simply showing that a device can safely achieve its intended purpose is inadequate. The manufacturer must also provide evidence of a meaningful and quantifiable benefit to using it. Two requirements in Annex XIV emphasise this: (I) the clinical evaluation plan includes "*a detailed description of intended clinical benefits to patients with relevant and specified clinical outcome parameters*". (II) the clinical evaluation is to "*analyse all relevant clinical data in order to reach conclusions about the safety and clinical performance of the device including its clinical benefits*" [32]. The MDR anticipates prioritising the performance of medical devices by mandating clinical validation before they enter the market.

2.2.2 Problems with implementing the Medical Device Regulation

Although the MDR will prioritise the performance of medical devices, several issues remain regarding implementing the MDR to the development of validation protocols.

2.2.2.1 Validation by independent parties is not mandatory

The MDR does not mandate that independent parties validate vital sign devices. The manufacturer could perform validation testing internally, with the potential of questionable expertise about the correct methodology and conflict of interest [10]. Therefore, it is recommended to set mandatory independent validation of vital sign devices.

2.2.2.2 Variable protocols can be used in validation studies

The MDR mandates no specific protocols to validate vital sign devices can lead to inconsistency in the quality and reliability of validation studies. Without standardised protocols, manufacturers may use different protocols and validation criteria for their validation studies, making it difficult for regulatory authorities to compare and assess the reliability of different devices. The lack of standardised protocols that are mandatory to use can confuse healthcare professionals in evaluating the performance of devices and making informed decisions regarding their use in clinical practice [10,35]. Two examples are provided to indicate the confusion when mandatory protocols are lacking:

Until 2018, no universal protocol for cuff-based non-invasive BP (NIBP) validation existed. Protocols from six different organisations were used for BP validation (*Association for the Advancement of Medical Instrumentatio*n [36], *British Hypertension Society* [37,38], *German Hypertension League* [39], *European Society of Hypertension* [40], *European Committee for Standardization* [41], and *International Organisation for Standardisation* [42]). While these validation protocols have similarities in concept, methodological differences include participant selection criteria, validation procedures, and criteria to pass. It was unclear which protocol was preferred when evaluating the reliability of cuff-based NIBP measuring devices [1,10]. Developing a universal protocol (ISO 81060-2:2018) [42] aimed to reduce confusion about validating vital sign devices. This protocol is expected to be adopted worldwide and provide standardised validation studies.

The ISO protocol (ISO 81060-2:2018) [42] is not applicable for cuffless BP device validation since it only validates intermittent cuff-based BP devices. The AAMI/ESH/ISO agreed that separate validation protocols are needed for cuffless BP monitors [43]. However, consensus on the appropriate protocol for validating cuffless NIBP devices is missing [10]. To our best knowledge, only the Institute of Electrical and Electronics Engineers (IEEE) developed a protocol for cuffless BP measurements (*IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2]), but this protocol is not universally accepted. Of the limited validation studies of cuffless NIBP devices, several studies use inappropriate protocols [44–47], confusing healthcare professionals in evaluating the performance of devices and making informed decisions regarding their use in clinical practice [10,35].

The two examples show the need for universal protocols to validate vital sign devices. We recommend that regulatory bodies establish mandatory requirements about the validation protocol that must be used. The protocol of choice should be universally accepted by researchers and healthcare professionals.

2.2.2.3 The validation setting does not represent the real-world setting

Validation protocols for vital sign devices may not reflect the real-world setting. Conditions can be carefully controlled in a validation setting to ensure that the equipment is operating correctly and the subjects are often healthy volunteers. However, vital signs measurements may be impacting the accuracy by various factors such as (I) underlying medical conditions, (II) medications that affect vital sign measurements, (III) patient distress in real-world situations, or (IV) movement during daily activities. For example, testing the accuracy while the patient is moving is crucial, as it is known that movements can impact the signal accuracy and validity of measurements [48–50], especially in light of the rise of wearable continuous monitoring of vital signs (introduced in section 2.2.3). The results may be more favourable in the current validation protocols than in real-world scenarios, where factors like patient movement can significantly impact signal accuracy. Therefore, it is recommended that the validation setting represents the real-world setting instead of only testing the performance under ideal conditions. Evaluations should include common daily activities, such as getting out of bed, walking, cycling, or climbing stairs, to represent the real-world setting [33].

2.2.2.4 Statistical assumptions in validation studies are violated

Ensuring that statistical assumptions are met is crucial in establishing the validity of the analysis. Nonetheless, Taffe et al. [51] reported that these assumptions are often disregarded in practice, such as (I) failing to account for a systematic relationship between the difference and mean of paired measurements and (II) neglecting to adjust for clustering when dealing with multiple measurements per patient [20,21,44–47]. Violating the LoA assumptions (explained in Appendix A) results in overly optimistic outcomes of accuracy or precision [13,14]. In order to ensure the validity of the analysis, it is essential to understand and comply with the statistical assumptions. Hence, we recommend using the toolbox presented in this thesis to avoid such mistakes.

2.2.3 Methodological problems in limits of agreement analysis

The LoA analysis is the preferred method for assessing agreement between two devices [16,17]. We discuss four methodological challenges with the classic LoA analysis [13], namely (I) clustering, (II) non-constant agreement over the measurement range, (III) combination of multiple effects, and (IV) non-constant agreement over time. First, we introduce continuous vital sign monitoring and medical service centres related to these challenges.

Technological innovations have resulted in lightweight wearable continuous monitoring devices capable of measuring vital signs. Continuous monitoring of vital signs with wearables may allow for a more comprehensive view of a person's health status, improving the quality of care [7,52–54]. Earlier detection of deterioration in patients allows for intervention before the patient's condition worsens, compared to the current practice of intermittent monitoring. The critical difference between continuous and intermittent vital sign monitoring is the frequency and duration of measurements. Continuous monitoring provides a real-time continuous stream of data, while intermittent monitoring provides periodic snapshots of the patient's physiological status. A patient can deteriorate unnoticed as the vital signs are typically manually registered only once every 8-hour shift at the hospital ward [55–57]. The first results with a wearable continuous monitoring system in the hospital ward showed a 1/3 reduction in unplanned ICU admission and rapid response teams calls [5].

The University Medical Centre Utrecht (UMC Utrecht) has established a new medical service centre called the *medisch regie centrum* (MRC) to monitor patients remotely using non-invasive continuous vital sign monitoring. This department aims to accelerate the development of digital health by utilising monitoring devices that allow patients to be monitored remotely by a team of trained medical students and healthcare professionals. The team can intervene promptly in case of any clinical deterioration.

2.2.3.1 Clustering

The first methodological challenge of the LoA analysis occurs when measurements are clustered, meaning multiple measurements within a subject are recorded. In the case of continuous monitoring, multiple vital sign measurements are frequently taken, which is generally not the case with intermittent vital sign measurements used in current practice. Multiple sequential measurements per subject will result in the measurements no longer being independent, as the current measurement will be correlated with the previous and subsequent measurements. Multiple measurements within a subject violate the independence assumption when using the classic LoA analysis [13]. In other words, only the between-subject variation is considered, and the within-subject variation is neglected, resulting in an underestimate of the 95% LoA in the classic LoA analysis [13,14].

2.2.3.2 Non-constant agreement over the measurement range

The second methodological challenge of the LoA analysis is the non-constant level of agreement across the measurement range. For example, the non-constant level of agreement may be caused by the floor effect in respiratory rate measurements [14]. The respiratory rate is unlikely to fall below a certain threshold, and the variability increases with the mean of the respiratory rate, as seen in Figure 2. When utilising the classic LoA analysis [13], it is assumed that there is a consistent agreement across the measurement range. Violating this assumption results in too wide 95% LoA for low values and too small 95% LoA for high values [13]. Therefore, the bias and 95% LoA are not representing the accuracy and precision between the measurements.

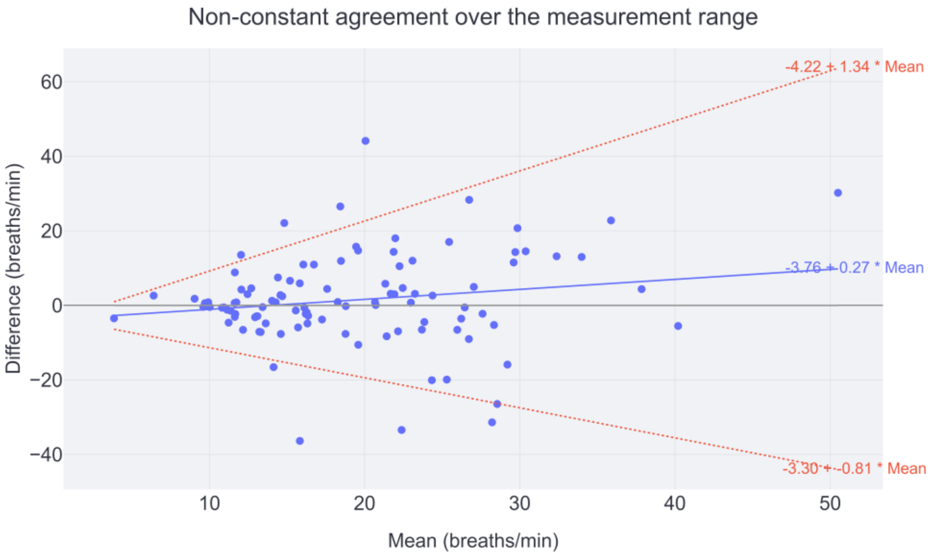


Figure 2. Example of a Bland-Altman plot with a non-constant agreement in respiratory rate measurements. The variability increases with the mean of the respiratory rate. The regression of difference LoA analysis represents the systematic relationship between the difference and mean.

2.2.3.3 Combination of multiple effects

The third methodological challenge of the LoA analysis is that correction for multiple effects is not possible in the current known methodological solutions. Variations of the classic LoA analysis are developed (introduced in section 2.4) but are inadequate to correct for multiple effects, such as clustering and non-agreement over the measurement range. For example, correcting for both effects is currently impossible in the case of respiratory rate measurement with repeated measurements within one patient (clustering) and a systematic relationship between the mean and difference (non-agreement over measurement range). Only one effect can be corrected, while the other is ignored, reducing the validity of the LoA analysis. A methodological solution that can correct for multiple effects is not currently known in the literature. The authors believe extending the mixed-effect LoA analysis may provide the solution to correct for multiple effects, as further discussed in section 4.4.3.4.

2.2.3.4 Non-constant agreement over time

The fourth methodological challenge of the LoA analysis is the non-constant level of agreement over time when there is a drift in the accuracy over time. Methodological solutions are missing to address the problem of non-constant agreement in LoA analysis. Drift refers to the gradual shift in baseline values of the measured physiological parameter over time [58]. In continuous vital sign monitoring, the accuracy could change over time compared to the calibration point, such as (I) sensor drift (e.g. less accurate measurements of the device after the moment of calibration) [59–62] and (II) patient drift (e.g. movement, positioning, health status, or medication) [48–50,63]. Although some potential factors contributing to drift are mentioned in the literature, evidence on the drift in vital sign devices remains limited. Unpublished M2 research of the authors revealed that the administration of norepinephrine increased the average discrepancy between continuous cuff-based NIBP measurements and arterial line measurements. Therefore we may conclude that patients receiving vasoactive medication are more susceptible to inaccurate BP readings. Five other studies [48,59–62] have shown that the agreement over time is non-constant, although the rationale for the non-constant agreement remains unknown in these studies. Drift is a potential issue in continuous monitoring. If drift is not detected and corrected, it can fail to detect changes in the patient's condition or result in false alarms. However, in the LoA analysis, non-constant agreement over time is not considered. In this thesis, we propose a longitudinal analysis that potentially enhances the quality of care.

2.3 THE POTENTIAL OF A LONGITUDINAL ANALYSIS

The detection of drift is essential for at least three target groups, namely (I) manufacturers of vital sign devices, (II) medical service centres utilising vital sign devices, and (III) regulatory bodies setting validation protocols for vital sign devices. Continuous, non-invasive vital sign devices are often used to monitor patient health over extended periods. These devices are typically calibrated to provide accurate measurements, but drift can be a problem for accurate measurements over time. This section discusses the potential of a longitudinal analysis for these three target groups to give insight into accuracy and precision over time.

2.3.1 Manufacturers of vital sign devices

Manufacturers have no insight into the decline of accuracy over time of their devices. By detecting and correcting drift over time, manufacturers can ensure that their devices continue to provide accurate readings and prevent potentially dangerous medical errors. If manufacturers have insight into the issues causing drift, they can target the specific issues contributing to the inaccuracies in vital sign monitoring.

2.3.2 Medical service centres using vital sign devices

Medical service centres do currently not have insight into the drift or other sources of inaccuracy in vital sign readings of the used devices. If medical service centres have to ability to observe these inaccuracies in their devices, they could take action, such as (I) notifying doctors and nurses to be more alert about the inaccurate vital sign reading, (II) initiating calibration of the device, or (III) initiating additional intermittent monitoring of vital signs. Before initiating action, the first step is identifying the patterns and reasons for inaccurate measurements (see examples in section 2.2.2.3). A longitudinal analysis may improve the best clinical practice as medical service centres have insight into the accuracy and precision over time of their used devices.

2.3.3 Regulatory bodies setting validation protocols for vital sign devices

The quality of validation studies could be improved when a longitudinal analysis is incorporated in validation studies. For example, in the *IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2], there is limited testing if the continuous BP device has constant agreement over time. The only requirement in the protocol is that the accuracy before the next calibration is similar to the measurements right after calibration. The factors mentioned in section 2.2.2.3 can cause drift, which is not tested in the IEEE standard. Therefore, the protocol does not guarantee that the agreement remains constant over time. We recommend that regulatory bodies incorporate a longitudinal analysis when setting the validation protocols. This thesis proposes integrating a longitudinal analysis into validation protocols to potentially improve the quality of care.

2.4 VARIANTS OF THE EXISTING LIMITS OF AGREEMENT ANALYSIS

Since the first publication of the LoA analysis [13], several variants of this method have been developed to assess the agreement between two measurement devices for different purposes. We relate the variants to their methodological problems (outlined in section 2.2.3). An elaborate explanation of the techniques can be found in Appendix A.

2.4.1 Clustering

When multiple measurements per subject are analysed in the LoA analysis, the classic LoA analysis [13] becomes inadequate due to the clustering problem (section 2.2.3.1). The classic variant only corrects for variation between the subject cluster, neglecting the within-subject variance. Therefore, the repeated measurements LoA analysis [16,64] was developed by Bland and Altman based on an Analysis Of Variance (ANOVA) model to correct for both the between-subject-variance as well as the within-subject variance.

An alternative to the repeated measurements variant is the mixed effect LoA analysis developed by Parker et al. [14]. This methodology accounts for the clustering of subjects by regarding subjects as a random effect. The total variation is the sum of the within- and between-subject variation. More information and visualisation of mixed-effect modelling can be found in Appendix B.

The mixed effect LoA analysis differs from the repeated measurements variant in that the subjects are seen as random effects. In contrast, in the repeated measurements, subjects are considered as fixed effects in the ANOVA model. If subjects are regarded as fixed effects, they represent the entire population of interest. Conversely, treating them as random effects recognises that they are a subset of a larger population [14].

2.4.2 Non-constant agreement over the measurement range

Suppose the agreement between measurements is not constant across the measurement range, as in the respiratory rate measurements discussed in section 2.2.3.2. In that case, the bias and 95% LoA may not accurately represent the precision and accuracy between the measurements. Bland and Altman developed the regression of difference LoA analysis for this purpose. The regression of difference LoA analysis involves regressing the difference between the measurements against the mean of the measurements. The regression line estimates the bias, and the 95% LoA can be regressed based on the residuals of the bias.

3 OBJECTIVES

The complexity of the LoA analysis causes inadequate validation of vital sign devices due to a lack of statistical expertise and programming skills. Current statistical software tools inadequately address the methodological issues required to apply LoA analysis in many situations. Therefore, we developed an open-source toolbox with correct statistical methods (based on the issues outlined in section) to assess the agreement between two devices using a Python package supplemented with a user-friendly graphical user interface. This toolbox will make it easier to validate the *first time right* [65]. Additionally, we validate a wearable continuous cuff-based BP device using the *IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2].

Main research question

How could we combine a coherent set of methods into a toolbox containing sufficient options to validate quantitative vital sign devices in a wide variety of clinical settings, including continuous monitoring devices?

Subquestion 1 (Chapter 4)

How could we develop an open-source toolbox with appropriate existing statistical methods to validate vital sign devices?

Subquestion 2 (Chapter 4)

How do we expand existing methods to determine the accuracy and precision over time in vital sign devices?

Subquestion 3 (Chapter 4)

How do we design the graphical user interface to make the toolbox easily accessible for end-users?

Subquestion 4 (Chapter 5)

How accurately can a wearable, cuffless, non-invasive continuous monitoring system measure blood pressure using the existing validation protocol?

4 TOOLBOX DEVELOPMENT AND VALIDATION

4.1 INTRODUCTION

Vital sign devices currently available on the market are inadequately validated. We developed the *ValidSense.py* Python package using the existing classic [13], repeated measurements [16,64], mixed-effect [14] and regression of difference [16] LoA analyses (outlined in section 2.4) to correct for the clustering and non-constant agreement over the measurement range. We also developed a user-friendly graphical user interface (GUI) to make the LoA analysis accessible to manufacturers and researchers. Tools for evaluating the statistical assumptions necessary for correctly utilising the LoA analysis and information regarding these assumptions are provided. Moreover, the longitudinal analysis is introduced as a new methodological solution to correct for non-constant agreement over time (outlined in sections 2.2.3.4 and 2.3). We seek to answer the following research questions in this section.

How could we develop an open-source toolbox with appropriate existing statistical methods to validate vital sign devices?

How do we expand existing methods to determine the accuracy and precision over time in vital sign devices?

How do we design the graphical user interface to make the toolbox easily accessible for end-users?

4.2 METHODS

The toolbox development consists of two parts: the development of the *ValidSense.py* package (in section 4.2.1) and a user-friendly GUI (in section 4.2.2). The development of the package is divided into three parts, where section 4.2.1.1 describes the implementation of the four existing LoA analyses, along with the Bland-Altman plot. Section 4.2.1.2 describes the development and implementation of the newly proposed longitudinal analysis, along with the agreement and time series plot. Section 4.2.1.3 outlines the statistical assumptions inspection related to the four LoA analyses, and (graphical) tools are provided to facilitate their verification. The GUI is constructed by utilising the package, with the initial step being the description of the requirements (section 4.2.2.1), followed by an explanation of the GUI workflow (section 4.2.2.2).

In addition, the correct implementation of the four existing LoA analyses was verified (in section 4.2.3). Also, two analyses were conducted, namely (I) the comparison between the four LoA analyses variants (in section 4.2.4) and (II) a simulation study aimed at exploring the benefits of the longitudinal analysis (in section 4.2.5).

4.2.1 Development of the ValidSense package

The Python package *ValidSense.py* (https://github.com/petervtooster/ValidSense) was modularly built to increase reusability. In the future, manufacturers or medical research centres can implement the individual functions of the Python package in their software to facilitate automatic analysis.

Each function in the Python package is designed to be error-proof according to the *fail-first* paradigm, which involves detecting and warning the user at the earliest stage. Errors are induced when (I) data type errors (e.g. argument is string instead of integer), (II) invalid argument error (e.g. argument is a negative integer where a positive integer is expected), (III) missing argument error (e.g. required argument for the function is missing) and (IV) missing data error (e.g. argument contains empty values). Appendix G contains detailed information on the specific warnings for each function and a description of the function, arguments and return values.

Along with the Python package, the permissive MIT license [66] is incorporated to encourage open collaboration and facilitate the sharing and modification of code, allowing for broader participation. The MIT licence also provides some legal protection and disclaims liability for the original developers.

The package is composed of three parts, namely (I) loading, (II) preprocessing, and (III) analysis. The loading part *ValidSense.load* consists of functions to load multiple files and merge them into one *pandas.dataframe* format, according to the following requirements:

1. Only CSV or XLSX files are allowed as extensions for the uploaded files.
2. If multiple files are being loaded, they must have identical variable names.
3. Variables are listed in columns, and the individual paired measurements are in rows. It is required to have variables for two measurement devices being compared and a clustering variable (e.g. subjects).
4. Date and time variables are optional and can be in one or two variables.
5. Files should not contain missing values.

The preprocessing part (*ValidSense.pre*) consists of four functions. First, rename variables that indicate the two devices for consistency in the following functions (*Dev1* and *Dev2*). Second, calculate the difference and mean between these two variables, as required for the LoA analysis. Third, transform the date and time variables to the *Numpy.datetime64* format. We allow date and time to be in one or two variables, allow for modification of the arguments using the *strftime* format, and allow for *UNIX* datetime. Fourth, delete rows that contain missing values.

The analysis part (*ValidSense.analysis*) consists of several functions regarding the existing LoA, the newly developed longitudinal analysis, and statistical assumptions checks. These are explained in the subsequent sections.

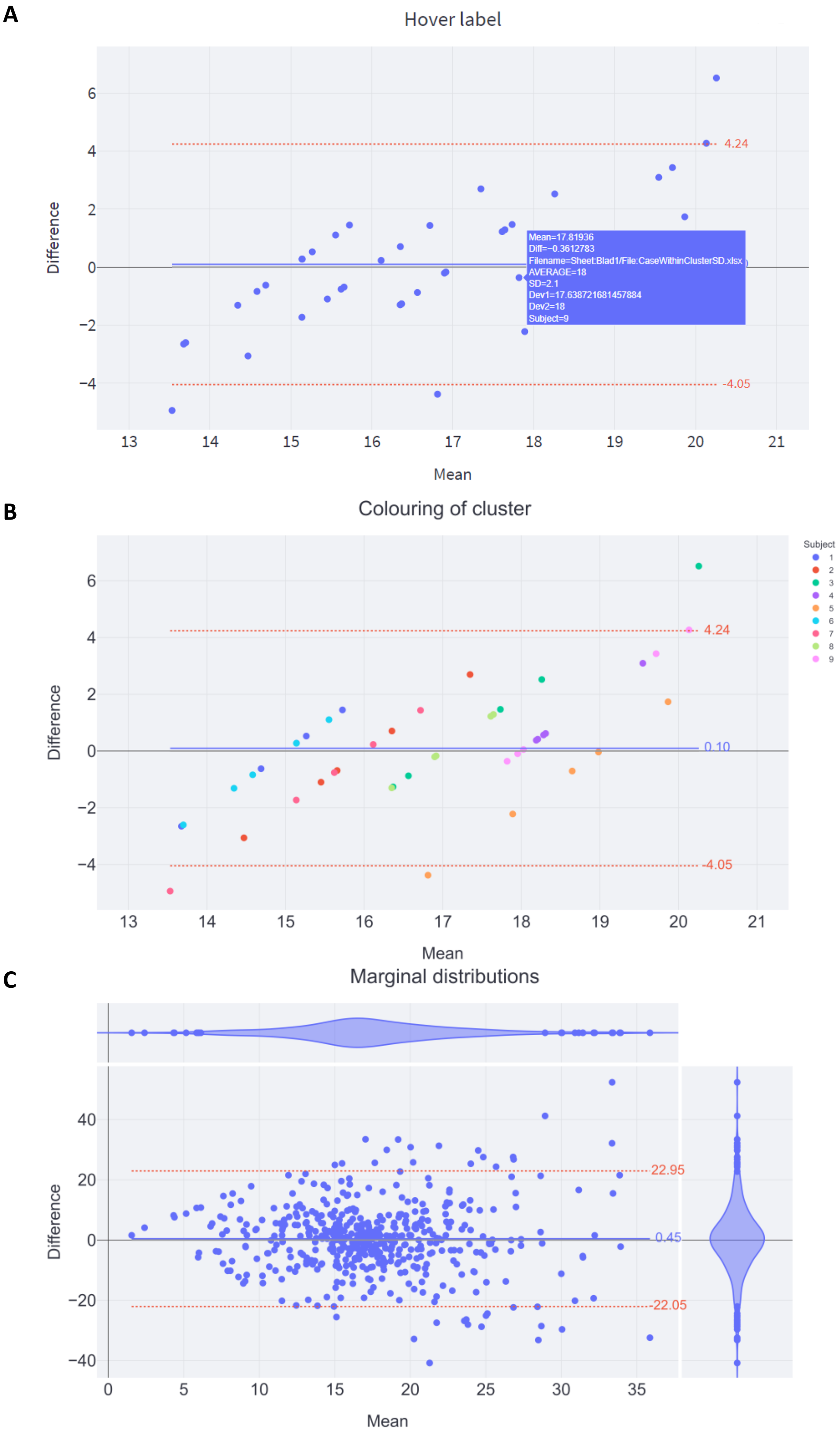
4.2.1.1 Existing limits of agreement analyses and Bland-Altman plot

The four LoA analyses (mentioned in section 2.4) have been implemented in Python, as no functions were available for the LoA analysis in the Python library. The methodology of these articles [13,14,16,64] is not elaborated on in this section but in Appendix A. We cover the additions related to specific requirements and options in the LoA analyses.

For the repeated measurements and mixed-effect LoA analysis, it is necessary to have at least two different subjects in the dataset. Additionally, there must be more measurements than the number of subjects (e.g. ten measurements distributed over ten subjects is insufficient, as within-subject variation cannot be calculated).

In the regression of difference LoA analysis, it is possible to correct for non-constant bias, non-constant LoA, or non-constant bias and LoA over the measurements range. These corrections are achieved by applying linear regression to the mean (detailed information in Appendix A). A first-degree function is expected to be sufficient, according to Bland and Altman [16]. If it turns out that a first-order function is insufficient, a second-order function can be added to obtain a quadratic regression function in future versions of ValidSense.

The outcomes of the four LoA analyses are numerical values for the bias and 95% LoA. The Bland-Altman plot (as explained in section 2.1) is used to visualise accuracy and precision. We have developed a function that generates scatter plots of paired measurements and draws lines for the bias and 95% LoA (based on the four LoA analyses). Moreover, additional functionalities were added (see Figure 3), namely (I) hover labels to provide additional information about specific measurements when the mouse cursor is placed over them to identify outliers, (II) the ability to colour clusters (e.g. subjects) to identify trends in a cluster, (III) marginal distributions of the difference and mean as subplots to enabling pattern recognition, and (IV) heatmap to show the density of measurements in different regions to prevent overplotting in large datasets.



D

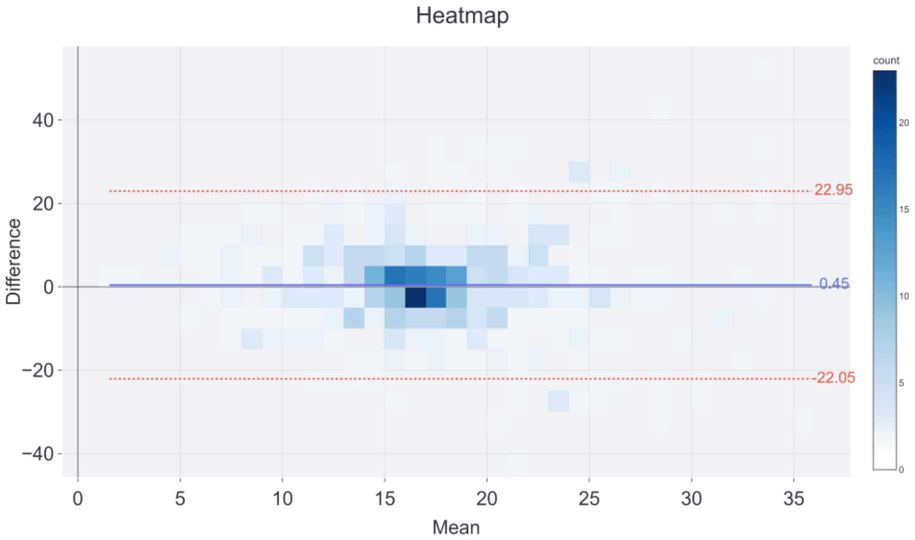


Figure 3. Bland-Altman plots with the additional functionalities incorporated in the ValidSense.py package. A) Hover label, providing additional information about specific measurements when the mouse cursor is placed over them to identify outliers. B) Colouring of a cluster to identify trends in the subject cluster. C) Marginal distribution of the difference and mean to enable pattern recognition. D) Heatmap (in blue) to show the density of measurements in different regions to prevent overplotting in large datasets.

4.2.1.2 Development of the new longitudinal analysis, agreement plot and time series plot

The development of the new longitudinal analysis was necessary to address non-constant agreement over time, as outlined in 2.2.3.4. Existing LoA analyses do not consider changes in accuracy and precision over time, which is why the longitudinal analysis was created.

The longitudinal analysis involves breaking down a dataset into smaller parts over time and applying existing LoA analysis to each part. A moving time window is applied, and based on the data included in the window, the bias and 95% LoA are calculated. The classic [13], repeated measurements [16,64], or mixed-effect [14] LoA analyses are used to calculate the bias and 95% LoA. A constant agreement over the measurement range is assumed.

The agreement plot was developed to visualisation the outcomes of the longitudinal analysis to provide insight into the accuracy and precision over time. Figure 4 provides an example of the agreement plot. The y-axis shows the differences between the two devices (similar to the Bland-Altman plot), while the x-axis represents the start time of the window. The bias- and 95% LoA-lines indicate the accuracy and precision over the time windows. The advantage of the agreement plot is that it facilitates the identification of trends or patterns over time that may go unnoticed otherwise. Exploring the cause of changes is the subsequent step, although this falls beyond the scope of this thesis.

The outcomes of the longitudinal analysis can also be used to visualise the Bland-Altman plot within the selected time window. The combination of both the agreement plot and the Bland-Altman plot of a selected time window could allow for examining over the measurement range and over the time window. In the GUI (section 4.2.2), a time-slider is used to navigate through the agreement plot and simultaneously display the Bland-Altman plot. In order to optimise computational efficiency and reduce processing time for the calculation of the Limits of Agreement (LoA) analysis, a strategic decision was made to decouple the longitudinal analysis from the graphical visualisations, specifically the agreement and Bland-Altman plots.

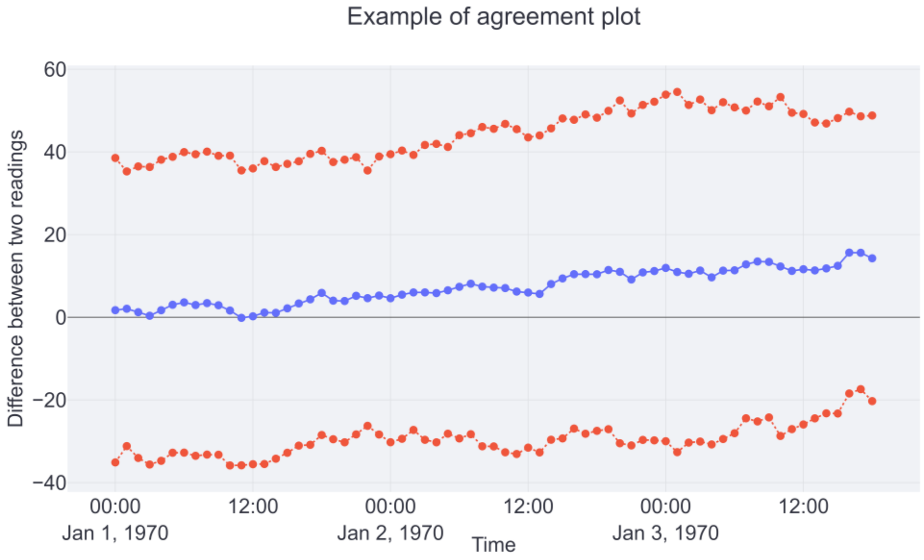


Figure 4. Example of an agreement plot indicating non-constant agreement over time. Bias (blue line) and 95% LoA (red dotted line) within the time window of six hours are presented.

In addition, we have added the time series plot to the *ValidSense.py* package to help identify changes over time (example given in Figure 5). The time series plot scatters the measurements from two devices, with the measurement value on the y-axis and the timestamp on the x-axis. The two devices are distinguished by different symbols, either a circle or a square. Three features are included, namely (I) hover labels to provide additional information about specific measurements when the mouse cursor is placed over them to identify outliers, (II) the ability to colour clusters (e.g. subjects) to identify trends in a cluster, and (III) the ability to include a moving median trend line to identify trends to smooth out the high variation between sequential measurements.

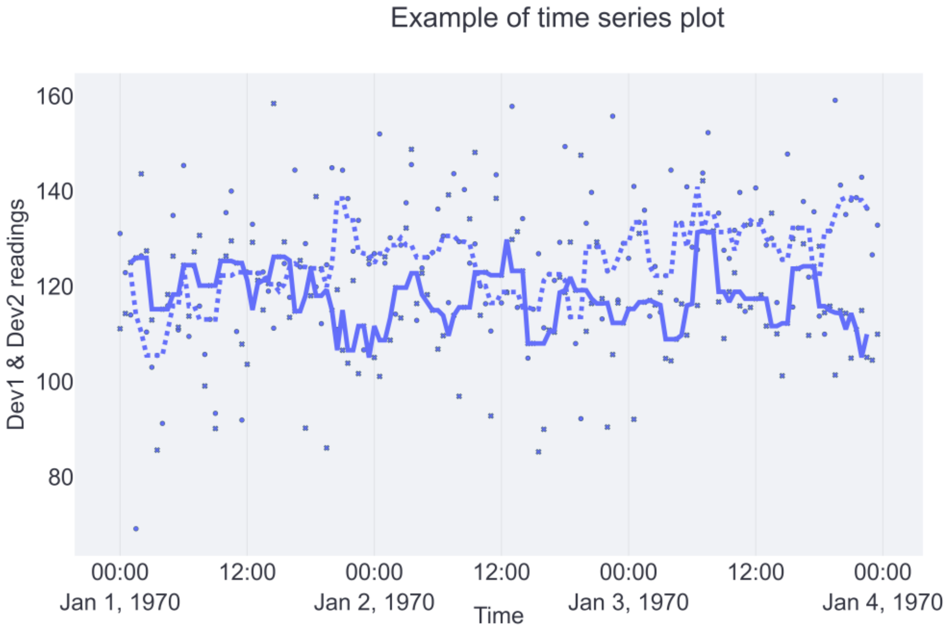


Figure 5. Example of time series plot of one subject with readings of the two devices (rounded and squared dots) and trendlines (median moving window of 30 measurements). Dev1 and Dev2 indicate the two measurement devices.

4.2.1.3 Inspect statistical assumptions in the limits of agreement analysis

Performing the LoA analysis can be complex as statistical expertise and programming skills are required. Making the analysis more accessible is promoted by the package, especially when combined with a GUI. However, another part of the problem is the lacking statistical knowledge, such as ignorance and violation of the statistical assumptions mentioned in section 2.2.2.4. Complying with these assumptions is important for the validity of the LoA analysis. The package guides on meeting the LoA analysis's assumptions and includes six built-in tools (histogram, Q-Q plot, scatter plot, within-cluster-SD plot, residual plot, and covariance) to verify these assumptions. Appendix A provides further information on the six tools and an explanation of the assumptions. The three assumptions that always must be checked are mentioned: First, the normal distribution of the differences can be checked using the histogram and Q-Q plot to ensure that 95% of paired measurements fall within the 95% LoA interval. Second, constant agreement over the measurement range to ensure that the bias and 95% LoA represent accuracy and precision. Third, independent measurements (e.g. violated in case of clustering) to ensure that the 95% LoA represent the precision between measurements. The authors want to emphasise that it is essential for users to check the statistical assumptions to ensure the validity of the LoA analysis.

4.2.2 Graphical user interface

The GUI makes the ValidSense.py Python package easy to use and guides end-users (e.g. researchers). General requirements and the workflow of the GUI are described.

4.2.2.1 General requirements

1. **Multiple pages**: The GUI should contain various pages to organise the functionalities.
2. **Loading and preprocessing**: The GUI should allow for the loading and preprocessing of the data.
3. **LoA analysis**: The GUI should contain the four existing LoA analyses and the Bland-Altman plot (as outlined in section 4.2.1.1.)
4. **Longitudinal analysis**: The GUI should contain a newly developed longitudinal analysis, agreement plot and time series plot (as outlined in section 4.2.1.2).
5. **Statistical assumptions**: The GUI should include information on the statistical assumptions for LoA analysis and the six tools to assess them by the assumptions (as outlined in 4.2.1.3).
6. **Data transportation**: To transfer data should be transferred from one page to another, the GUI must save the data between pages.
7. **Input and visualisation**: The GUI should allow users to input data in the sidebar, whereas the main screen displays visualisations.
8. **Accessibility**: The GUI should be easily accessible via a webpage, initially on the UMC Utrecht Posit Connect server. When the GUI is improved, it should be made open source.
9. **Error handling**: The GUI should notify the user of warnings generated by functions in the ValidSense.py package and any warnings specific to the GUI that may arise when the user provides incorrect or missing inputs.
10. **Hover capability**: The GUI should allow hover capability in graphs to provide the user with additional information about the data.
11. **Download capability**: The GUI should allow for saving high-resolution figures in PNG extension.
12. **Working mechanism**: The GUI should provide a summary of the working mechanism of the GUI.
13. **Contact information**: The GUI should provide contact information.
14. **GitHub link**: The GUI should link to the ValidSense.py package on GitHub.
15. **Licence information**: The GUI should provide information about the MIT licence.

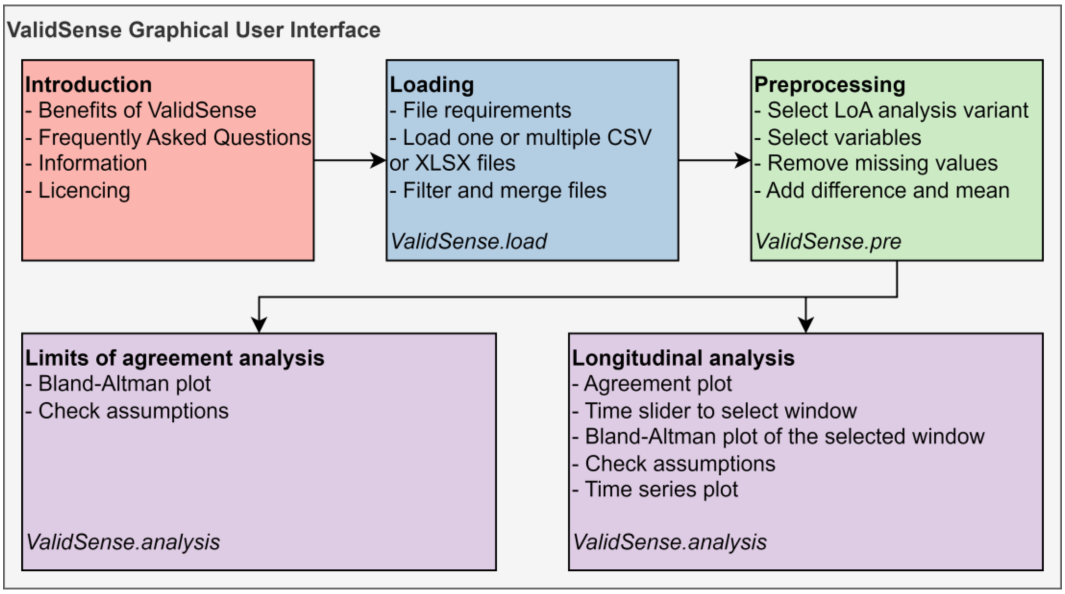


Figure 6. Overview of the ValidSense graphical user interface, built from the ValidSense.py Python package (modules indicated in italics). The multipage toolbox consists of five pages, indicated by the boxes.

4.2.2.2 Workflow of the graphical user interface

Easy accessibility for end-users in the GUI was achieved using Streamlit (https://streamlit.io/). An overview of the GUI is provided in Figure 6. The user interacts via the GUI on five pages described and visualised in Figure 7. The session state of Streamlit is used to store information which persists across multiple pages, enabling the sharing of variables between pages in the ValidSense toolbox. Warning for the several pages are included, as described in Table 2.

Information page: An explanation of the workflow of the toolbox is provided. Information about the variants of the LoA analysis, longitudinal analysis, and statistical assumptions are provided in the form or frequently asked questions. Moreover, the contact information and licence information is provided.

Loading page: The user interacts via the toolbox to (I) access information about the file requirements for loading (as mentioned in section 4.2.1), (II) enables the loading of one or multiple CSV and XLSX files, with the ability to merge multiple files into a single table, and (III) providing the option to customise the delimiter for CSV files.

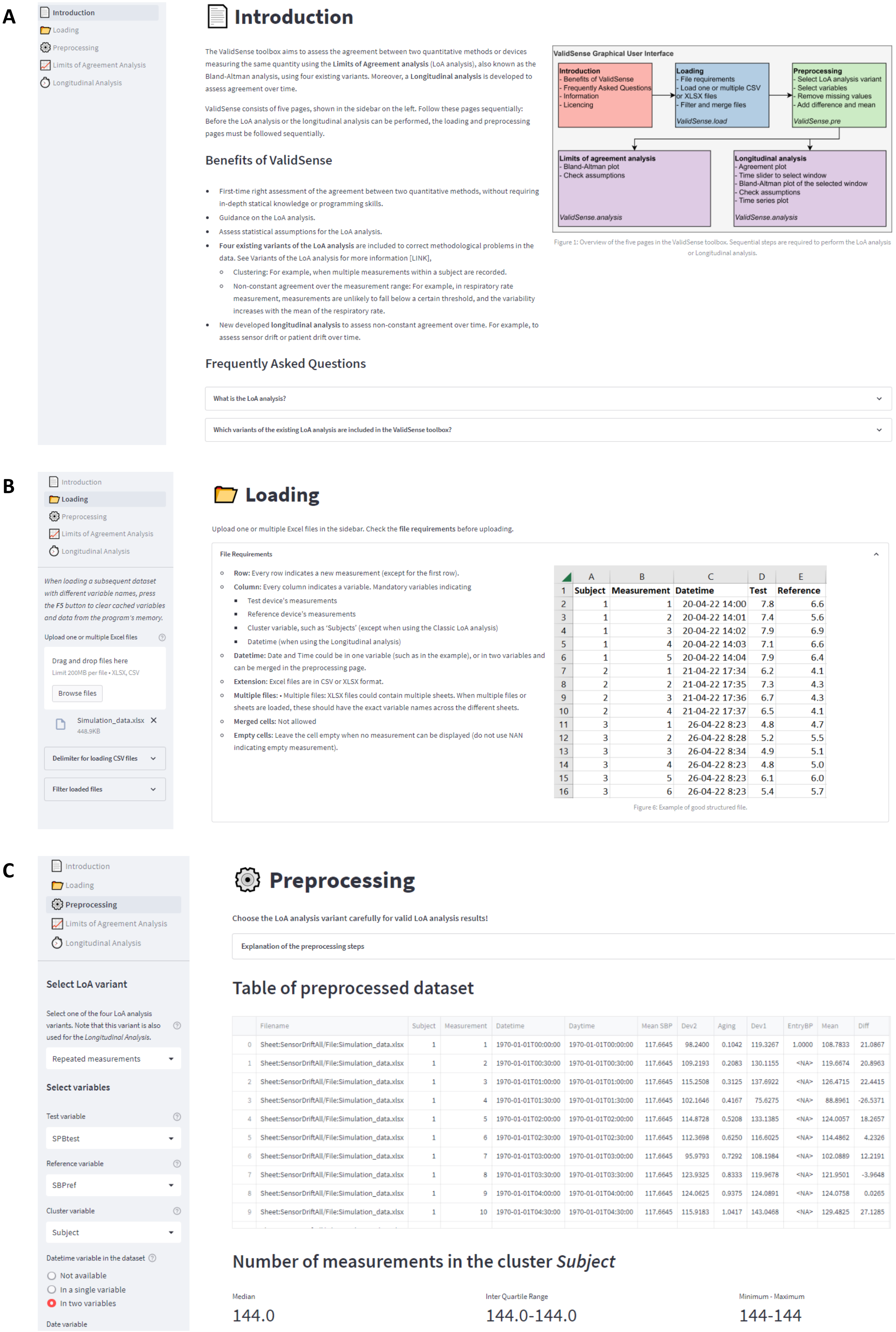
Preprocessing page: The user interacts by selecting one of the four LoA analysis variants used in the LoA Analysis and Longitudinal Analysis page. In addition, the data is standardised by (I) selecting and renaming the two variables indicating the two measurement devices to 'Dev1' and 'Dev2', (II) selecting and renaming a cluster variable (such as 'subjects'), (III) if variables indicating date and time were provided, these were converted to the *numpy.datetime64* format and renamed as ‘datetime', (IV) computing and appending the mean and difference variables for 'Dev1' and 'Dev2' to the table, (V) removing measurements that contain missing values. Moreover, the number of deleted measurements due to missing values and the median, interquartile range, minimum and maximum measurements over the clusters are reported.

Limits of Agreement Analysis page: The LoA analysis outcomes are presented in a table and Bland-Altman plots. The GUI incorporates hover labels, cluster colour-coding, marginal distributions, and heatmaps (as outlined in section 4.2.1.1). The GUI informs the user of the statistical assumptions and offers six tools for assumption assessment (as outlined in section 4.2.1.3 and Appendix A).

Longitudinal Analysis page: The user selects the time window size as input for the longitudinal analysis. The longitudinal analysis outcomes are presented in a table and agreement plot. A time-slider is provided to navigate through the agreement plot and show the Bland-Altman plot of a selected time window. In addition, a time series plot is included, allowing the user to filter the subjects to be included. Information and assessment of the statistical assumptions are provided to the user.

Table 2. Warning for the different pages in the graphical user interface of ValidSense.

Page	Warnings
Loading	• No files have been uploaded. • The Excel file could not be read. • Filtered files or sheets to be merged are empty.
Preprocessing	• Data not loaded. Go back to the loading page. • Select Test and Reference device before continuing. • Select a variable indicating cluster (e.g. 'Subjects'). • Select a datetime variable before continuing. • Select a date and time variable before continuing. • Conversion of Datetime is not possible. • Dev1, Dev2 and cluster variables are not selected in *select variables utilised in the LoA analysis* to remove these measurements.
Limits of Agreement Analysis	• Data not preprocessed. Go back to the preprocessing page. • Select minimally one random variable for bias before continuing. • Select minimally one random variable for 95% LoA before continuing. • Select a variable to visualise distribution in the histogram before continuing. • Select a variable to visualise distribution in the Q-Q plot before continuing. • Select variables indicating fitted values or residuals for the residual plot before continuing. • Select the cluster variable in the within-cluster-SD plot before continuing. • Select the variables for covariance before continuing.
Longitudinal Analysis	• Data not preprocessed. Go back to the preprocessing page. • The *Regression of difference* variant of the LoA analysis is unsuitable for longitudinal analysis, as constant agreement over the measurement range is assumed. • Select minimally one random variable for bias before continuing. • Select minimally one random variable for 95% LoA before continuing. • Window size is not of type integer or is empty. • Window size is not a positive number. • The cluster variable for the longitudinal analysis filtering is empty. • The window size exceeds the time range covered by the data, or the number of the cluster variable included after filtering is too low. • The window size contains all data points. • Select a variable to visualise distribution in the histogram before continuing. • Select a variable to visualise distribution in the Q-Q plot before continuing. • Select variables indicating fitted values or residuals for the residual plot before continuing. • Select the cluster variable in the within-cluster-SD plot before continuing. • Select the variables for covariance before continuing.



Introduction
Loading
Preprocessing
Limits of Agreement Analysis
Longitudinal Analysis

Introduction

The ValidSense toolbox aims to assess the agreement between two quantitative methods or devices measuring the same quantity using the **Limits of Agreement analysis** (LoA analysis), also known as the Bland-Altman analysis, using four existing variants. Moreover, a **Longitudinal analysis** is developed to assess agreement over time.

ValidSense consists of five pages, shown in the sidebar on the left. Follow these pages sequentially: Before the LoA analysis or the longitudinal analysis can be performed, the loading and preprocessing pages must be followed sequentially.

Benefits of ValidSense

- First-time right assessment of the agreement between two quantitative methods, without requiring in-depth statical knowledge or programming skills.
- Guidance on the LoA analysis.
- Assess statistical assumptions for the LoA analysis.
- **Four existing variants of the LoA analysis** are included to correct methodological problems in the data. See Variants of the LoA analysis for more information [LINK],
  - Clustering: For example, when multiple measurements within a subject are recorded.
  - Non-constant agreement over the measurement range: For example, in respiratory rate measurement, measurements are unlikely to fall below a certain threshold, and the variability increases with the mean of the respiratory rate.
- New developed **longitudinal analysis** to assess non-constant agreement over time. For example, to assess sensor drift or patient drift over time.

Frequently Asked Questions

What is the LoA analysis?

Which variants of the existing LoA analysis are included in the ValidSense toolbox?

ValidSense Graphical User Interface

Introduction
- Benefits of ValidSense
- Frequently Asked Questions
- Information
- Licencing

Loading
- File requirements
- Load one or multiple CSV or XLSX files
- Filter and merge files

ValidSense.load

Preprocessing
- Select LoA analysis variant
- Select variables
- Remove missing values
- Add difference and mean

ValidSense.pre

Limits of agreement analysis
- Bland-Altman plot
- Check assumptions

ValidSense.analysis

Longitudinal analysis
- Agreement plot
- Time slider to select window
- Bland-Altman plot of the selected window
- Check assumptions
- Time series plot

ValidSense.analysis

Figure 1: Overview of the five pages in the ValidSense toolbox. Sequential steps are required to perform the LoA analysis or Longitudinal analysis.

Introduction
Loading
Preprocessing
Limits of Agreement Analysis
Longitudinal Analysis

When loading a subsequent dataset with different variable names, press the F5 button to clear cached variables and data from the program's memory.

Upload one or multiple Excel files

Drag and drop files here
Limit 200MB per file • XLSX, CSV

Browse files

Simulation_data.xlsx ✕
448.9KB

Delimiter for loading CSV files

Filter loaded files

Loading

Upload one or multiple Excel files in the sidebar. Check the **file requirements** before uploading.

File Requirements

- **Row:** Every row indicates a new measurement (except for the first row).
- **Column:** Every column indicates a variable. Mandatory variables indicating
  - Test device's measurements
  - Reference device's measurements
  - Cluster variable, such as 'Subjects' (except when using the Classic LoA analysis)
  - Datetime (when using the Longitudinal analysis)
- **Datetime:** Date and Time could be in one variable (such as in the example), or in two variables and can be merged in the preprocessing page.
- **Extension:** Excel files are in CSV or XLSX format.
- **Multiple files:** • Multiple files: XLSX files could contain multiple sheets. When multiple files or sheets are loaded, these should have the exact variable names across the different sheets.
- **Merged cells:** Not allowed
- **Empty cells:** Leave the cell empty when no measurement can be displayed (do not use NAN indicating empty measurement).

	A	B	C	D	E
1	Subject	Measurement	Datetime	Test	Reference
2	1	1	20-04-22 14:00	7.8	6.6
3	1	2	20-04-22 14:01	7.4	5.6
4	1	3	20-04-22 14:02	7.9	6.9
5	1	4	20-04-22 14:03	7.1	6.6
6	1	5	20-04-22 14:04	7.9	6.4
7	2	1	21-04-22 17:34	6.2	4.1
8	2	2	21-04-22 17:35	7.3	4.3
9	2	3	21-04-22 17:36	6.7	4.3
10	2	4	21-04-22 17:37	6.5	4.1
11	3	1	26-04-22 8:23	4.8	4.7
12	3	2	26-04-22 8:28	5.2	5.5
13	3	3	26-04-22 8:34	4.9	5.1
14	3	4	26-04-22 8:23	4.8	5.0
15	3	5	26-04-22 8:23	6.1	6.0
16	3	6	26-04-22 8:23	5.4	5.7

Figure 6: Example of good structured file.

Introduction
Loading
Preprocessing
Limits of Agreement Analysis
Longitudinal Analysis

Select LoA variant

Select one of the four LoA analysis variants. Note that this variant is also used for the *Longitudinal Analysis.*

Repeated measurements

Select variables

Test variable: SPBtest

Reference variable: SBPref

Cluster variable: Subject

Datetime variable in the dataset
○ Not available
○ In a single variable
● In two variables

Date variable

Preprocessing

Choose the LoA analysis variant carefully for valid LoA analysis results!

Explanation of the preprocessing steps

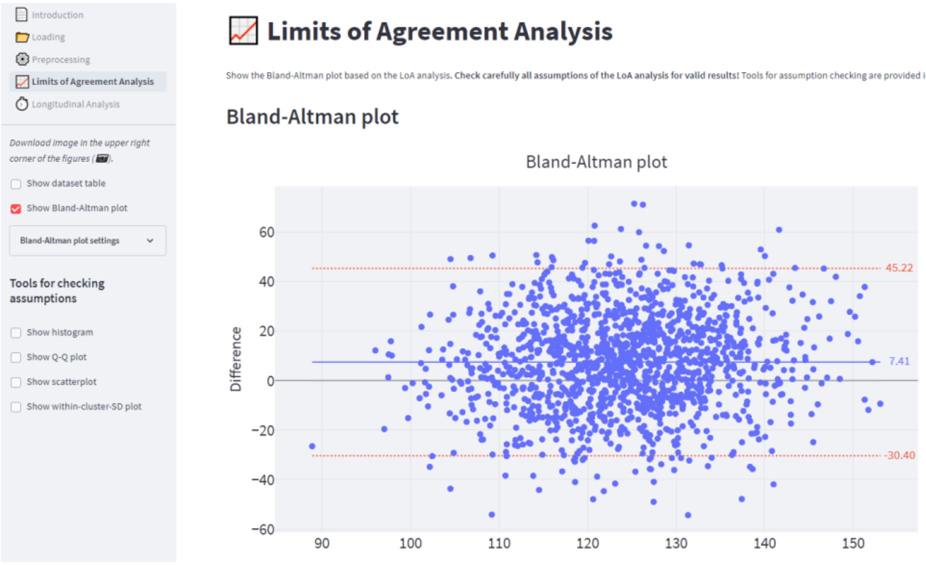
Table of preprocessed dataset

	Filename	Subject	Measurement	Datetime	Daytime	Mean SBP	Dev2	Aging	Dev1	EntryBP	Mean	Diff
0	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	1	1970-01-01T00:00:00	1970-01-01T00:00:00	117.6645	98.2400	0.1042	119.3267	1.0000	108.7833	21.0867
1	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	2	1970-01-01T00:30:00	1970-01-01T00:30:00	117.6645	109.2193	0.2083	130.1155	<NA>	119.6674	20.8963
2	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	3	1970-01-01T01:00:00	1970-01-01T01:00:00	117.6645	115.2508	0.3125	137.6922	<NA>	126.4715	22.4415
3	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	4	1970-01-01T01:30:00	1970-01-01T01:30:00	117.6645	102.1646	0.4167	75.6275	<NA>	88.8961	-26.5371
4	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	5	1970-01-01T02:00:00	1970-01-01T02:00:00	117.6645	114.8728	0.5208	133.1385	<NA>	124.0057	18.2657
5	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	6	1970-01-01T02:30:00	1970-01-01T02:30:00	117.6645	112.3698	0.6250	116.6025	<NA>	114.4862	4.2326
6	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	7	1970-01-01T03:00:00	1970-01-01T03:00:00	117.6645	95.9793	0.7292	108.1984	<NA>	102.0889	12.2191
7	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	8	1970-01-01T03:30:00	1970-01-01T03:30:00	117.6645	123.9325	0.8333	119.9678	<NA>	121.9501	-3.9648
8	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	9	1970-01-01T04:00:00	1970-01-01T04:00:00	117.6645	124.0625	0.9375	124.0891	<NA>	124.0758	0.0265
9	Sheet:SensorDriftAll/File:Simulation_data.xlsx	1	10	1970-01-01T04:30:00	1970-01-01T04:30:00	117.6645	115.9183	1.0417	143.0468	<NA>	129.4825	27.1285

Number of measurements in the cluster *Subject*

Median	Inter Quartile Range	Minimum - Maximum
144.0	144.0-144.0	144-144

D



Download image in the upper right corner of the figures (📷).

□ Show dataset table

☑ Show Bland-Altman plot

Bland-Altman plot settings

Tools for checking assumptions

□ Show histogram

□ Show Q-Q plot

□ Show scatterplot

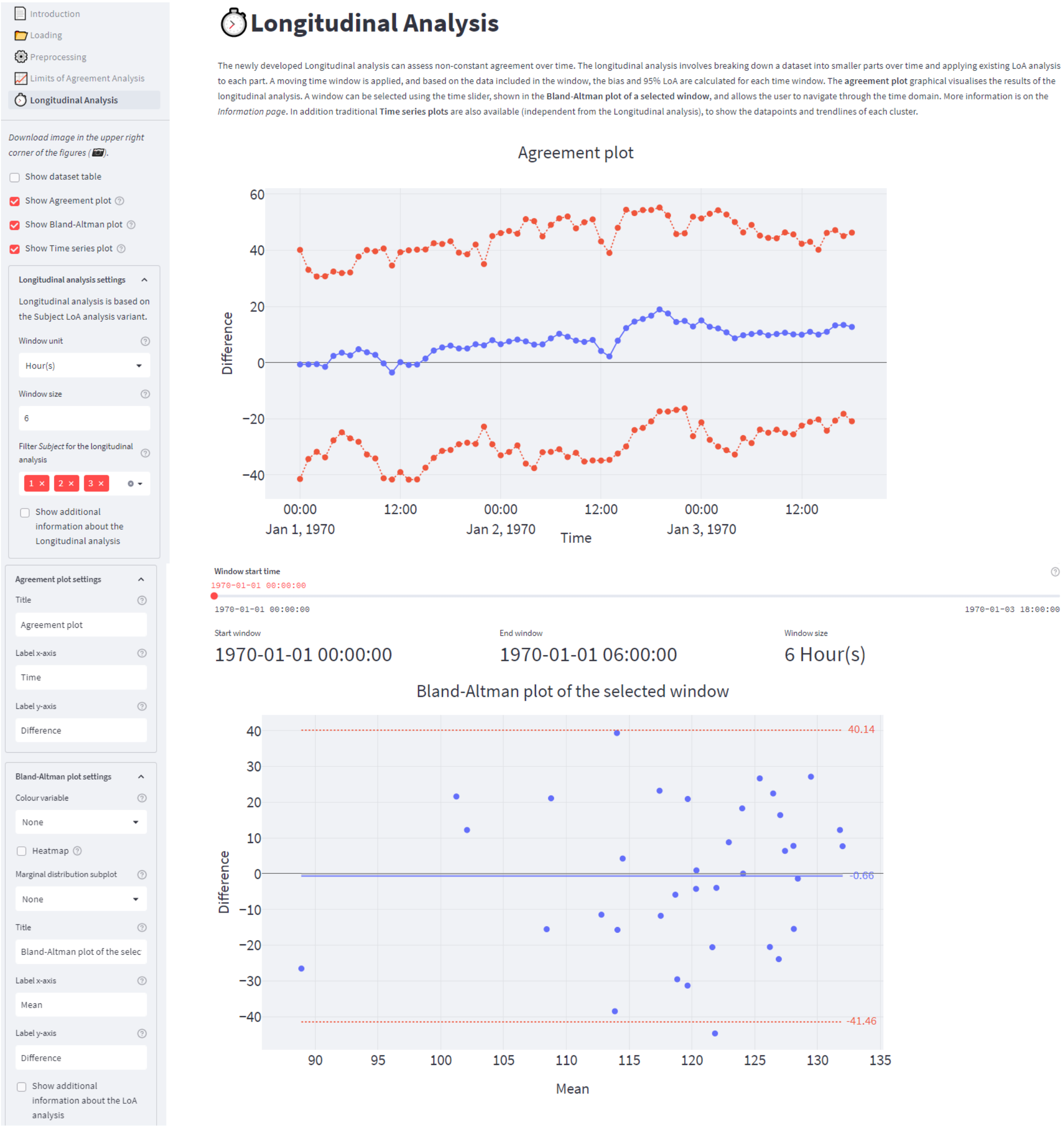
□ Show within-cluster-SD plot

Limits of Agreement Analysis

Show the Bland-Altman plot based on the LoA analysis. **Check carefully all assumptions of the LoA analysis for valid results!** Tools for assumption checking are provided in the sidebar.

Bland-Altman plot

E



Download image in the upper right corner of the figures (📷).

□ Show dataset table

☑ Show Agreement plot

☑ Show Bland-Altman plot

☑ Show Time series plot

Longitudinal analysis settings

Longitudinal analysis is based on the Subject LoA analysis variant.

Window unit: Hour(s)

Window size: 6

Filter *Subject* for the longitudinal analysis: 1 × 2 × 3 ×

□ Show additional information about the Longitudinal analysis

Agreement plot settings

Title: Agreement plot

Label x-axis: Time

Label y-axis: Difference

Bland-Altman plot settings

Colour variable: None

□ Heatmap

Marginal distribution subplot: None

Title: Bland-Altman plot of the selec

Label x-axis: Mean

Label y-axis: Difference

□ Show additional information about the LoA analysis

Longitudinal Analysis

The newly developed Longitudinal analysis can assess non-constant agreement over time. The longitudinal analysis involves breaking down a dataset into smaller parts over time and applying existing LoA analysis to each part. A moving time window is applied, and based on the data included in the window, the bias and 95% LoA are calculated for each time window. The **agreement plot** graphical visualises the results of the longitudinal analysis. A window can be selected using the time slider, shown in the **Bland-Altman plot of a selected window**, and allows the user to navigate through the time domain. More information is on the *Information page*. In addition traditional **Time series plots** are also available (independent from the Longitudinal analysis), to show the datapoints and trendlines of each cluster.

Window start time

1970-01-01 00:00:00

1970-01-01 00:00:00 — 1970-01-03 18:00:00

Start window: 1970-01-01 00:00:00

End window: 1970-01-01 06:00:00

Window size: 6 Hour(s)

Figure 7. Impression of the different pages of the GUI, with A) Introduction page, B) Loading page, C) Preprocessing page, D) Limits of Agreement Analysis page, and E) Longitudinal Analysis page.

4.2.3 Verifying the correct implementation of the four limits of agreement analysis

To verify whether the four existing LoA analyses were correctly implemented in the *ValidSense.py* package, the outcomes of the four existing LoA analyses were compared with those reported in the original articles [11,12,15,63]. We provide a summary of the datasets that were utilised.

Classic LoA analysis: Table 2 from Bland and Altman's 1999 paper [16] was preferred over the one presented in their 1986 paper [13] due to its larger sample size and more precise measurements to allow for accurate verification. The BP measurement of the first measurement of the sphygmomanometer was compared to the first measurement by an observer from this dataset (J1 vs S1).

Repeated measurements LoA analysis: Table 4 in Bland and Altman's 1999 paper [16] describes 60 independent pairs of measurements taken from 12 subjects. Two different methods (IC vs RV) measured the cardiac ejection fraction (%)[64]. Table 4 in Bland and Altman's 1999 paper [16] and the 2007 paper [64] both describe this dataset, but there are inconsistencies between the papers, as pointed out by Matsubayashi [67]. These include incorrect presentation of between- and within-subject variance, total standard deviation, bias, and LoA in the 1999 paper, slight changes in subject numbering between the two papers, and incorrect calculation of between- and within-subject variance in the 2007 paper. To ensure correct verification, we compared the between- and within-subject variance presented by Matsubayashi [67].

Mixed-effect LoA analysis: Parker's dataset [14] compared the respiratory rate measured by the chest band (RRcb) to that of the gold standard respiratory rate monitor (RRox) without removing any outliers. Parker considered the subject as a random effect in both the bias- and 95% LoA-models but considered activity only as a fixed effect in the 95% LoA-model.

Regression of difference LoA analysis: Table 3 from Bland and Altman's paper [16] was utilised, which contains paired measurements of fat content in human milk. The enzymic procedure was used to compare Gerber to the Triglycerides method. It should be noted that Figure 8 of the 1999 paper [16] erroneously states that the difference of paired measurements is calculated as Triglycerides minus Gerber. The model for bias and LoA relies solely on the mean of the paired measurements, resulting in parallel LoA to the bias line in the Bland-Altman plot.

4.2.4 Analysis I: Comparing the four limits of agreement analyses

We performed the four LoA analyses of the *ValidSense.py* package to compare the outcomes of the four LoA analyses. A large dataset was used to compare measurements from a continuous blood pressure monitor with NIBP measurements. We describe the study design, setting, and population (note that the same dataset was used in section 5.2.1 for the validation study, but only the static SBP measurements were used to compare LoA analyses variants).

Volunteers were measured using the CPC device with earclip PPG sensor (Checkpoint Cardio, Kazanluk, Bulgaria) and intermittent using auscultatory NIBP (Microlife WatchBP Office AFIB, Widnau, Switzerland) by two experienced research nurses employed by Checkpoint Cardio. The data collection was performed in the medical research centre of Checkpoint Cardio in Kazanluk, Bulgaria, without specific inclusion or exclusion criteria. The manufacturer aimed to include more than 1600 volunteers to get a representative sample of the Bulgarian population, with volunteers included that (I) are free from haemodynamic problems, (II) diagnosed with hypertension, (III) diagnosed with hypotension. Volunteers were first given 10 minutes to relax on a chair after the CPC and the reference NIBP device were used for the static test. In this test, volunteers sat on a chair, and five measurements with the CPC and NIBP were taken simultaneously at 5-minute intervals between each measurement. The research nurse rounded the auscultatory NIBP measurements to the nearest five mmHg. CPC measurements with similar timestamps to the NIBP measurements were paired based on the nearest timestamp and saved in a CSV file. More information regarding the rationale for auscultatory NIBP measurements and details about the CPC measurement system can be found in Appendix C and D.

The paired measurements were used for the four LoA analyses variants to compare the methodologies based on the outcomes of the LoA analysis. These outcomes are the bias, lower and upper 95% LoA, between- and within-subject-SD, and total-SD.

4.2.5 Analysis II: A simulation study to explore the advantages of longitudinal analysis

We set up a simulation study to investigate the potential benefit of a longitudinal analysis, which has not yet been possible but could be a potential issue for continuous monitoring. Drift could be related to the sensor or the patient, as outlined in section 2.2.3.4. The authors recognise the potential for a longitudinal analysis for (I) manufacturers of vital sign devices, (II) medical service centres using these devices, and (III) regulatory bodies setting validation protocols (outlined in section 2.3).

The simulation study involves a hypothetical example in which a medical service centre wants to compare the accuracy of a new wearable continuous NIBP device to a Holter BP device as a reference, measuring two times per hour [68] over three days in ten subjects. First, we explain and simulate the initial situation after which sensor or patient drift was added to the simulated wearable readings, specified in five scenarios. An overview of these scenarios is visualised in Figure 8.

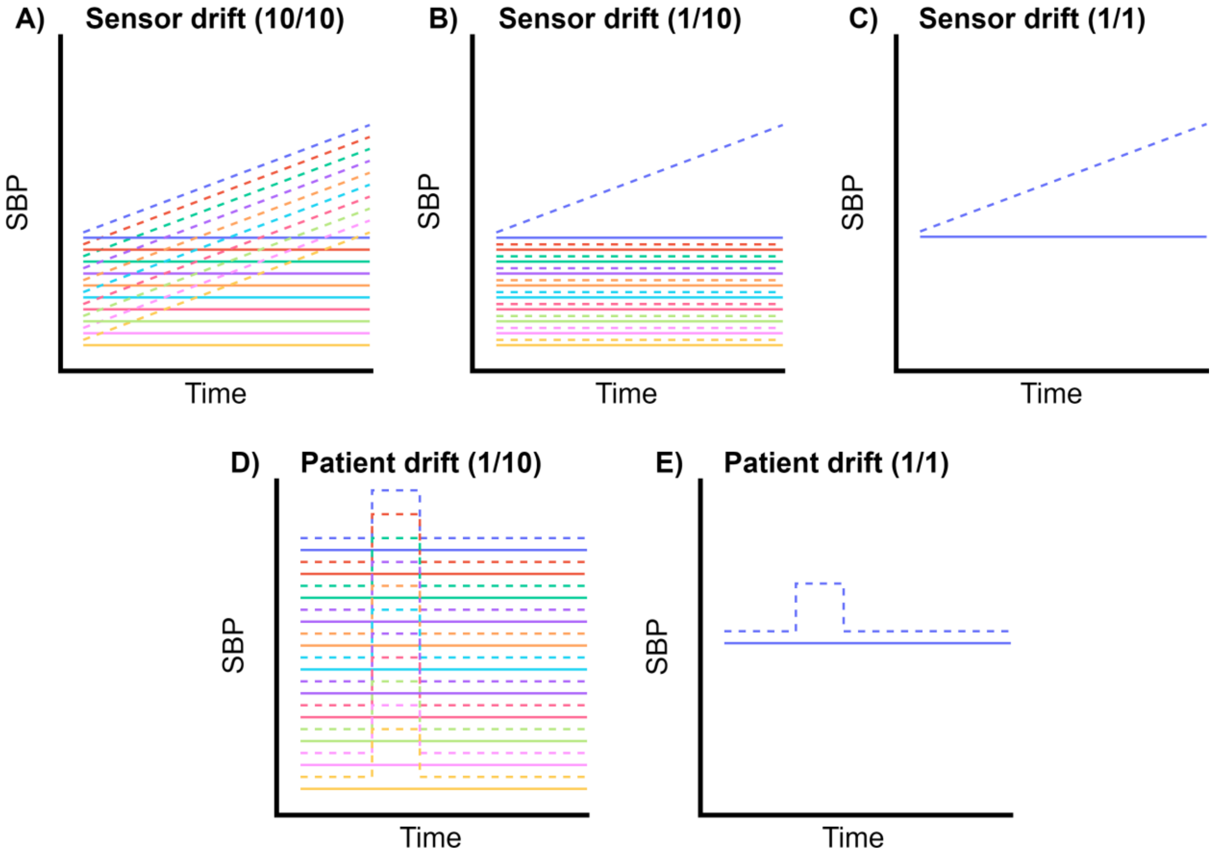


Figure 8. Illustration of the five scenarios of the simulation study. A) Sensor drift in all ten wearables. B) Sensor drift in one out of ten wearables. C) Sensor drift in single wearables. D) Patient drift in one out of ten wearables. E) Patient drift in single wearables.

Initial situation: The simulated SBP was chosen with a mean of 119 mmHg, with a between-subject SD of 3 mmHg [69]. We chose a within-subject SD of 13 mmHg to account for (I) daily activity variations [70] and (II) inaccuracies of the wearable and Holter BP device, as neither can measure true BP [35,71–74]. We used the *NORM.INV* function in Microsoft Excel (Microsoft Corporation, version 2301) to simulate the initial readings based on the mentioned mean, between-subject, and within-subject variations in SBP.

Sensor drift: BP sensors may become less accurate over time after the moment of calibration, which is a well-known phenomenon [59–62]. The sensor drift could be caused by a systematic malfunction in all simulated wearables or a specific malfunction in only one wearable. We assumed that the wearable is subject to a linear decline of 5 mmHg per day in accuracy due to sensor drift.

Patient drift: The administration of vasoactive medication could result in a sudden accuracy decline in BP readings. It is known that vasoactive medication [63] can affect the pulse transit time (PTT) in continuous BP devices (see Appendix D for more information about PTT). As a result, the wearable may provide inaccurate BP readings while the Holter device remains reliable. We assumed in our simulation that the Holter BP reading remains constant even when exposed to norepinephrine. We simulate a single patient receiving a dose of 0.26 µg/kg/min of norepinephrine, resulting in a 20 mmHg increase in the SBP reading [75] during 6 hours (on day one from 12:00 to 18:00).

Scenario A – Sensor drift in all ten wearables: We simulate a linear sensor drift of 5 mmHg per day after the calibration point on Day 1 at 0:00 in all ten wearables. This scenario could represent a systematic malfunction in all wearables.

Scenario B – Sensor drift in one out of ten wearables: We simulate a linear sensor drift of 5 mmHg per day after the calibration point on Day 1 at 0:00 in only one out of ten wearables. This scenario could represent a specific malfunction in one wearable.

Scenario C – Sensor drift in single wearables: We simulate a linear sensor drift of 5 mmHg per day after the calibration point on Day 1 at 0:00 in only one wearable.

Scenario D – Patient drift in one out of ten wearables: We simulate a sudden patient drift of 20 mmHg between Day 1 at 12:00 till 18:00 in one out of ten wearables.

Scenario E – Patient drift in single wearables: We simulate a sudden patient drift of 20 mmHg between Day 1 at 12:00 till 18:00 in only one wearable.

The longitudinal analysis utilised these five scenarios and was depicted through an agreement plot. A time window of six hours was set for the longitudinal analysis.

4.3 Results

4.3.1 Verifying the correct implementation of the four limits of agreement analysis

Correct implementation of the four LoA analyses in the *ValidSense.py* package was verified by comparing the outcomes to the original articles. Table 3 illustrates the clustering of measurements across multiple subjects. Table 4 shows identical results from the four LoA analyses in the ValidSense toolbox compared to those mentioned in the original articles. Bland-Altman plots of these four LoA analyses can be found in Appendix F.

Table 3. Characteristics of the four datasets utilised to verify correct implementation.

LoA analysis	Number of measurements	Number of subjects
Classic	85	85
Repeated measurements	60	12
Mixed-effect	385	21
Regression of difference	45	45

Table 4. Verification of the correct implementation of the four LoA analyses variants.

LoA analysis variant	Bias	Lower 95% LoA	Upper 95% LoA
Classic			
Original article	-16.29	-54.7	22.1
ValidSense toolbox	-16.29	-54.7	22.1
Repeated measurements			
Original article	0.6022	-1.3395	+2.5438
ValidSense toolbox	0.6022	-1.3395	+2.5438
Mixed-effect			
Original article	-1.60	-9.99	6.80
ValidSense toolbox	-1.60	-9.99	6.80
Regression of difference			
Original article	0.079 - 0.0283×M	-0.078 - 0.0283×M	0.236 - 0.0283×M
ValidSense toolbox	0.079 - 0.0283×M	-0.078 - 0.0283×M	0.236 - 0.0283×M

Decimals are rounded, similar to the original articles. M: mean of paired measurements.

4.3.2 Analysis I: Comparing the four variants of the limits of agreement analysis

Volunteers were enrolled from August 2019 to May 2022, resulting in 5854 paired measurements in 1411 subjects (the subject's characteristics can be found in Table 8 in section 5.3, as the same dataset is utilised). The outcomes of the LoA analyses are shown in Table 5 (the Bland-Altman plot of the four LoA analyses is shown in Appendix F). The repeated measurements and mixed-effect LoA analyses correct for the clustering of subjects, with the repeated measurements showing a slightly smaller total variability than the mixed-effect LoA analysis. The repeated measurements LoA analysis showed a higher between-subject-SD but smaller within-subject-SD than the mixed-effect LoA analysis. In addition, the classic LoA analysis shows less variation between measurements since it does not account for clustering, in contrast to the larger variations in the repeated measurements and mixed-effect LoA analysis. Finally, an almost constant agreement over the measurement range for the regression of difference LoA analysis is observed.

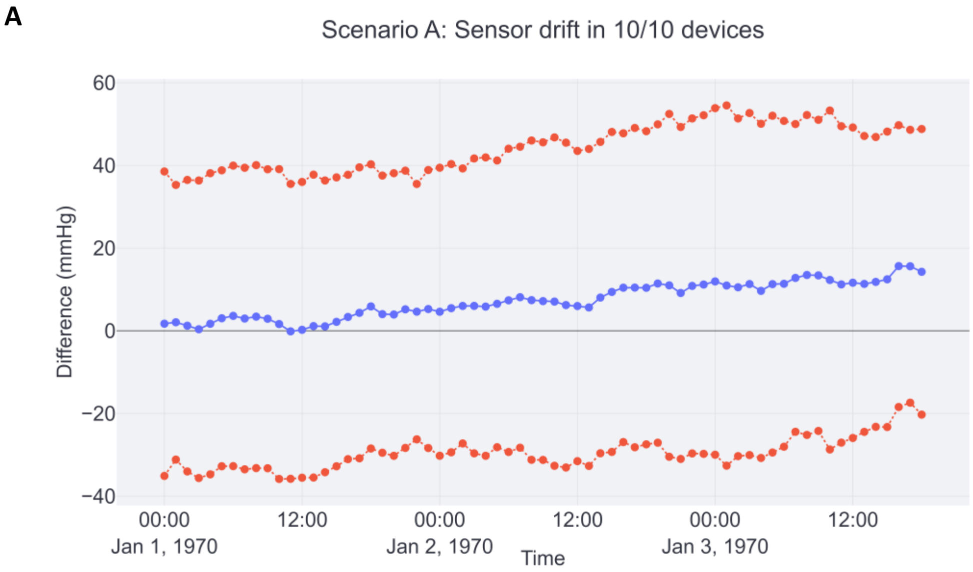
Table 5. Comparison of the four LoA analysis variants.

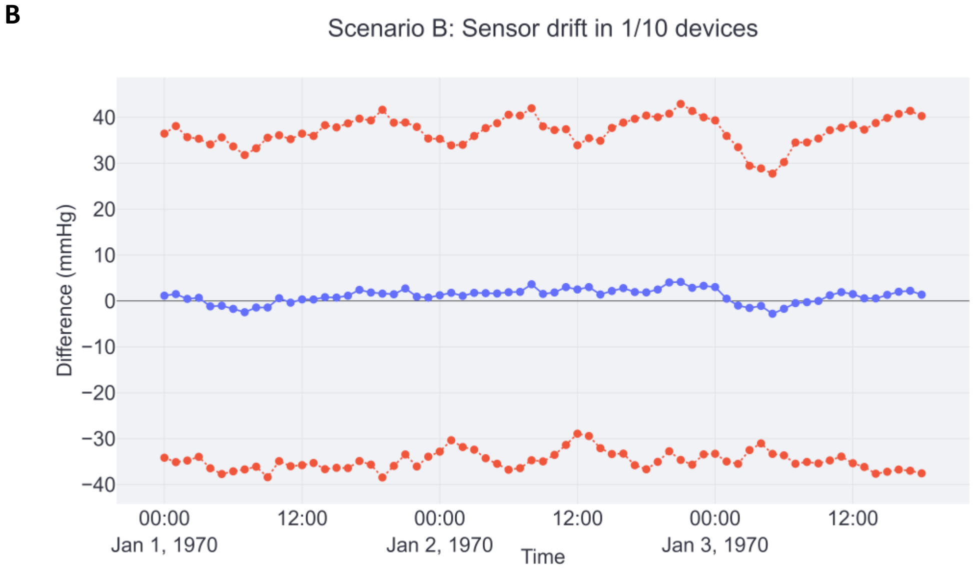
LoA analysis	Bias	Lower 95% LoA	Upper 95% LoA	Within-subject-SD	Between-subject-SD	Total-SD
Classic	2.8861	-10.0845	15.8567	-	-	6.6176
Repeated measurements	2.8861	-10.0868	15.8590	4.6007	4.7584	6.6188
Mixed-effect	2.8187	-10.1988	15.8361	4.7620	4.6297	6.6416
Regression of difference	1.7507 + 0.0085×M	-9.6222 - 0.0026×M	13.1236 + 0.0196×M	-	-	6.6160*

Asterisk indicates that the standard deviation (SD) is based on the bias model only. M, mean of paired differences.

4.3.3 Analysis II: A simulation study to explore the advantages of the longitudinal analysis

The simulation study investigates the potential benefit of the longitudinal analysis, which is not yet possible but could be a potential issue for continuous monitoring. An increasing bias and 95% LoA in the agreement plot (Figure 9A-C) due to sensor drift can be distinguished in scenarios A and C but not noticeable in scenario B. The traditional time series (Figure 9D) illustrates the increasing discrepancy in accuracy between the two devices, but it is harder to detect drift than the agreement plot.





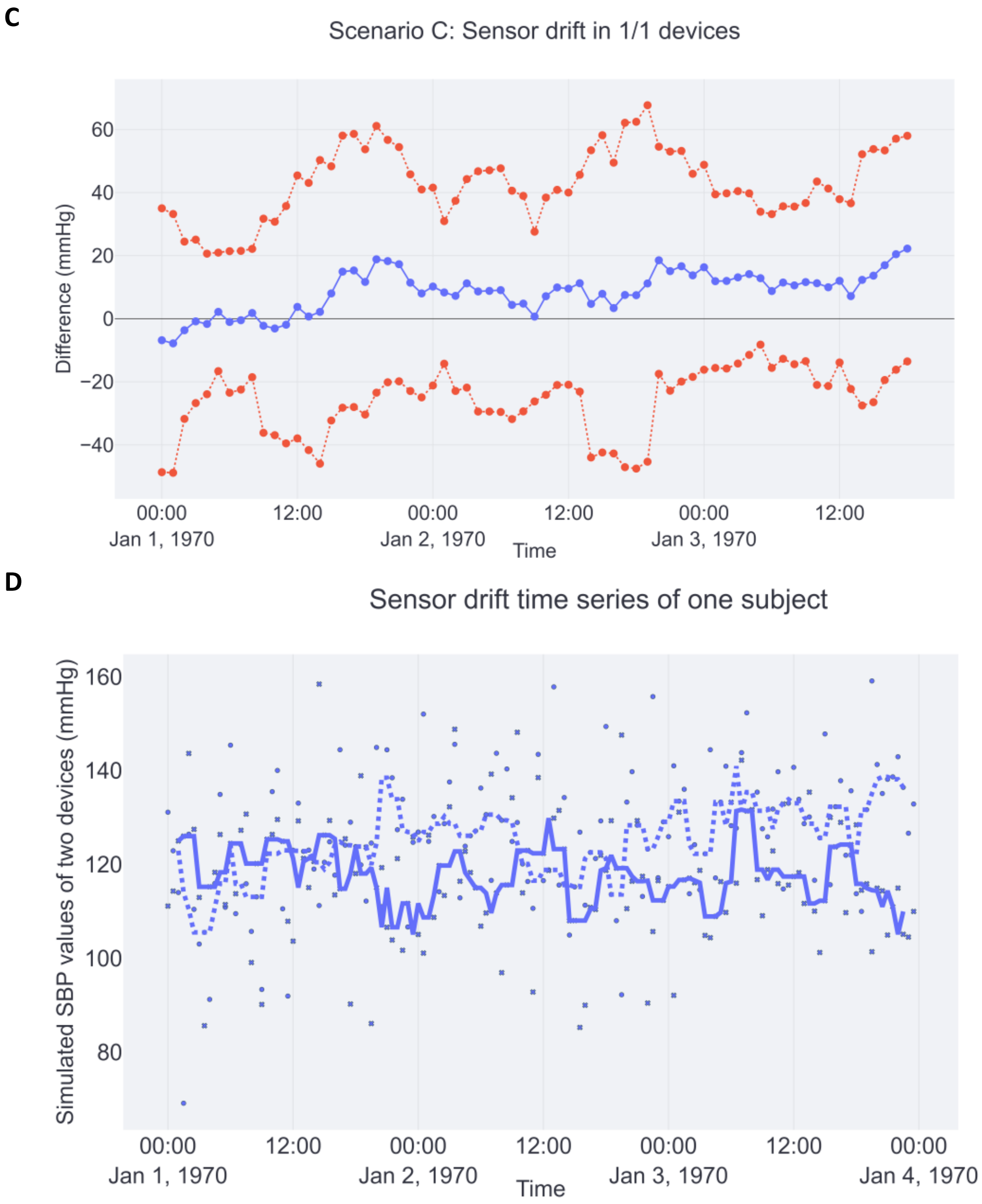
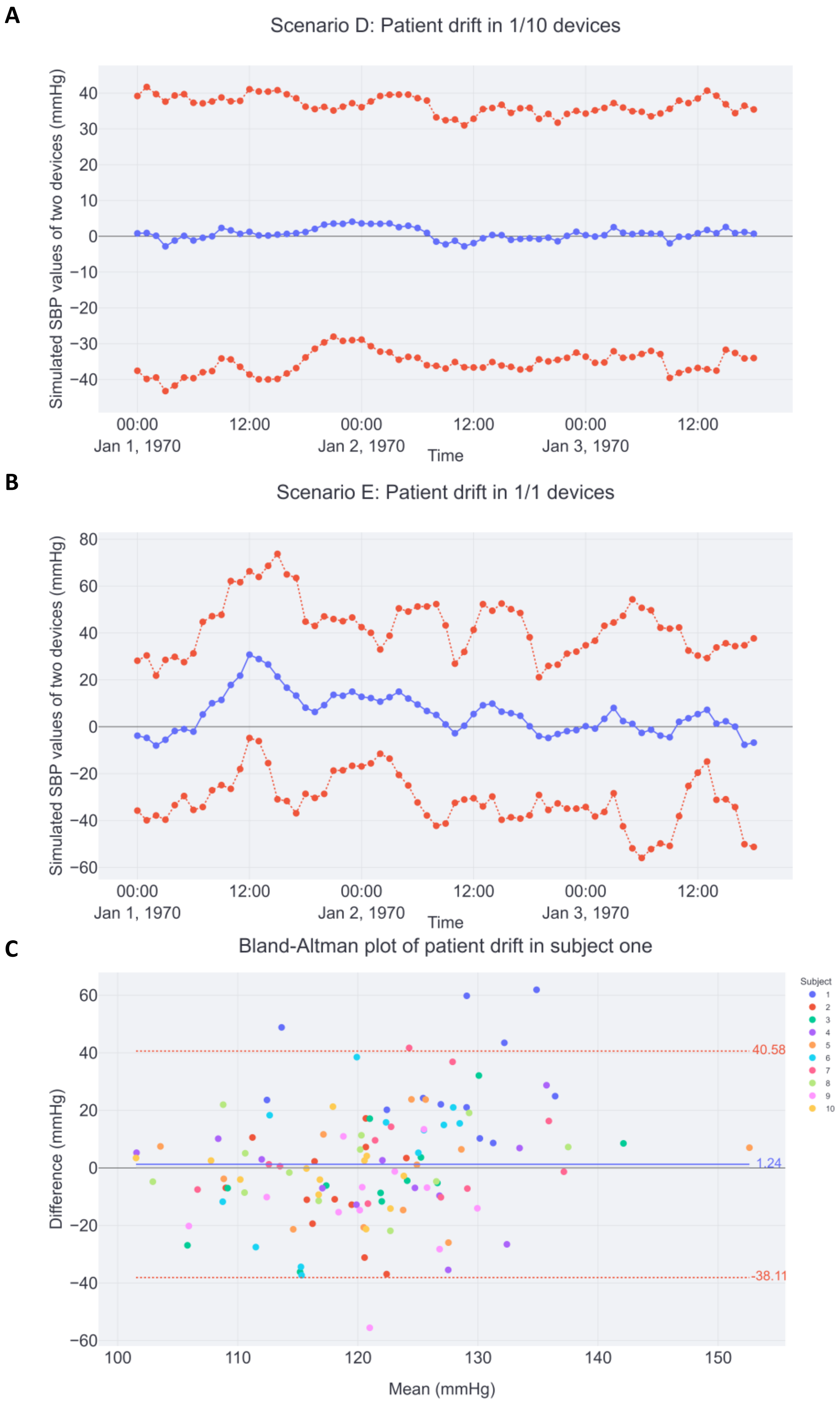


Figure 9. Sensor drift. A) Scenario A: Agreement plot of sensor drift in all ten wearables. B) Scenario B: Agreement plot of sensor drift in one out of ten wearables. C) Scenario C: Agreement plot of sensor drift in single wearable. D) Time series with simulated SBP readings (points) and trendline (median moving window of 30 measurements).

An increasing bias and 95% LoA in the agreement plot (Figure 10A-B) due to patient drift can be distinguished in scenario E but not noticeable in scenario D. An increased bias for subject one in the Bland-Altman plot can be recognised, although this drift could easily be missed (see Figure 10C). The time series plot illustrates the discrepancy in accuracy between the two devices (see Figure 10C).


A
Scenario D: Patient drift in 1/10 devices
Simulated SBP values of two devices (mmHg)
40
30
20
10
0
−10
−20
−30
−40
00:00
Jan 1, 1970
12:00
00:00
Jan 2, 1970
12:00
00:00
Jan 3, 1970
12:00
Time
B
Scenario E: Patient drift in 1/1 devices
Simulated SBP values of two devices (mmHg)
80
60
40
20
0
−20
−40
−60
00:00
Jan 1, 1970
12:00
00:00
Jan 2, 1970
12:00
00:00
Jan 3, 1970
12:00
Time
C
Bland-Altman plot of patient drift in subject one
Difference (mmHg)
60
40
20
0
−20
−40
−60
40.58
1.24
-38.11
100
110
120
130
140
150
Mean (mmHg)
Subject
1
2
3
4
5
6
7
8
9
10

D

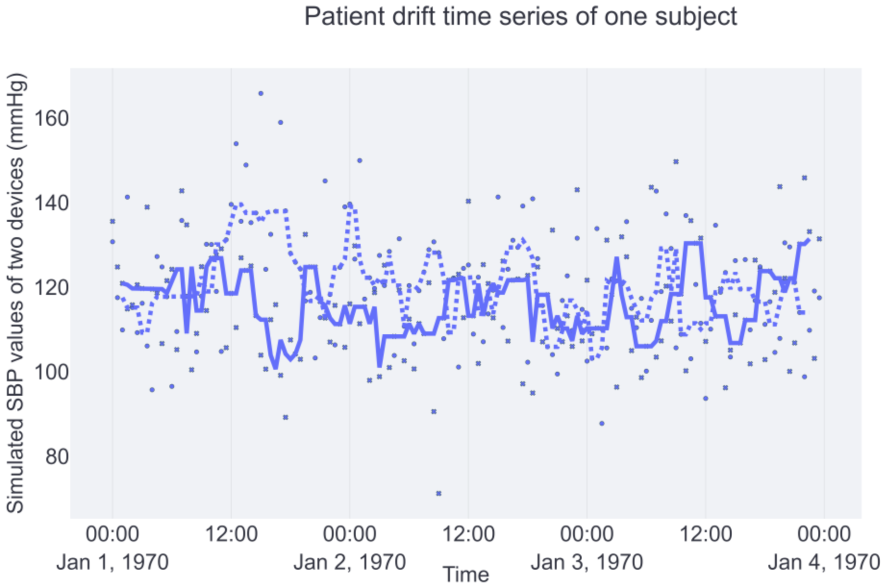


Figure 10. Patient drift on day one from 12.00 to 18.00. A) Scenario D: Agreement plot of patient drift in all ten wearables. B) Scenario E: Agreement plot of patient drift in one out of ten wearables. C) Scenario D: Bland-Altman plot with coloured subjects over the time window on day one from 12.00 to 18.00. D) Time series with simulated SBP readings (points) and trendline (median moving window of 30 measurements).

4.4 Discussion

We designed a toolbox consisting of a Python package and GUI to validate vital sign devices. We utilised the existing LoA analyses to correct for clustering and non-constant agreement across the measurement range. Additionally, we developed a new methodology for the longitudinal analysis to correct for non-constant agreement over time in vital sign devices.

4.4.1 Limits of agreement analysis

Four existing LoA analyses were incorporated in the toolbox, which are the classic [13], repeated measurements [16,64], mixed-effect [14] and regression of difference [16] LoA analysis. These four LoA analyses were correctly implemented, as verified in this study. Analysis I compared the methodological differences between the four types of LoA analysis. As there was only minimal constant agreement over the measurement range in the dataset, this effect is disregarded in the discussion of analysis I. Instead, the emphasis is on correcting for clustering within subjects since multiple measurements were taken per subject.

The classic LoA analysis does not account for the clustering effect in data. Measurements are no longer independent when multiple measurements per subject are taken. The precision is underestimated when utilising the classic LoA analysis since this analysis produces a too-narrow 95% LoA range [13,14]. The importance of correction for repeated measurements within subjects (clustering) is often lacking in validation studies, as explained in section 2.2.2.4. Therefore, we incorporated the repeated measurements and mixed-effect LoA analyses in our toolbox. Only then does the 95% LoA represent the precision by accounting for within-subject variance. Two differences in the correction for clustering between these LoA analyses were observed, namely (I) the mixed-effect LoA analysis had a wider 95% LoA range than the repeated measurements LoA analysis, and (II) the ratio between the between-subject and within-subject-SD varied. The different methodologies caused these effects: The repeated measurements considered subjects as fixed-effects, whereas the mixed-effect LoA analysis considered subjects as random effects. The authors believe that the mixed-effect LoA analysis is preferred for correcting for clustering for two reasons.

1. The included subjects are a representative sample of the population of interest in scientific studies, including validation studies. Considering subjects as random effects recognises that they are a subset of a larger population [14] in contrast to a fixed-effect approach assuming that the included subjects represent the entire population. Therefore, the random effects approach is preferred above the fixed effect approach for these studies.
2. The mixed-effect LoA analysis can correct an unbalanced dataset, such as in Parker's [14], where subjects performing different activities made varying numbers of observations. Our study's dataset was not unbalanced, so no correction was required. If this correction is not made, smaller sample sizes will be overrepresented in the bias. In contrast to the repeated measurements LoA analysis, the mixed-effect LoA analysis can correct for smaller sample sizes by treating factors that indicate smaller sample sizes as fixed effects in the bias model.

Therefore, we believe that the mixed-effect approach should be used to correct clustering. The other LoA analyses are also included in our toolbox and can be used.

4.4.2 Longitudinal analysis

Drift can be detected in the agreement plot when it affects all devices simultaneously (scenario A) or when only the subject with drift is analysed (scenarios C and E). If the drift only occurs in one out of ten subjects (scenarios B and D), the drift is barely noticeable. The drift only influenced the within-subject variance of one out of ten subjects and is therefore not visible in the agreement plot. The longitudinal analysis was used to detect non-constant agreement over time, as the emergence of continuous monitoring gives rise to the potential problem of drift (as outlined in section 2.2.3.4). Drift can only be detected in the collective subjects simultaneously (scenario A) or in a single subject when only that subject is analysed in the longitudinal analysis (scenarios C and E).

The representativeness of the simulation study used in the longitudinal analysis might be reduced. We acknowledge that various limitations could affect the representativeness of the simulation to regular physiological changes in humans. First, the simulation study regards sensor and patient drift, simulating a non-constant accuracy over time. There were no scenarios included simulating non-constant precision over time. We recommend testing if non-constant precision over time could be noticed in the longitudinal analysis. Secondly, all subjects' variations throughout the day were consistent but should be reevaluated. For instance, a patient's activity level may impact within-subject variation, as those lying in bed may exhibit less variation than those who are active and monitored at home. We recommend simulating more scenarios, such as the effect of day-night rhythm [63] on the longitudinal analysis. Third, the parameters of the increase of norepinephrine were based on a population in intensive care, with several factors not representing the situation in vital signs monitoring (such as mechanical ventilation, mean age of 72 years, and various morbidities). For a complete, generable simulation, more emphasis should be on the chosen scenarios, within-subject variation and the population of interest parameters. The focus was not on the representativeness of the simulations but on determining whether a drift could be observed in the longitudinal analysis. We showed that the longitudinal analysis could detect and quantify drift and recommend further exploration of the benefits in non-simulated settings.

4.4.3 Strengths, limitations and recommendations of the limits of agreement analysis and the longitudinal analysis

To date, no package has been created that includes multiple LoA analyses variants. This package's strength lies in three main areas: (I) offering access to various LoA methods, (II) providing tools for evaluating the statistical assumptions required for accurate utilisation of the LoA analysis, and (III) developing a new longitudinal analysis that offers benefits to different end-users (as outlined in section 2.3). The toolbox makes *first time right* [65] validation possible. Specific strengths, limitations, and recommendations of the LoA analysis and the longitudinal analysis are discussed:

4.4.3.1 Precision in the longitudinal analysis

The longitudinal analysis has the benefit of estimating both accuracy and precision over time. In traditional time series, a trend line can indicate accuracy but is not used for precision. The introduction of the longitudinal analysis can quantify both accuracy and precision between measurements.

4.4.3.2 The longitudinal analysis does not incorporate non-constant agreement over the measurement range

The regression of difference LoA analysis is not incorporated in the current version of the longitudinal analysis since the visualisation in the agreement plot required constant agreement over the measurement range. However, this methodological limitation should not prevent incorporating the regression of difference LoA analysis in the longitudinal analysis. Quantifying the non-constant agreement over the measurement range and over time is possible, even though the agreement plot cannot be visualised.

4.4.3.3 Longitudinal analysis is limited to the time window size

The development of the longitudinal analysis provides a framework for assessing accuracy and precision across multiple dimensions, namely over time while accounting for clustering. These effects can be visualised by simultaneously using the agreement and Bland-Altman plots. However, in the longitudinal analysis, only a single time window can be considered at any given time, ruling out the possibility of analysing trends within time windows. The duration of the time window sets a limit on the maximum observable trends, reducing the data resolution [76]. Rather than filtering data into time windows, it would be more effective to incorporate time as a factor in the mixed-effect LoA analysis to analyse trends independent of the chosen time window size.

4.4.3.4 Multiple methodological challenges

The included LoA analyses, as well as the longitudinal analysis, could allow for correction to (I) clustering, (II) non-constant agreement over the measurement range, and (III) non-constant agreement over time, as an extension to the classic LoA analysis. However, the methodological challenge becomes even more complex when correction should be applied for multiple effects. The longitudinal analysis incorporates only correction to clustering and non-constant agreement over time. The other situations are also discussed, as well as the combination of all three effects:

Measurement range and clustering: The current package version does not allow for the correction of clustering and non-constant agreement over the measurement range. Nonetheless, the mixed-effect LoA analysis can simultaneously account for both effects if subjects are treated as a random effect and the mean between the two measurement methods is considered a fixed effect. This feature has not been included yet but can be added in future versions.

Measurement range, over time, and clustering: The mixed-effect LoA analysis could be further extended to correct for all three mentioned effects. The disadvantage of this method is the complex interpretation, as a graphical representation of the data becomes impossible. The user should interpret the agreement regarding the measurement range, time domain and clustering. Further research is needed if this analysis is feasible, and interpretation of this analysis without graphical representations may become too challenging to comprehend. Humans prefer two-dimensional plots over three-dimensional plots [76], and (II) prefer visualisation over numeric values for better understanding [77].

4.4.3.5 The confidence interval is missing

In the current version of the LoA analysis, the 95% confidence intervals of the bias and 95% LoA are not incorporated. We recommend incorporating the confidence intervals in future versions. These confidence intervals show how likely the bias and LoA represent the accuracy and precision between the devices [78].

4.4.4 Graphical user interface

Our GUI simplifies the LoA analysis and the longitudinal analysis, making it accessible to end-users. It also includes tools to evaluate the statistical assumptions required for accurately utilising the LoA analysis and information on these assumptions. With the help of our user-friendly GUI, the complex analysis of the *ValidSense.py* Python package can be performed without requiring a high level of statistical expertise or programming skills. Users can easily follow step-by-step instructions, upload files, and download visualisations. The added value of the GUI is evaluated based on the scientific visualisation guidelines, which are principles and best practices for effectively communicating complex data and scientific findings through visualisations. Scientific visualisation guidelines ensure that visualisations accurately convey scientific information to a wide audience, including experts and non-experts [76]. Images are more accessible for the brain to interpret than numbers [77], making it crucial to follow the guidelines for effective visualisations [76]. The effectiveness of the Bland-Altman plot, agreement plot, and time series plot is evaluated:

4.4.4.1 Bland-Altman plot

The Bland-Altman plot illustrates the agreement between the two measurement devices. In line with the scientific visualisation guidelines of emphasizing the *visualisation of patterns* [76], the Bland-Altman plot provides information about the accuracy and precision of the data. It can expose four types of misbehaviour, namely (I) accuracy (bias), (II) precision (95% LoA), (III) proportional error (trend), (IV) inconsistent variability, and (V) excessive or erratic variability [79]. In addition, the Bland-Altman plot is a straightforward graph that conveys the essential information to assess the agreement between two measurement devices. Therefore, the Bland-Altman plot is an easy-to-interpret plot for assessing accuracy and precision, following the scientific data visualisation guideline of *creating the simplest graph that conveys the information* [76].

Moreover, the authors improved the Bland-Altman plot in the GUI by adding extra features for more in-depth analysis. First, the user can add a heatmap or subplots of the data distribution relative to the y- and x-axis to show the density differences relevant in large datasets. Information about the measurements' density in different regions complies with the guideline of *making density differences apparent in case of overlapping data points* [76]. Second, the user can hover their mouse cursor over a data point to reveal a hover label containing information about that specific point. Third, the user can colour clustered data, such as subjects. The colour clustering and hovering are for outliers detection and trends identification in clusters, following the guideline for *visualising patterns* [76].

However, we think the Bland-Altman plot could be improved by allowing users to identify irregular cluster agreements. It is recommended to provide the user with the option to set trendlines for each cluster to highlight these trends within clusters in the Bland-Altman plot. Adding these cluster trendlines might improve the identification of patient drift between clusters, which was, for example, unclear by assessing the Bland-Alman plot in scenario D of the simulation study.

4.4.4.2 Agreement plot

The agreement plot was developed to visualise the agreement over time. Contrary to the time series plot, the agreement plot displays more relevant information as the accuracy and precision are visualised. Similar to the Bland-Alman plot, the agreement plot is based on the scientific data visualisation guidelines of *visualising patterns* and *creating the simplest graph that conveys the information* [76] by showing accuracy and precision over time.

4.4.5 Recommendations on the graphical user interface

It is also recommended to undergo more elaborate testing procedures to improve the usability of the toolbox. The focus of this thesis was not on the GUI; therefore, we performed only usability testing using the 'think aloud method' [80] on two volunteers. Improvement of usability could be achieved by increasing the number of subjects in the usability testing and setting up a simulation data set to test the error handling process (e.g., empty data or missing data) for improvement in usability.

Moreover, we did not consider the possibility of abusing the toolbox by users. The purpose of the toolbox was to make the LoA and the longitudinal analysis more accessible for end-users, as a high level of statistical expertise or programming skills are not required. There is a potential risk that users perform the LoA analysis without knowing what they are doing or assessing the statical assumptions. In the GUI, we highlight the importance of assessing statical assumptions. Users can ignore the statistical testing of their data, but this comes with the risk of violating assumptions and may lead to decreased validity of their analysis [13,14]. In a future version, we recommend hardcode text in the Bland-Alman plots that statistical assumptions were violated or not tested, which may reduce the risks of abusing the toolbox.

4.4.6 Alternative analysis and visualisations

This section discusses the other analyses and visualisations not incorporated in this toolbox, namely the LoA analysis of Myles, machine learning for trend analysis, correlation plots, four-quadrant plots, Clarke error grid and cycle plot.

LoA analysis of Myles: We did not incorporate the LoA analysis of Myles et al. [15] in our package since the methodology is incorrect in the authors' view. Myles [15] proposed alternative techniques for computing the 95% LoA. They modelled time as a random effect, which was reasonable due to the limited number of independent time points (only seven). However, when using this approach in general, three issues may arise. First, time is a continuous variable with one observation at each time point, leading to autocorrelation between time points based on their proximity. Second, time points may be non-random and fixed by the study design, which makes it challenging to meet the mixed model assumptions of independent and normally distributed random effects with constant variance [14]. Consequently, utilising subjects as random effects is preferable instead of time. The authors believe that the mixed-effect LoA analysis of Parker [14] is the most suitable method for estimating bias and 95% LoA, allowing correcting for clustering (as outlined in section 4.4.1).

Machine learning for trend analysis: As an alternative to the longitudinal analysis, machine learning might provide to identify trends between measurements. The benefit of this approach is that it can uncover the most dominant trends that humans may miss. For example, machine learning could analyse whether subgroups in the data contribute to inaccuracy (such as patients in a specific hospital department) [81]. Once these factors are identified, we can test whether there is a significant relationship between them and inaccuracy. Further investigation is recommended to evaluate the additional benefits of machine learning techniques compared to the suggested longitudinal analysis.

Correlation plots: Correlation assesses the relationship's strength, not the agreement's quantification, as in the LoA analysis. Therefore the methodology of the correlation is inappropriate for assessing the agreement (as elaborately outlined in section 2.1)

Four-quadrant plots: A four-quadrant plot can assess the trending ability between consecutive measurements of two devices. However, this technique has limitations because (I) there are no clear-cut-off values for the definition of clinical agreement [82], (II) when a time delay between the two devices is present, the four-quadrant plot loses its power, and (III) outcomes of the four-quadrant plot are unrelated to their clinical scale (poor trending ability has more severe health risks in the extreme ranges of the measurement range). Therefore, we recommend not using the four-quadrant plot and the longitudinal analysis proposed in this thesis.

Clarke error grid: The Clarke error grid is a tool that assesses the risk of severe consequences associated with measurement inaccuracies in various regions. Initially developed for evaluating the clinical accuracy of glucose sensors [83], it allows for identifying potential regions where inaccurate measurements could impact treatment decisions [33]. The concept of evaluation based on the risk might be beneficial to include in the Bland-Atlman plot, or the Clarke-error grid could be added as a tool in validation studies. We recommend further investigation into the application of marking high-risk regions.

Cycle plot: A cycle plot is a graphical representation of time series data, which displays the data over time by breaking it down into individual cycles. Each cycle represents a set of data points for a given period (such as a month). Agreement plot and time series plot utilising a form of window averaging reduces the data resolution, which is preserved in the cycle plot. Therefore, the cycle plot might be beneficial in case of multiple trends in the data (e.g. sensor drift and day-night rhythm), visualising both trends. An example of a cycle graph is shown in Figure 11, where time-series data is repeated at different time scales, such as monthly data over many years, to visualise both long-term and short-term trends [84]. Although further investigation is required, the cycle plot might reveal trends across multiple time scales in a single display, which may be more difficult to discern from either the agreement plot or the time series plot.

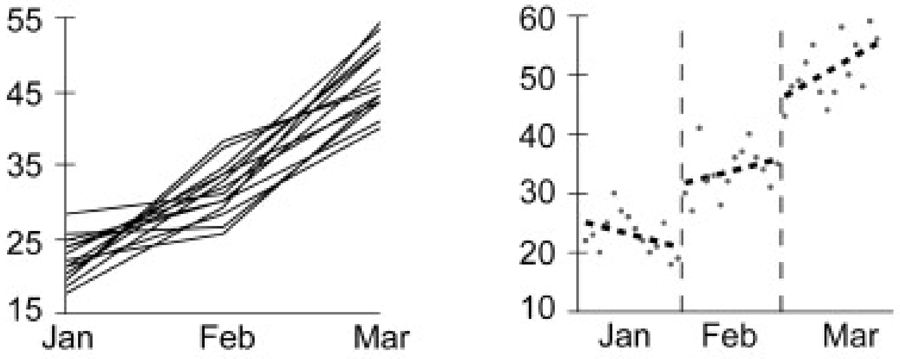


Figure 11. Example of cycle plot (right), compared to traditional time series (left). Both long-term (years) and short-term (monthly) trends can be seen in the cycle plot. Figure derived from [76].

4.5 Conclusion

We developed an open-source toolbox consisting of a Python package and a user-friendly graphical user interface to assess the agreement between two devices. The toolbox allows for validating vital signs monitoring devices the *first time right*, without requiring high-level statical knowledge of programming skills. The four existing LoA analyses are correctly implemented in the toolbox and allow for the correction of multiple measurements per subject (clustering) and non-constant agreement over the measurement range. In addition, the new methodology of the longitudinal analysis is developed to assess non-constant agreement over time, such as sensor and patient drift. Further research is needed to improve the longitudinal analysis and show the benefits of the longitudinal analysis in a real-world setting.

5 VALIDATION OF WIRELESS MULTIPARAMETER SYSTEM

5.1 INTRODUCTION

Changes in blood pressure (BP) are an essential indicator of physiological decline, providing opportunities for early recognition and intervention [3,5–8]. Unvalidated devices with questionable validity are available on the market, as validation studies are often missing [10,11,33]. Especially in cuffless BP monitors, the lack of validation is true, as the appropriate and universal validation protocol is lacking, or existing standards show serious limitations.

Checkpoint Cardio (CPC) developed a continuous, cuffless multiparameter vital sign monitoring system that calculates the SBP and DBP based on photoplethysmographic (PPG) signals. Continuous monitoring enables real-time monitoring and allows for mobility and home monitoring without the inconvenience of any cables. Additionally, ambulatory and home BP measurements are cost-effective, prevent the white-coat effect, and better predict cardiovascular events and mortality [85–88]. Cuffless and continuous BP monitoring could improve healthcare quality in hospitals and at home. So far, it is unknown how the performance of the CPC monitor differs from the gold standard of NIBP measurements. Currently, the only standard available for validating a cuffless BP monitor as the CPC device is the *IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2]. The main objective of the standard is to facilitate innovation and growth in the development of wearable BP monitoring technology while ensuring that the devices are safe and effective for consumers. The standard outlines requirements for an acceptable device accuracy based on three levels of tests, namely (I) static tests, (II) BP-inducing tests, and (III) tests before the next calibration. This study aims to assess whether a wearable and continuous monitoring system can reliably measure blood pressure using the existing validation standard.

5.2 METHODS

5.2.1 Study design, setting and population

We performed a retrospective observational methods comparison study of BP measurements in volunteers using the *IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2]. Volunteers were measured using the CPC device with earclip PPG sensor (Checkpoint Cardio, Kazanluk, Bulgaria) and intermittent using auscultatory NIBP (Microlife WatchBP Office AFIB, Widnau, Switzerland) by two experienced research nurses employed by Checkpoint Cardio. The data collection was performed in the medical research centre of Checkpoint Cardio in Kazanluk, Bulgaria, without specific inclusion or exclusion criteria. The manufacturer aimed to include more than 1600 volunteers to get a representative sample of the Bulgarian population, with volunteers included that (I) are free from haemodynamic problems, (II) diagnosed with hypertension, (III) diagnosed with hypotension. Volunteers were first given 10 minutes to relax on a chair, followed by the administration of three tests, performed by both the CPC device as the reference NIBP device:

1. **Static test**: Volunteers sat on a chair, and five measurements with the CPC and NIBP were taken simultaneously at 5-minute intervals between each measurement.
2. **Exercise test**: Volunteers underwent an exercise protocol that raised their BP. Before the test, their initial BP was measured. Volunteers cycled on a velo ergometer, gradually increasing power from 25 to 50 to 75 watts. The BP was measured simultaneously every 2 minutes during exercise by the CPC and NIBP. Ten minutes after the exercise, BP was measured again.

3. **Medication test**: Antihypertensive medication was administered to volunteers with hypertension to reduce their blood pressure. Different drugs were given to each volunteer, and their BP was monitored for at least three hours using both CPC and NIBP at 10-20 minutes intervals, depending on the medication's administration time and effectiveness in reducing BP to a normal range.

The research nurse rounded the auscultatory NIBP measurements to the nearest five mmHg. CPC measurements with similar timestamps to the NIBP measurements were paired based on the nearest timestamp and saved in a CSV file. More information regarding the rationale for auscultatory NIBP measurements and details about the CPC measurement system can be found in Appendix C and D.

5.2.2 Data preprocessing

The collected data were processed using Python in four steps: (I) measurements with missing values for SBP or DBP in either CPC device or reference device measurements were removed. (II) Measurements of subjects in the medication group without indicated medication time were removed. (III) non-physiological measurements for SBP or DBP in CPC or NIBP were removed. We established physiological ranges for DBP (40-140 mmHg), SBP (80-250 mmHg), and pulse pressure (>20 mmHg). These cut-off values were based on BP data from 19.000 US adults [89] to eliminate non-physiological BP values. (IV) The first paired measurements were designated entry-BP, representing the BP right after calibration.

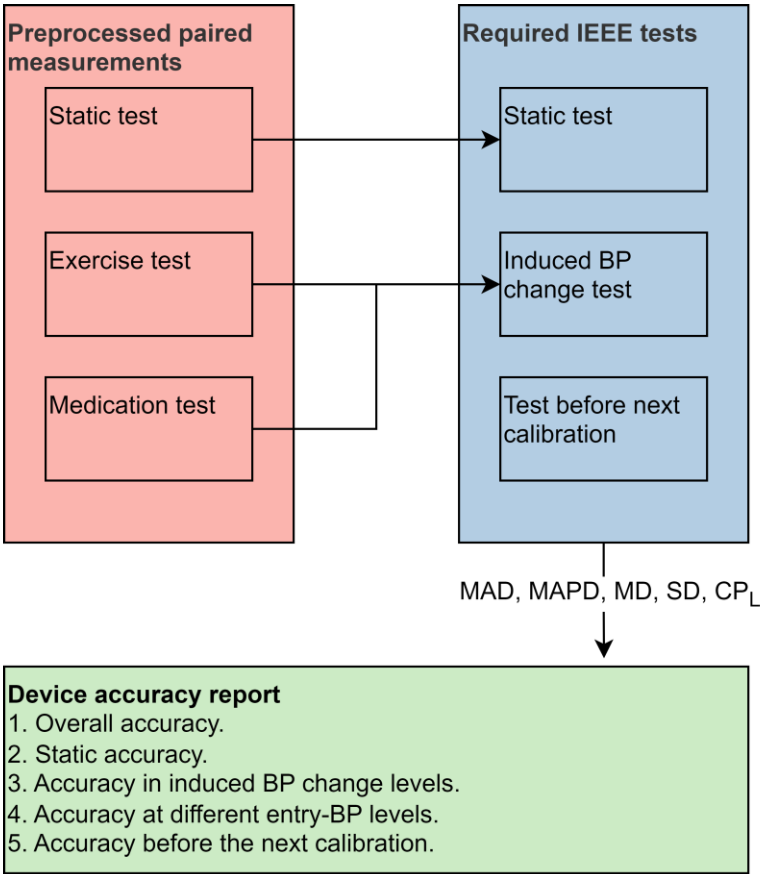

Preprocessed paired measurements
Static test
Exercise test
Medication test
Required IEEE tests
Static test
Induced BP change test
Test before next calibration
MAD, MAPD, MD, SD, CP$_L$
Device accuracy report
1. Overall accuracy.
2. Static accuracy.
3. Accuracy in induced BP change levels.
4. Accuracy at different entry-BP levels.
5. Accuracy before the next calibration.

***Figure 12**. Overview of the preprocessing of the paired measurements, required IEEE tests and device accuracy report. MAD: mean absolute difference. MAPD: mean absolute percentage difference. MD: mean difference. SD: standard deviation. CP$_L$: cumulative percentages of differences falling within the limit of L. n: number or paired measurements.*

5.2.3 Analysis

The preprocessed CPC measurements were analysed using the IEEE *Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2], as summarised in Figure 12. Our study differs on five points from the requirements of the IEEE:

1. We use two subsets of CPC data (exercise and medication tests) for the induced BP change test the IEEE requires.
2. We conducted only one pair of measurements within a short period since we had only one available.
3. The usual practice is to take three measurements for the entry-BP and calculate the average value to remove outliers. However, we have only one measurement available per timestamp and utilise the first measurement for the entry-BP calculation.
4. We cannot assess changes in accuracy over time since we lack data to test its accuracy before the next scheduled calibration.

The data was analysed by introducing the mean absolute difference (MAD), mean absolute percentage difference (MAPD), mean difference (MD), standard deviation (SD), and cumulative percentage of differences falling within the limit of L (CP_L). The MAD, MAPD, MD, SD and CP_L were calculated as,

Equation 1:

$$MAD = \sum_{i=1}^{n} |t_i - r_i| /n,$$

Equation 2:

$$MAPD = \sum_{i=1}^{n} |(t_i - r_i)/r_i| /n \ast 100,$$

Equation 3:

$$MD = \sum_{i=1}^{n} t_i - r_i /n,$$

Equation 4:

$$SD = \sqrt{\sum_{i=1}^{n} (t_i - r_i)^2/(n-1)},$$

Equation 5:

$$CP_L = m/n \ast 100,$$

where t_i is the test device measurement, r_i is the reference device measurement, *m* is the number of measurements where the difference falls within the limit of L, and *n* is the total number of measurements. The IEEE requires a device accuracy report, in which the five statistical outcomes (Equation 1 till Equation 5) were analysed in five categories, including:

1. **Overall accuracy**: using measurements of all three tests. An accuracy level of grade D for the MAD is considered unacceptable (see Table 7).
2. **Static accuracy**: using only the static test.
3. **Accuracy in induced BP change levels**: using the induced BP change test. This test shows the device's accuracy in response to an induced BP change compared to the calibration point. An accuracy level of MAD should be below seven mmHg.
4. **Accuracy at different entry-BP levels**: using the static test. The entry-BP categorises hypertension in subjects as per Table 6 The MAD should be below six mmHg, except for stage 2 hypertension subjects.
5. **Accuracy before the next calibration**: using the static test. The accuracy measures should be consistent with the overall accuracy.

To simplify the analyse and enhance reproducibility, we developed a Python function (*IEEEcufflessBP.py*), included in the *ValidSense* Python package (Appendix G). We utilised this package to analyse our dataset. Moreover, to show the outcomes of both the static and induced BP change levels as required by the IEEE, we utilised the *ValidSense* toolbox (introduced in section 4).

Table 6. Blood pressure classification and requirements.

Blood pressure classification	SBP (mmHg)		DBP (mmHg)
Normal	<120	and	<80
Prehypertension	120-140	or	80-90
Stage 1 hypertension	140-160	or	90-100
Stage 2 hypertension	≥160	or	≥100

SBP: systolic blood pressure. DBP: diastolic blood pressure. Table derived from [2].

Table 7. IEEE grading on overall accuracy, where grading D is regarded as unacceptable.

MAD	Grading
<5	A
5-6	B
6-7	C
≥7	D

MAD: mean absolute difference. Table derived from [2].

5.3 Results

Volunteers were enrolled from August 2019 to May 2022, resulting in 5854 paired measurements in 1411 subjects. An overview of the subject characteristics is summarised in Table 8. As shown in Table 9, the percentages of measurements falling within the extreme ranges specified by the IEEE standard [2] do not meet the required values of 13.6% for all four ranges (-30 to -15 mmHg and 15 to 30 mmHg for SBP, and -20 to -10 mmHg and 10 to 20 mmHg for DBP). However, it should be noted that the total number of measurements is more thfan seven times the required value according to the IEEE standard. Therefore, we can justify the reduced percentage of measurements in the extreme ranges as sufficient.

Table 8. Subject characteristics.

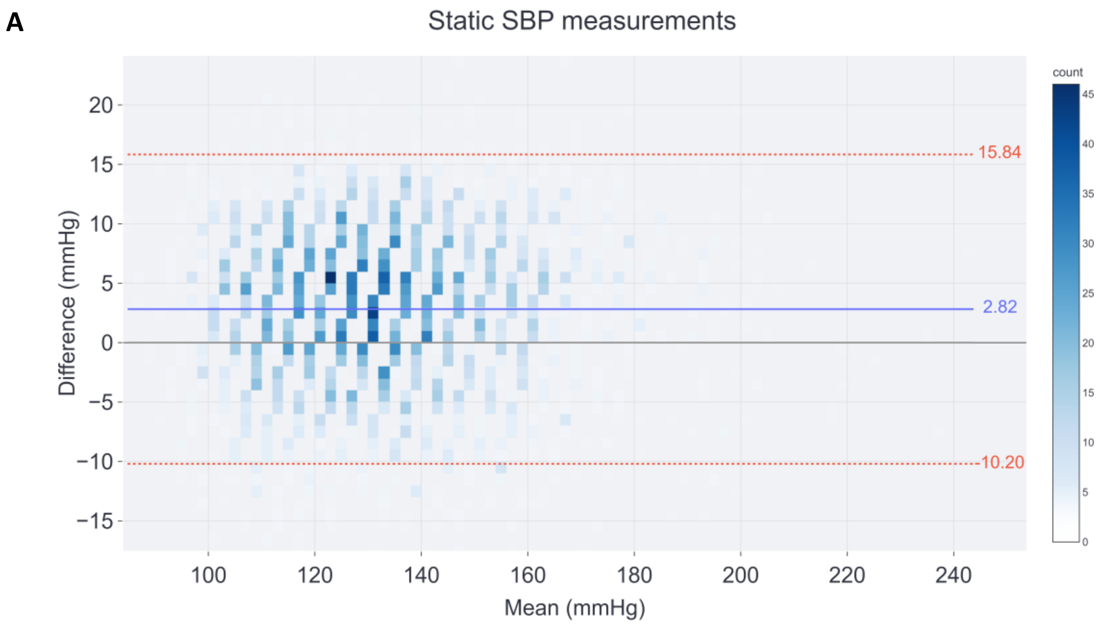
Total number	Subjects, n	1411
	Measurements, n	5854
	Measurements per subject, median [IQR]	4.0 [3.0-5.0]
Entry-BP levels	Normal, n (%)	242 (17.2)
	Prehypertension, n (%)	416 (29.5)
	Stage 1 hypertension, n (%)	321 (22.7)
	Stage 2 hypertension, n (%)	216 (15.3)
Gender	Male, n (%)	631 (44.7)
	Female, n (%)	780 (55.3)
Age	Years, mean (SD)	58.7 (15.4)
	Years, minimum-maximum	15-93
Duration of BP measurements	Minutes, median [IQR]	5.6 [5.0-7.0]

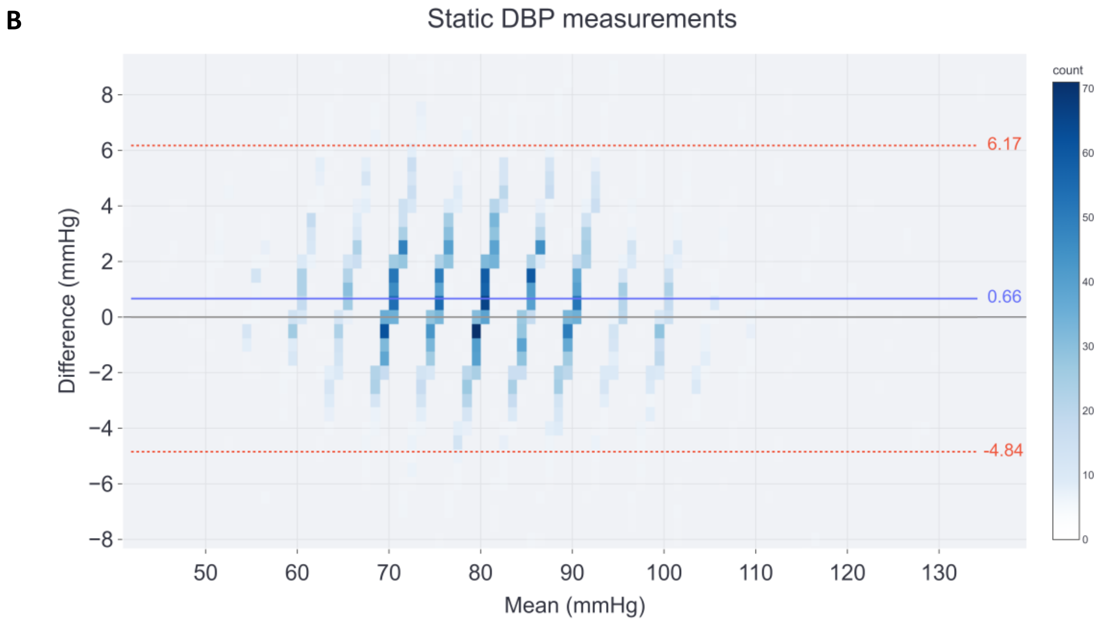
n: number or paired measurements. IQR: interquartile range. SD: standard deviation.

Table 9. The percentage of measurement falling within the blood pressure change level induced from the calibration point.

	Induced BP change level (mmHg)	Exercise test & Medication test	Exercise test	Medication test
SBP	-30 to -15	10.1 %	3.2 %	26.2 %
	-15 to 0	33.0 %	29.1 %	42.4 %
	0 to 15	47.1 %	56.4 %	25.2 %
	15 to 30	9.8 %	11.3 %	6.2 %
DBP	-20 to -10	5.3 %	2.1 %	12.8 %
	-10 to 0	31.8 %	30.1 %	35.9 %
	0 to 10	59.6 %	64 %	49.1 %
	10 to 20	3.3 %	3.8 %	2.1 %

The overall accuracy according to the IEEE standard grades level C for the SBP and level A for the DBP, both passing the test. Accuracy at different BP entry levels passes the test for both SBP and DBP, as the MAD is below the six mmHg. Accuracy before the next moment of calibration cannot be determined. The induced BP change levels' accuracy passes the DBP test as the MAD is below the seven mmHg. However, the SBP fails the test in all induced BP change levels. The mean and SD of both SBP and DBP are higher in the induced BP change levels than in the other IEEE tests depicted in the IEEE accuracy reports (see Table 10 and Table 11). The cumulative percentages revealed that around 18% of the measurements had an SBP error reading of more than fifteen mmHg in extreme ranges of SBP inducement, and even 50% had an SBP error of more than ten mmHg. The SBP measurements of the CPC device fail to pass the IEEE tests. However, the DBP measurements pass the test (neglecting the missing measurements for the accuracy test before the next moment of calibration). The LoA analysis based on the mixed-effect LoA analysis shows a bias (95% LoA) of 2.8 (-10.2 – 15.8) mmHg in the static SBP measurements, 0.7 (-4.8 – 6.2) mmHg in the static DBP measurements, 2.9 (-12.0 – 17.9) mmHg in the induced SBP measurements, and 0.7 (-5.0 – 6.3) mmHg in the induced DBP measurements, with Bland-Altman plots shown in Figure 13. These results in an increased inaccuracy and precision in the SBP measurements compared to the DBP measurements. The time series seen in Appendix F shows that the CPC device fails to detect the trend exhibited in the reference device, with an increasing inaccuracy over time. Additional figures in the same appendix demonstrate that the statistical assumptions of the mixed-effect LoA analysis are satisfied.


A
Static SBP measurements
count
45
40
35
30
25
20
15
10
5
0
15.84
2.82
-10.20
Difference (mmHg)
20
15
10
5
0
−5
−10
−15
100
120
140
160
180
200
220
240
Mean (mmHg)


B
Static DBP measurements
count
70
60
50
40
30
20
10
0
6.17
0.66
-4.84
Difference (mmHg)
8
6
4
2
0
−2
−4
−6
−8
50
60
70
80
90
100
110
120
130
Mean (mmHg)

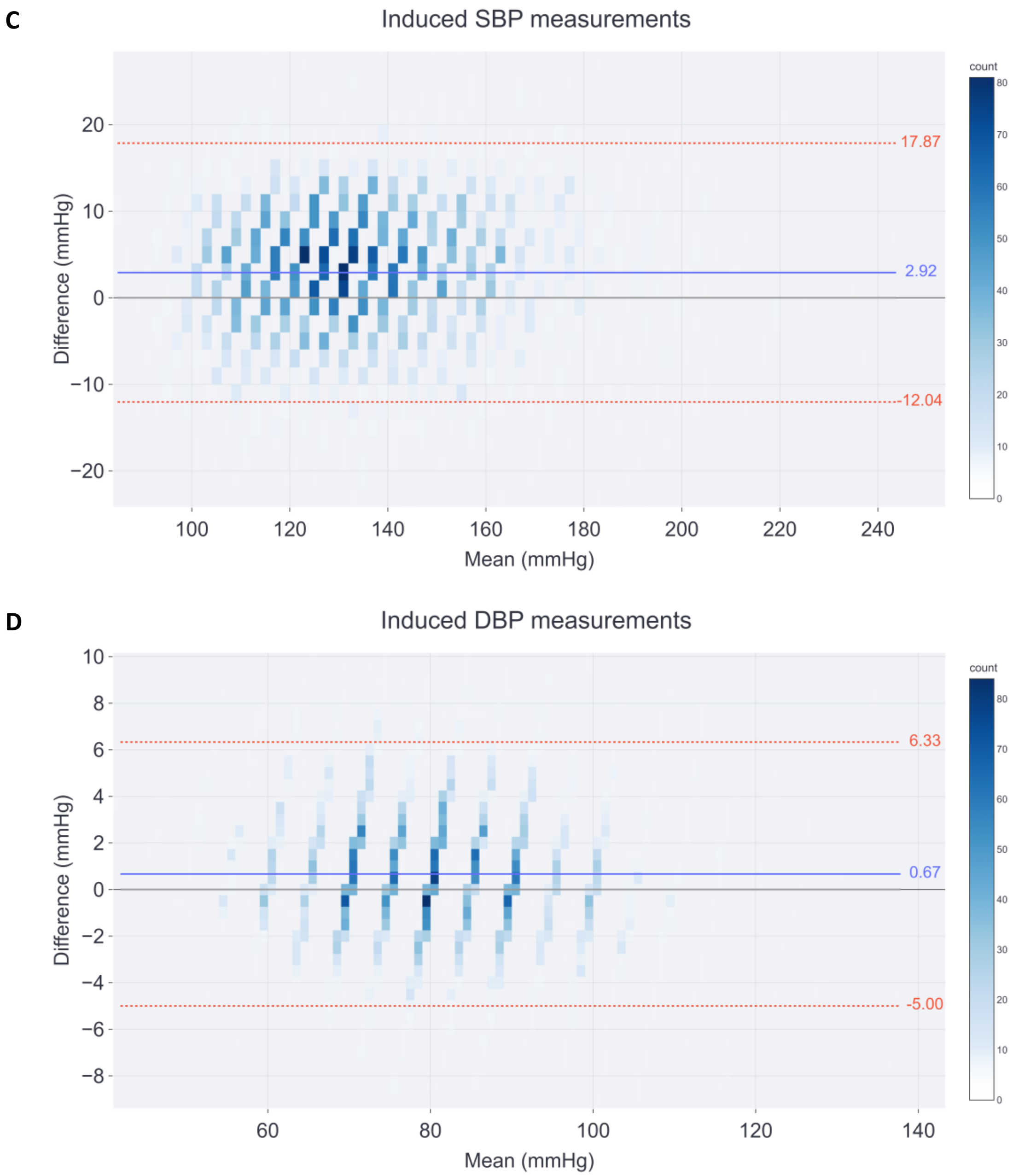


Figure 13. Bland-Altman plot according to the mixed-effect LoA analysis of A) the static SBP measurements, B) static DBP measurements, C) induced SBP levels, and D) induced DBP levels. SBP: systolic blood pressure. DBP: diastolic blood pressure. Heatmap (in blue) to show the density of measurements in different regions to prevent overplotting.

Table 10. Device accuracy report for systolic blood pressure.

Accuracy test	SBP level (mmHg)	n	MAD (mmHg)	MAPD (%)	MD (mmHg)	SD (mmHg)	CP5 (%)	CP10 (%)	CP15 (%)
Overall		5854	6.4	4.9	3.0	7.6	45.7	78.9	95.3
Static		4814	5.9	4.6	2.9	6.6	48.3	82.9	97.4
Induced BP change level Exercise test & Medication test	-30 to -15	71	9.8	7.2	8.0	7.5	21.1	45.1	85.9
	-15 to 0	233	9.8	7.5	8.4	8.4	24.5	54.9	85.8
	0 to 15	332	7.7	5.7	2.5	9.7	42.5	69.0	88.9
	15 to 30	69	9.8	6.5	-0.2	11.5	26.1	50.7	82.6
	All	705	8.8	6.5	4.7	9.8	32.8	60.1	87.0
Induced BP change level Exercise test	-30 to -15	16	9.5	7.2	4.3	10.6	31.2	56.2	81.2
	-15 to 0	144	10.5	8.1	9.0	9.4	22.9	51.4	80.6
	0 to 15	279	7.8	5.7	2.5	9.9	43.0	69.2	88.2
	15 to 30	56	9.7	6.4	0.9	11.6	28.6	51.8	78.6
	All	495	8.9	6.5	4.3	10.4	35.2	61.6	84.6
Induced BP change level Medication test	-30 to -15	55	9.9	7.2	9.1	6.0	18.2	41.8	87.3
	-15 to 0	89	8.6	6.6	7.3	6.4	27.0	60.7	94.4
	0 to 15	53	7.4	5.4	2.4	8.7	39.6	67.9	92.5
	15 to 30	13	10.4	7.2	-4.9	10.3	15.4	46.2	100
	All	210	8.7	6.5	5.8	8.1	27.1	56.7	92.4
BP entry level	Normal	242	5.2	4.9	3.4	5.3	51.2	89.7	99.2
	Prehypertension	416	4.9	3.9	2.6	5.7	59.4	87.5	98.8
	Stage 1 hypertension	321	5.0	3.5	2.3	6.0	58.6	90.0	97.5
	Stage 2 hypertension	216	5.1	3.1	1.3	6.2	53.2	91.2	97.7
Before next calibration		0	-	-	-	-	-	-	-

Table based on the IEEE standard for wearable, cuffless blood pressure measuring devices [1,2]. *MAD: mean absolute difference. MAPD: mean absolute percentage difference. MD: mean difference. SD: standard deviation.* CP_L*: the cumulative percentages of paired differences that fall within a specific limit L. n: number or paired measurements.*

Table 11. Device accuracy report for diastolic blood pressure.

Accuracy test	DBP level (mmHg)	n	MAD (mmHg)	MAPD (%)	MD (mmHg)	SD (mmHg)	CP5 (%)	CP10 (%)	CP15 (%)
Overall		5854	2.2	2.9	0.7	2.9	90.6	99.5	100
Static		4814	2.2	2.8	0.7	2.8	91.4	99.6	100
Induced BP change level Exercise test & Medication test	-20 to -10	42	4.5	6.3	4.0	3.6	54.8	97.6	100
	-10 to 0	252	2.9	3.8	2.3	2.9	81.3	99.2	100
	0 to 10	473	2.2	2.7	-0.1	2.8	92	100	100
	10 to 20	26	3.7	3.8	-2.6	3.4	76.9	100	100
	All	793	2.6	3.3	0.8	3.3	86.1	99.6	100
Induced BP change level Exercise test	-20 to -10	12	3.7	5.5	3.6	3.7	58.3	91.7	100
	-10 to 0	168	2.7	3.6	2.2	2.8	84.5	98.8	100
	0 to 10	358	2.2	2.6	-0.1	2.8	91.6	100	100
	10 to 20	21	3.7	3.7	-2.9	3.0	76.2	100	100
	All	559	2.4	3.0	0.5	3.1	88.2	99.5	100
Induced BP change level Medication test	-20 to -10	30	4.8	6.6	4.2	3.6	53.3	100	100
	-10 to 0	84	3.3	4.3	2.6	3.1	75.0	100	100
	0 to 10	115	2.1	2.7	-0.2	2.7	93.0	100	100
	10 to 20	5	3.6	4.1	-1.5	4.8	80.0	100	100
	All	234	2.9	3.8	1.3	3.4	81.2	100	100
BP entry level	Normal	242	1.8	2.7	0.8	2.4	96.3	99.2	100
	Prehypertension	416	1.8	2.4	0.5	2.3	95.0	100	100
	Stage 1 hypertension	321	2.0	2.4	0.5	2.6	92.8	100	100
	Stage 2 hypertension	216	1.9	2.1	0.2	2.6	94.4	99.5	100
Before next calibration		0	-	-	-	-	-	-	-

Table based on the IEEE standard for wearable, cuffless blood pressure measuring devices [1,2]. *DBP: diastolic blood pressure. MAD: mean absolute difference. MAPD: mean absolute percentage difference. MD: mean difference. SD: standard deviation. CP_L: the cumulative percentages of paired differences that fall within a specific limit L. n: number or paired measurements.*

5.4 DISCUSSION

The CPC device systematically tends to overestimate the BP compared to the reference device. Especially for the SBP, the overestimation is more significant compared to the overestimation of the DBP. The imprecision between the two measurement devices is also higher for the SBP than the DBP measurements.

The validation failed for the SBP device but passed for the DBP device, according to the *IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2]. These results indicate that the CPC device cannot accurately measure induced SBP measurements (caused by antihypertensive medication and physical activity) compared to the reference device. This inaccuracy could result in severe consequences for patients. Overestimation could lead to overmedication with adverse side effects, anxiety and increased costs [31,35,90,91], and underestimation could lead to a missed opportunity to lower cardiovascular risks with lifestyle changes and therapeutics [35]. Therefore, the potential risk of over- and underestimation of hypertension is at risk for patients.

5.4.1 Comparison to other studies

Biobeat (BB-613WP, Biobeat Technologies LTD, Petah Tikva, Israel) developed a continuous BP device. Results of that study [92] showed a lower bias (95% LoA) of -0.1 (7.1 – 6.9) mmHg for SBP and 0.0 (-6.9 – 6.9) mmHg for DBP. The SBP measurements of the CPC system are less accurate but more precise compared to the Biobeat system. When subjects perform physical exercises, the Biobeat sensor becomes less accurate when subjects perform physical exercises, similar to the CPC system. The bias (95% LoA) of the Biobeat system is 0.5 (-7 to 8) for SBP and -1 (-10 – 8) for DBP, showing a similar increase in LoA when subjects are compared to static conditions.

We found two studies comparing studies utilising the IEEE standard to validate continuous BP monitoring devices. First, the research of Kim et al. [93] shows that the MAD threshold for the overall accuracy of seven mmHg is exceeded, which was not exceeded in our study. The SBP during exercise shows a bias (95% LoA) of 7.9 mmHg (-24.9 to 40.7), which is less accurate and less precise than that of our study. Second, Islam et al. [62] also failed to meet the IEEE validation criteria, showing a bias (95% LoA) of 0.1 mmHg (-20.61 to 20.77), indicating more accuracy but less precise SBP measurements compared to the CPC device.

However, the comparison between these studies is limited due to several factors. We mention several differences in the validation study of the Biobeat system, compared to the CPC system: (I) The exercise duration is shorter in the Biobeat study (5 minutes vs at least 30 minutes in this study). (II) The type and effort of the exercise are not defined. (III) Different study populations with less hypertensive and younger subjects. Stage 1 and 2 hypertension of 9% compared to 38%, and a mean (SD) age of 35.1 (23.8) years compared to 58.7 (15.4) years) (IV) Different reference devices probably have different accuracy and precision. (V) LoA analysis is probably based on the classic LoA analysis (although not mentioned), compared to the mixed-effect LoA analysis. (VI) Exact values for the bias and 95% LoA are not mentioned and were visually guessed from the Bland-Altman plot, reducing the reliability of this comparison. All these six factors indicate that an equivalent comparison could not be performed. In addition, the comparison between Kim’s and Islam’s study may also not be valid, as correction for multiple measurements was not performed in the LoA analysis. In addition, the comparison between Kim's and Islam's studies may be invalid as (I) correction for multiple measurements was not applied in the LoA analysis, (II) smaller sample sizes were used in these two studies, and (III) all the IEEE standard tests were not performed. Therefore, standardised validation studies are necessary to enable equivalent comparisons. It is also important to use the correct LoA analysis variants, which is improved by using the *ValidSense* toolbox.

5.4.2 Limitations

There are a few limitations to this study. One major factor is that the accuracy over time of the CPC was not tested, as measurements before the next calibration were not performed. As a result, we could not determine if there was a decrease in accuracy performance according to the IEEE standard or could not perform the longitudinal analysis of the *ValidSense* toolbox. Previous studies [59,60,62,94] have reported accuracy drift in cuffless devices, underscoring the importance of conducting accuracy testing over time.

Another limitation of this study is that most of the readings of the reference device are rounded to the nearest five mmHg (for example, a measurement of 78 mmHg is rounded to 80 mmHg) to avoid fluctuations in the BP readings. For example, if the actual BP shifts from 77 to 78 mmHg, the rounding effect results in an observed BP shift from 75 to 80 mmHg due to only one mmHg increase in the actual BP. The rounding effect can lead to an error of up to 5 mmHg. Unfortunately, the manufacturer responsible for this database did not prevent this rounding step, resulting in unnecessary errors in the reference measurements. These rounding patterns can be observed as stripe patterns in the Bland-Altman plot.

The third limitation of this study is that the sphygmomanometer method was used for the reference device. While this method has been used since Riva Rocci's invention in 1896 and refinement by Korotkoff in 1905 [95], the fundamental measuring principles of cuff-BP devices have remained largely unchanged. While the cuff BP method is a time-honoured technique, some authors question whether this antique method is still the best tool for delivering optimal care (and used as a reference in validation studies) for patients in the 21st century [35,96]. As Appendix C mentions, cuff-BP underestimates the SBP by 5.7 mmHg and overestimates the DBP by 5.5 mmHg [35]. Therefore, we can conclude that the reference device used in this study does not represent the true intra-arterial BP values. Some of the observed inaccuracies may be related to the selected reference device.

The final limitation of this study is that the IEEE recommends taking three pairs of measurements shortly after each other per test to filter out any outliers. However, we only have one measurement per timestamp, which means that outliers in the reference device were not excluded. Therefore, this could potentially lead to falsely worse outcomes for accuracy.

5.4.3 Recommendations

For future research, we suggest incorporating measurements before the next calibration point in the data acquisition process and three sets of measurements for each timestamp to perform the complete IEEE standard to enhance the quality of this validation study.

Furthermore, we observed that the CPC system accurately estimates DBP but requires further improvement in measuring SBP, particularly during induced BP changes. Before reliable use of this system in a clinical environment, further investigation is necessary to understand the reasoning behind these inaccuracies. Then, the manufacturer should improve the reliability of the CPC device for safe clinical usage.

Our study proposes several recommendations for enhancing the current IEEE standard. Firstly, we recommend reducing the degree of freedom in the validation procedure. The standard lacks a defined method for inducing BP changes, which can lead to inconsistencies between studies and devices [10]. In addition, we suggest that the evaluation of devices should include not only accuracy but also precision, which is currently absent from the IEEE standard. The AAMI/ESH/ISO standard for validation of cuff-BP devices incorporates the assessment of the standard deviation (should be below eight mmHg for passing the test), which can be used as a benchmark for comparison [43].

Furthermore, we propose that the assessment criteria be aligned with the device's intended purpose. The intended use for screening, diagnosis, or treatment should define the thresholds for accuracy and precision that must be met in validation studies [94,96]. The consequences of inaccurate or imprecise measurements are increased when monitoring a frailty population [35].

Lastly, we recommend incorporating real-world scenarios into the test protocols to assess the discrepancy between the validation and actual settings where devices are employed. The CPC device is commonly used for monitoring patients in hospitals where vasoactive medications are administered and in-home monitoring settings where patients are physically active. The study included healthy volunteers, while patients in hospitals or eligible for home monitoring are often vulnerable and at higher risk of deterioration. Hence, the authors believe that the device's accuracy must be even more accurate and precise when measuring extreme changes in BP levels since over- and underestimation of hypertension has several clinical implications. We suggest including an assessment of daily activities, as wireless sensors are susceptible to movement during these activities [48]. The impact of these activities should be evaluated to provide a more accurate reflection of accuracy in real-world settings. Evaluations should involve everyday daily activities, such as getting out of bed, walking, cycling, or climbing stairs, to simulate real-world conditions [33].

5.5 CONCLUSION

We performed a validation study according to the IEEE standard to assess the BP measurements of the CPC device compared to the NIBP as a reference device. The study revealed that the SBP measurements obtained by the CPC device do not meet the IEEE standard, whereas the DBP measurements do. The CPC device produced less accurate and less precise measurements for (I) the SBP compared to the DBP and (II) induced SBP by medication or physical activity compared to static SBP. Improvement of the algorithm estimating the SBP reading is recommended to ensure reliable measurements on which physicians and patients could rely.

6 GENERAL DISCUSSION

6.1 MAIN FINDINGS OF THIS THESIS

The market is flooded with vital sign monitoring devices of unknown or questionable quality, leading to potential inaccuracies in the readings and posing a significant risk to patient safety [10–12]. The Limits of Agreement (LoA) analysis is the preferred methodology to assess the agreement between two devices, but it requires high-level statical expertise and programming skills. We developed the *ValidSense* toolbox, an open-source Python package supplemented with a user-friendly graphical user interface, to make the analysis more accessible and user-friendly. Users are guided step-by-step through the LoA analysis, informed about the different LoA analysis variants to correct for clustering and non-constant agreement over time, can perform the correct LoA analysis, and can test statistical assumptions underlying the validity of the LoA analysis. In addition, we developed a longitudinal analysis to assess the agreement over time. We performed a simulation study that showed that sensor and patient drift can be detected. The longitudinal analysis seems promising but requires further study to show the benefits in clinical settings.

The validation of the CPC device has revealed that the SBP algorithm requires further improvement, particularly in cases where SBP is induced. The device appears to be more inaccurate and imprecise under such circumstances. We recommend that the manufacturer enhance their algorithm and use the *ValidSense* toolbox to assess whether the improvements increase accuracy and precision.

We believe that improving the accuracy of vital sign device monitoring devices is essential to ensure that doctors and patients have confidence in these devices. Continuous monitoring of wearable vital sign devices is promising to improve healthcare, but providing evidence that accurate and precise measurements are obtained from these devices is essential for trust in these devices. The *ValidSense* toolbox may improve the quality of vital sign monitoring devices by providing an easy way to assess the agreement between two devices. However, we believe that hurdles regarding the validation setting and protocols also need to be overcome to enhance the quality of vital sign monitoring devices. We elaborate on these discussion points hereafter.

6.2 HURDLES TO OVERCOME TO IMPROVE THE QUALITY OF VITAL SIGNS MONITORING DEVICES

Validation protocols must represent the real-world setting instead of only testing the performance under ideal conditions. In the current practice, validation studies are performed in a controlled environment. However, the usage in clinical practice may be subject to disturbing factors (e.g. patient movement, vasoactive medication, and sensor drift). Therefore, validation protocols should (I) include common daily activities (such as getting out of bed, walking, cycling, or climbing stairs [33]), (II) be validated in clinical conditions, and (III) include assessment over time to show that devices are not subject to a sensor or patient drift. Only then could validation studies show that measurements represent the true physiological state of patients.

Validation protocols need to be more standardised to make studies comparable. For example, we mentioned manufacturers' different methods to induce blood pressure in the IEEE standard or failing to correct for multiple observations per subject in the LoA analysis. Therefore, we believe there is a need for standardisation to make the results of validation studies comparable to each other [43,94].

Clinical consideration should be the bases of the acceptable agreement intervals of the LoA analysis [17]. However, there is a lack of consensus regarding the acceptable ranges [25,53]. It would be desirable for the scientific community to reach a consensus regarding acceptable levels of accuracy and precision in vital sign measurements. In addition, we believe that these ranges should be based on the value of the vital sign measured and the risk of deterioration in the patient population. For example, a very high SBP of 180 mmHg is relevant if there is an inaccuracy of ten mmHg than for an SBP of 120 mmHg. A methodology based on Clark's error grid might help classify the inaccuracy based on the risks of clinical deterioration. Besides, inaccurate measurements have an increased risk of severe clinical outcomes in patients at risk of deterioration compared to healthy subjects. Therefore, stricter cut-off values for acceptable accuracy and precision should be based on the value of the vital sign measured and the risk of deterioration in the patient population.

6.3 A GLIMPSE OF THE FUTURE: VALIDATION OF VITAL SIGN MONITORING DEVICES IN 2030

The hypothetical start-up *VitalWatch Technologies* case (presented in the Introduction, section 1) shows the need for an easily accessible toolbox to validate vital signs monitoring devices. In the box below, we explored the future perspectives of validating vital sign monitoring devices.

Seven years ago, in 2023, *VitalWatch Technologies* received the tools to demonstrate the reliability of their multiparameter wireless vital sign monitoring device for the first time. Before 2023 they lacked the statistical knowledge and programming skills to perform validation study. Using the *ValidSense* toolbox, they could show agreement between their device and a reference device in a simple and accessible way by providing the company with the correct analysis tools and guidance.

Furthermore, *VitalWatch Technologies* found that the device was prone to sensor drift by analysing the accuracy and precision over time. The longitudinal analysis of the *ValidSense* toolbox showed that the inaccuracy increases over time. Therefore, *VitalWatch Technologies* prompted to improve their sensors and algorithms. After each update, the company showed that the device gets more accurate and precise measurements. The *ValidSense* toolbox forms the foundation for trust in the device by doctors and patients. By 2030, continuous vital sign monitoring have reduced nurses' workload, lowered costs and, most importantly, can detect patient deterioration earlier to prompt interventions.

REFERENCES

[1] IEEE Engineering in Medicine and Biology Society. IEEE Standard for Wearable , Cuffless Blood Pressure Measuring Devices IEEE Engineering in Medicine and Biology Society. vol. 2019. 2014.

[2] IEEE Engineering in Medicine and Biology Society. IEEE Standard for Wearable , Cuffless Blood Pressure Measuring Devices IEEE Engineering in Medicine and Biology Society. vol. Amendment. 2019.

[3] Churpek MM, Yuen TC, Winslow C, Robicsek AA, Meltzer DO, Gibbons RD, et al. Multicenter development and validation of a risk stratification tool for ward patients. Am J Respir Crit Care Med 2014;190:649–55. https://doi.org/10.1164/rccm.201406-1022OC.

[4] Brekke IJ, Puntervoll LH, Pedersen PB, Kellett J, Brabrand M. The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review. PLoS One 2019;14:1–13. https://doi.org/10.1371/journal.pone.0210875.

[5] Eddahchouri Y, Peelen R V., Koeneman M, Touw HRW, van Goor H, Bredie SJH. Effect of continuous wireless vital sign monitoring on unplanned ICU admissions and rapid response team calls: a before-and-after study. Br J Anaesth 2022:1–7. https://doi.org/10.1016/j.bja.2022.01.036.

[6] Young MP, Gooder VJ, McBride K, James B, Fisher ES. Inpatient transfers to the intensive care unit: Delays are associated with increased mortality and morbidity. J Gen Intern Med 2003;18:77–83. https://doi.org/10.1046/j.1525-1497.2003.20441.x.

[7] Churpek M, Wendlandt B, Zadravecez F, Adhikari R, Winslow C ED. Association Between ICU Transfer Delay and Hospital Mortality: A Multicentre Investigation. J Hosp Med 2016;11:757–62. https://doi.org/10.1002/jhm.2630.Association.

[8] Ludikhuize J, Brunsveld-Reinders AH, Dijkgraaf MGW, Smorenburg SM, De Rooij SEJA, Adams R, et al. Outcomes associated with the nationwide introduction of rapid response systems in The Netherlands. Crit Care Med 2015;43:2544–51. https://doi.org/10.1097/CCM.0000000000001272.

[9] Taenzer AH, Pyke JB, McGrath SP. A review of current and emerging approaches to address failure-to-rescue. Anesthesiology 2011;115:421–31. https://doi.org/10.1097/ALN.0b013e318219d633.

[10] Sharman JE, O'Brien E, Alpert B, Schutte AE, Delles C, Hecht Olsen M, et al. Lancet Commission on Hypertension group position statement on the global improvement of accuracy standards for devices that measure blood pressure. J Hypertens 2020;38:21–9. https://doi.org/10.1097/HJH.0000000000002246.

[11] Picone DS, Deshpande RA, Schultz MG, Fonseca R, Campbell NRC, Delles C, et al. Nonvalidated Home Blood Pressure Devices Dominate the Online Marketplace in Australia: Major Implications for Cardiovascular Risk Management. Hypertension 2020;75:1593–9. https://doi.org/10.1161/HYPERTENSIONAHA.120.14719.

[12] Jung MH, Kim GH, Kim JH, Moon KW, Yoo KD, Rho TH, et al. Reliability of home blood pressure monitoring: In the context of validation and accuracy. Blood Press Monit 2015;20:215–20. https://doi.org/10.1097/MBP.0000000000000121.

[13] Bland JM, Altman DG. Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement. Lancet 1986;327:307–10. https://doi.org/10.1016/S0140-6736(86)90837-8.

[14] Parker RA, Weir CJ, Rubio N, Rabinovich R, Pinnock H, Hanley J, et al. Application of mixed effects limits of agreement in the presence of multiple sources of variability: Exemplar from the comparison of several devices to measure respiratory rate in COPD patients. PLoS One 2016;11:1–15. https://doi.org/10.1371/journal.pone.0168321.

[15] Myles PS, Cui J. I. Using the Bland-Altman method to measure agreement with repeated measures. Br J Anaesth 2007;99:309–11. https://doi.org/10.1093/bja/aem214.

[16] Bland JM, Altman DG. Measuring agreement in method comparison studies with heteroscedastic measurements. Stat Methods Med Res 1999;8:135–60. https://doi.org/10.1177/096228029900800204.

[17] Giavarina D. Understanding Bland Altman analysis. Biochem Medica 2015;25:141–51. https://doi.org/10.11613/BM.2015.015.

[18] International Organization for Standardization. ISO 5725-1:1994 - Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions. 1994.

[19] Megnigbeto AC, Niakara A, Nebie LVA, Ouédraogo NA, Zagré NM. Validation de la méthode de mesure de la tension artérielle pour une enquête en population sur l’hypertension artérielle chez le Noir africain. Cah d’études Rech Francoph / Santé 2002;12:313–7.

[20] Borges MA, Prado M, Santini TR de S, Barbosa AHP, Moreira AC, Ishibe EI, et al. Development and clinical validation of a non-invasive, beat-to-beat blood pressure monitoring device, compared to invasive blood pressure monitoring during coronary angiography. Einstein (Sao Paulo) 2019;17:eAO4156. https://doi.org/10.31744/einstein_journal/2019AO4156.

[21] Wankum PC, Thurman TL, Holt SJ, Hall RA, Simpson PM, Heulitt MJ. Validation of a noninvasive blood pressure monitoring device in pediatric intensive care patients. Annu Int Conf IEEE Eng Med Biol - Proc 2002;3:1944–5. https://doi.org/10.1109/IEMBS.2002.1053106.

[22] Franco Pessana E, Ramiro S, Gustavo L, Micaela M, Oscar M, Agustin R, et al. A New Approach to Validate the Use of Brachial Blood Pressure to Assess Non-Invasive Aortic Pressure in Human Beings. J Hypertens Manag 2021;7:064. https://doi.org/10.23937/2474-3690/1510064.

[23] Seidlerová J, Tůmová P, Rokyta R, Hromadka M. Factors influencing the accuracy of non-invasive blood pressure measurements in patients admitted for cardiogenic shock. BMC Cardiovasc Disord 2019;19:1–10. https://doi.org/10.1186/s12872-019-1129-9.

[24] Montenij LJ, Buhre WF, Jansen JR, Kruitwagen CL, Waal EE De. Methodology of method comparison studies evaluating the validity of cardiac output monitors : a stepwise approach and checklist †. BJA 2016;116:750–8. https://doi.org/10.1093/bja/aew094.

[25] Leenen JPL, Leerentveld C, van Dijk JD, van Westreenen HL, Schoonhoven L, Patijn GA. Current evidence for continuous vital signs monitoring by wearable wireless devices in hospitalized adults: Systematic review. J Med Internet Res 2020;22. https://doi.org/10.2196/18636.

[26] Appelboom G, Camach E, Abraham ME, Bruce SS, Dumont EL, Zacharia BE, et al. Smart wearable body sensors for patient self-assessment and monitoring. Arch Public Heal 2014;72:1–9.

[27] Iqbal MH, Aydin A, Brunckhorst O, Dasgupta P, Ahmed K. A review of wearable technology in medicine. J R Soc Med 2016;109:372–80. https://doi.org/10.1177/0141076816663560.

[28] Medaval Ltd. Blood Pressure Monitors n.d. https://medaval.ie/resources/EN/pages/bpm-manufacturers-pie-charts.html (accessed May 9, 2023).

[29] European Medicines Agency. Medical devices | European Medicines Agency n.d.

https://www.ema.europa.eu/en/human-regulatory/overview/medical-devices (accessed April 4, 2023).

[30] The Medical Devices Directive (93/42/EEC). Counc Dir 93/ 42/EEC 14 June 1993 Concern Med Devices 1993:18–81. https://doi.org/10.4324/9780080523156-6.

[31] O'Brien E, Alpert BS, Stergiou GS. Accurate blood pressure measuring devices : Influencing users in the 21st century 2020:1138–41. https://doi.org/10.1111/jch.13278.

[32] Smirthwaite A. Clinical evaluation under EU MDR 2021.

[33] Breteler M. A Safer Care Pathway from ICU to Home with wearable & wireless monitoring. 2021.

[34] Medical Device Regulations (MDR). Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council. 2017 n.d. https://eur-lex.europa.eu/legal-content/EN/ALL/?uri= uriserv:OJ.L_.2017.117.01.0001.01.ENG (accessed May 16, 2022).

[35] Sharman JE, Marwick TH. Accuracy of blood pressure monitoring devices: a critical need for improvement that could resolve discrepancy in hypertension guidelines. J Hum Hypertens 2019;33:89–93. https://doi.org/10.1038/s41371-018-0122-6.

[36] White WB, Berson AS, Robbins C, Jamieson MJ, Prisant LM, Roccella E, et al. National standard for measurement of resting and ambulatory blood pressures with automated sphygmomanometers. Hypertension 1993;21:504–9. https://doi.org/10.1161/01.HYP.21.4.504.

[37] O'Brien E, Petrie J, Littler W, De Swiet M, Padfield PL, O'Malley K, et al. The British Hypertension Society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems. J Hypertens 1990;8:607–19. https://doi.org/10.1097/00004872-199007000-00004.

[38] O'brien E, Petrie J, Littler W, De Swiet M, Padfield PL, Altmanu DG, et al. The British Hypertension Society protocol for the evaluation of blood pressure measuring devices n.d.

[39] Tholl U, Lüders S, Bramlage P, Dechend R, Eckert S, Mengden T, et al. The German Hypertension League (Deutsche Hochdruckliga) Quality Seal Protocol for blood pressure-measuring devices: 15-year experience and results from 105 devices for home blood pressure control. Blood Press Monit 2016;21:197–205. https://doi.org/10.1097/MBP.0000000000000186.

[40] O'Brien E, Pickering T, Asmar R, Myers M, Parati G, Staessen J, et al. Working Group on Blood Pressure Monitoring of the European Society of Hypertension International Protocol for validation of blood pressure measuring devices in adults. Blood Press Monit 2002;7:3–17. https://doi.org/10.1097/00126097-200202000-00002.

[41] European Committee for Standardization. BS EN 1060-4:2004 | 6 Oct 2004 | BSI Knowledge n.d. https://knowledge.bsigroup.com/products/non-invasive-sphygmomanometers-test-procedures-to-determine-the-overall-system-accuracy-of-automated-non-invasive-sphygmomanometers/standard (accessed March 27, 2023).

[42] International Organization for Standardization. ISO - ISO 81060-2:2018 - Non-invasive sphygmomanometers — Part 2: Clinical investigation of intermittent automated measurement type n.d. https://www.iso.org/standard/73339.html (accessed November 21, 2022).

[43] Stergiou GS, Alpert B, Mieke S, Asmar R, Atkins N, Eckert S, et al. A universal standard for the validation of blood pressure measuring devices: Association for the Advancement of Medical

Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Collaboration Statement. Hypertension 2018;71:368–74. https://doi.org/10.1161/HYPERTENSIONAHA.117.10237.

[44] Vybornova A, Polychronopoulou E, Wurzner-Ghajarzadeh A, Fallet S, Sola J, Wuerzner G. Blood pressure from the optical Aktiia Bracelet: A 1-month validation study using an extended ISO81060-2 protocol adapted for a cuffless wrist device. Blood Press Monit 2021;0:305–11. https://doi.org/10.1097/MBP.0000000000000531.

[45] Wu N, Zhang X, Wang W, Zhang H. Validation of the Andon KD5031 for clinical use and self-measurement according to the European Society of Hypertension International Protocol. Blood Press Monit 2016;21:310–2. https://doi.org/10.1097/MBP.0000000000000204.

[46] Liu ZH, Liu XY, Wu WJ. Validation of Transtek LS808-B for self/home measurement according to the European Society of Hypertension International Protocol revision 2010. Blood Press Monit 2016;21:352–5. https://doi.org/10.1097/MBP.0000000000000212.

[47] Reshetnik A, Gohlisch C, Zidek W, Tölle M, Van Der Giet M. Validation of the Tel-O-GRAPH, a new oscillometric blood pressure-measuring device, according to the British Hypertension Society protocol. Blood Press Monit 2016;21:307–9. https://doi.org/10.1097/MBP.0000000000000195.

[48] Haveman ME, van Rossum MC, Vaseur RME, van der Riet C, Schuurmann RCL, Hermens HJ, et al. Continuous Monitoring of Vital Signs With Wearable Sensors During Daily Life Activities: Validation Study. JMIR Form Res 2022;6:1–16. https://doi.org/10.2196/30863.

[49] McManus R, Lacy P, Clark C, Chapman N, Lewis P. Reporting of blood pressure monitor validation studies. Blood Press Monit 2018;23:214–5. https://doi.org/10.1097/MBP.0000000000000334.

[50] Elgendi M. Optimal signal quality index for photoplethysmogram signals. Bioengineering 2016;3:1–15. https://doi.org/10.3390/bioengineering3040021.

[51] Taffé P, Zuppinger C, Burger GM, Nusslé SG. The Bland-Altman method should not be used when one of the two measurement methods has negligible measurement errors. PLoS One 2022;17:1–12. https://doi.org/10.1371/journal.pone.0278915.

[52] Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. Resuscitation 2016;102:1–5. https://doi.org/10.1016/j.resuscitation.2016.02.005.

[53] Breteler MJM, KleinJan EJ, Dohmen DAJ, Leenen LPH, van Hillegersberg R, Ruurda JP, et al. Vital signs monitoring with wearable sensors in high-risk surgical patients a clinical validation study. Anesthesiology 2020:424–39. https://doi.org/10.1097/ALN.0000000000003029.

[54] Subbe CP, Duller B, Bellomo R. Effect of an automated notification system for deteriorating ward patients on clinical outcomes. Crit Care 2017;21:1–9. https://doi.org/10.1186/s13054-017-1635-z.

[55] Taenzer AH, Spence BC. The Afferent Limb of Rapid Response Systems: Continuous Monitoring on General Care Units. Crit Care Clin 2018;34:189–98. https://doi.org/10.1016/j.ccc.2017.12.001.

[56] Sessler DI, Saugel B. Beyond ‘failure to rescue’: the time has come for continuous ward monitoring. Br J Anaesth 2019;122:304–6. https://doi.org/10.1016/j.bja.2018.12.003.

[57] Webster CS, Scheeren TWL, Wan YI. Patient monitoring, wearable devices, and the healthcare

information ecosystem. Br J Anaesth 2022;128:756–8. https://doi.org/10.1016/j.bja.2022.02.034.

[58] Morris AS, Langari R. Instrument Types and Performance Characteristics. Meas Instrum 2012:11–37. https://doi.org/10.1016/B978-0-12-381960-4.00002-4.

[59] Hahnen C, Freeman CG, Haldar N, Hamati JN, Bard DM, Murali V, et al. Accuracy of vital signs measurements by a smartwatch and a portable health device: Validation study. JMIR MHealth UHealth 2020;8. https://doi.org/10.2196/16811.

[60] Pickering TG, Hall JE, Appel LJ, Falkner BE, Graves J, Hill MN, et al. Recommendations for blood pressure measurement in humans and experimental animals: Part 1: Blood pressure measurement in humans - A statement for professionals from the Subcommittee of Professional and Public Education of the American Heart Association Co. Circulation 2005;111:697–716. https://doi.org/10.1161/01.CIR.0000154900.76284.F6.

[61] Hu J-R, Martin G, Iyengar S, Kovell LC, Plante TB, Helmond N van, et al. Validating cuffless continuous blood pressure monitoring devices. Cardiovasc Digit Heal J 2023;4:9–20. https://doi.org/10.1016/j.cvdhj.2023.01.001.

[62] Islam SMS, Cartledge S, Karmakar C, Rawstorn JC, Fraser SF, Chow C, et al. Validation and acceptability of a cuffless wrist-worn wearable blood pressure monitoring device among users and health care professionals: Mixed methods study. JMIR MHealth UHealth 2019;7:1–11. https://doi.org/10.2196/14706.

[63] Kallioinen N, Hill A, Horswill MS, Ward HE, Watson MO. Sources of inaccuracy in the measurement of adult patients' resting blood pressure in clinical settings: A systematic review. J Hypertens 2017;35:421–41. https://doi.org/10.1097/HJH.0000000000001197.

[64] Bland JM, Altman DG. Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 2007;17:571–82. https://doi.org/10.1080/10543400701329422.

[65] Kalaria DR, Parker K, Reynolds GK, Laru J. An industrial approach towards solid dosage development for first-in-human studies: Application of predictive science and lean principles. Drug Discov Today 2020;25:505–18. https://doi.org/10.1016/j.drudis.2019.12.012.

[66] Open Source Initiative. The MIT License n.d. https://opensource.org/license/mit/ (accessed April 7, 2023).

[67] Matsubayashi J. Letter to the editor: Inconsistencies between example data and calculation results in Bland, J. M., & Altman, D. G. (2007). Agreement between methods of measurement with multiple observations per individual. J Biopharm Stat 2021;31:868–70. https://doi.org/10.1080/10543406.2021.1968894.

[68] Health Quality Ontario. Twenty-Four-Hour Ambulatory Blood Pressure Monitoring in Hypertension: An Evidence-Based Analysis. Ont Health Technol Assess Ser 2012;12:1.

[69] Morris CJ, Hastings JA, Boyd K, Krainski F, Perhonen MA, Scheer FAJL, et al. Day/night variability in blood pressure: Influence of posture and physical activity. Am J Hypertens 2013;26:822–8. https://doi.org/10.1093/ajh/hpt026.

[70] Marshall TP. Blood pressure variability: The challenge of variation. Am J Hypertens 2008;21:3–4. https://doi.org/10.1038/ajh.2007.20.

[71] Pour-Ghaz I, Manolukas T, Foray N, Raja J, Rawal A, Ibebuogu UN, et al. Accuracy of non-invasive and minimally invasive hemodynamic monitoring: where do we stand? Ann Transl Med

2019;7:421–421. https://doi.org/10.21037/atm.2019.07.06.

[72] Bur A, Hirschl MM, Herkner H, Oschatz E, Kofler J, Woisetschläger C, et al. Accuracy of oscillometric blood pressure measurement according to the relation between cuff size and upper-arm circumference in critically ill patients. Crit Care Med 2000;28:371–6. https://doi.org/10.1097/00003246-200002000-00014.

[73] Kuck K, Baker PD. Perioperative noninvasive blood pressure monitoring. Anesth Analg 2017;127:408–11. https://doi.org/10.1213/ANE.0000000000002619.

[74] Lewis PS, Chapman N, Chowienczyk P, Clark C, Denver E, Lacy P, et al. Oscillometric measurement of blood pressure: a simplified explanation. A technical note on behalf of the British and Irish Hypertension Society. J Hum Hypertens 2019;33:349–51. https://doi.org/10.1038/s41371-019-0196-9.

[75] Dubin A, Pozo MO, Casabella CA, Pálizas F, Murias G, Moseinco MC, et al. Increasing arterial blood pressure with norepinephrine does not improve microcirculatory blood flow: A prospective study. Crit Care 2009;13:1–8. https://doi.org/10.1186/cc7922.

[76] Kelleher C, Wagener T. Ten guidelines for effective data visualization in scientific publications. Environ Model Softw 2011;26:822–7. https://doi.org/10.1016/j.envsoft.2010.12.006.

[77] Ware C. Information Visualization: Perception for Design. Elsevier; 2020. https://doi.org/10.1016/C2016-0-02395-1.

[78] Carkeet A. A Review of the Use of Confidence Intervals for Bland-Altman Limits of Agreement in Optometry and Vision Science. Optom Vis Sci 2020;97:3–8. https://doi.org/10.1097/OPX.0000000000001465.

[79] Riffenburgh RH, Gillen DL. Techniques to Aid Analysis. Stat Med 2020:631–49. https://doi.org/10.1016/b978-0-12-815328-4.00027-9.

[80] Güss CD. What is going through your mind? Thinking aloud as a method in cross-cultural psychology. Front Psychol 2018;9:1–11. https://doi.org/10.3389/fpsyg.2018.01292.

[81] Wu A, Wang Y, Shu X, Moritz D, Cui W, Zhang H, et al. AI4VIS: Survey on Artificial Intelligence Approaches for Data Visualization. IEEE Trans Vis Comput Graph 2022;28:5049–70. https://doi.org/10.1109/TVCG.2021.3099002.

[82] Saugel B, Grothe O, Wagner JY. Tracking changes in cardiac output: Statistical considerations on the 4-quadrant plot and the polar plot methodology. Anesth Analg 2015;121:514–24. https://doi.org/10.1213/ANE.0000000000000725.

[83] Mondal H, Mondal S. Clarke Error Grid Analysis on Graph Paper and Microsoft Excel. J Diabetes Sci Technol 2020;14:499. https://doi.org/10.1177/1932296819890875.

[84] Bradstreet TE. Naomi B. Robbins: Creating more effective graphs. Comput Stat 2007;22:661–3. https://doi.org/10.1007/S00180-007-0064-X.

[85] Myers MG, Haynes RB, Rabkin SW. Canadian Hypertension Society Guidelines for Ambulatory Blood Pressure Monitoring 1999;7061:1149–57.

[86] Imai Y. Prognostic significance of ambulatory blood pressure. Blood Press Monit 1999;4:249–56.

[87] McGrath BP. Ambulatory blood pressure monitoring of healthy schoolchildren with a family history of hypertension. Ren Fail 2010;32:535–40. https://doi.org/10.3109/08860221003706966.

[88] Chobanian A V., Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, et al. Seventh report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure. Hypertension 2003;42:1206–52. https://doi.org/10.1161/01.HYP.0000107251.49515.c2.

[89] Wright JD, Hughes JP, Ostchega Y, Yoon SS, Nwankwo T. Mean systolic and diastolic blood pressure in adults aged 18 and over in the United States, 2001-2008. Natl Health Stat Report 2011:2001–8.

[90] Campbell NRC, McKay DW. Accurate blood pressure measurement: Why does it matter? C Can Med Assoc J 1999;161:277–8.

[91] Haynes RB, Sackett DL, Taylor DW, Gibson ES, Johnson AL. Increased Absenteeism from Work after Detection and Labeling of Hypertensive Patients. Http://DxDoiOrg/101056/NEJM197810052991403 2010;299:741–4. https://doi.org/10.1056/NEJM197810052991403.

[92] Nachman D, Gepner Y, Goldstein N, Kabakov E, Ishay A Ben, Littman R, et al. Comparing blood pressure measurements between a photoplethysmography-based and a standard cuff - based manometry device. Sci Rep 2020:1–9. https://doi.org/10.1038/s41598-020-73172-3.

[93] Kim S, Lee JD, Park JB, Jang S, Kim J, Lee S-S. Evaluation of the Accuracy of a New Cuffless Magnetoplethysmography Blood Pressure Monitor in Hypertensive Patients. Pulse 2018;6:9–18. https://doi.org/10.1159/000484940.

[94] Hu J-R, Martin G, Iyengar S, Kovell LC, Plante TB, Helmond N van, et al. Validating cuffless continuous blood pressure monitoring devices. Cardiovasc Digit Heal J 2023;4:9–20. https://doi.org/10.1016/j.cvdhj.2023.01.001.

[95] F Cartwright FFARC S PF, Jeremy Booth by. A short history of blood pressure measurement. Proc R Soc Med 1977;70:793. https://doi.org/10.1177/003591577707001112.

[96] Mukkamala R, Yavarimanesh M, Natarajan K, Hahn JO, Kyriakoulis KG, Avolio AP, et al. Evaluation of the Accuracy of Cuffless Blood Pressure Measurement Devices: Challenges and Proposals. Hypertension 2021:1161–7. https://doi.org/10.1161/HYPERTENSIONAHA.121.17747.

[97] Brown VA. An Introduction to Linear Mixed-Effects Modeling in R. Adv Methods Pract Psychol Sci 2021;4. https://doi.org/10.1177/2515245920960351.

[98] Frey ME, Petersen HC, Gerke O. Nonparametric Limits of Agreement for Small to Moderate Sample Sizes: A Simulation Study. Stats 2020;3:343–55. https://doi.org/10.3390/stats3030022.

[99] Maas CJM, Hox JJ. The influence of violations of assumptions on multilevel parameter estimates and their standard errors. Comput Stat Data Anal 2004;46:427–40. https://doi.org/10.1016/j.csda.2003.08.006.

[100] Michael H. Kutner, Christopher J. Nachtsheim, John Neter WL. Applied Linear Statistical Models. vol. 29. IRWIN, The McGraw-Hill Companies, Inc.; 1996. https://doi.org/10.1080/00224065.1997.11979760.

[101] Landgraf J, Wishner SH, Kloner RA. Comparison of automated oscillometric versus auscultatory blood pressure measurement. Am J Cardiol 2010;106:386–8. https://doi.org/10.1016/j.amjcard.2010.03.040.

[102] Saugel B, Kouz K, Meidert AS, Schulte-Uentrop L, Romagnoli S. How to measure blood pressure using an arterial catheter: A systematic 5-step approach. Crit Care 2020;24:1–10.

https://doi.org/10.1186/s13054-020-03093-0.

[103] Nightingale-consortium. Nightingale Public Final Report 2021:1–8.

[104] Nightingale. Main software / data document. Software design : AI models for vital sign parameters 2018:1–6.

[105] Elgendi M. On the Analysis of Fingertip Photoplethysmogram Signals. Curr Cardiol Rev 2012;8:14–25. https://doi.org/10.2174/157340312801215782.

[106] Block RC, Yavarimanesh M, Natarajan K, Carek A, Mousavi A, Chandrasekhar A, et al. Conventional pulse transit times as markers of blood pressure changes in humans. Sci Rep 2020;10:1–9. https://doi.org/10.1038/s41598-020-73143-8.

[107] Maharaj R, Raffaele I, Wendon J. Rapid response systems: A systematic review and meta-analysis. Crit Care 2015;19:1–15. https://doi.org/10.1186/s13054-015-0973-y.

[108] van Loon K. Monitoring vital instability in patients outside high care facilities 2016:162.

[109] Sandroni C, D'Arrigo S, Antonelli M. Rapid response systems: Are they really effective? Crit Care 2015;19. https://doi.org/10.1186/s13054-015-0807-y.

[110] Difonzo M. Performance of the Afferent Limb of Rapid Response Systems in Managing Deteriorating Patients: A Systematic Review. Crit Care Res Pract 2019;2019. https://doi.org/10.1155/2019/6902420.

[111] Taenzer AH, Pyke J, Herrick MD, Dodds TM, McGrath SP. A comparison of oxygen saturation data in inpatients with low oxygen saturation using automated continuous monitoring and intermittent manual data charting. Anesth Analg 2014;118:326–31. https://doi.org/10.1213/ANE.0000000000000049.

[112] Weenk M, Koeneman M, van de Belt TH, Engelen LJLPG, van Goor H, Bredie SJH. Wireless and continuous monitoring of vital signs in patients at the general ward. Resuscitation 2019;136:47–53. https://doi.org/10.1016/J.RESUSCITATION.2019.01.017.

[113] Posthuma LM, Downey C, Visscher MJ, Ghazali DA, Joshi M, Ashrafian H, et al. Remote wireless vital signs monitoring on the ward for early detection of deteriorating patients: A case series. Int J Nurs Stud 2020;104:103515. https://doi.org/10.1016/J.IJNURSTU.2019.103515.

APPENDICES

A. LIMITS OF AGREEMENT ANALYSIS

The limits of agreement (LoA) analysis was first reported in British Medical Journal in 1986 [13]. This method was used to compare two measurement techniques to determine their agreement (as explained in section 2.1). Since then, several variants of the LoA analyses have been introduced to serve different purposes. This appendix provides an in-depth examination of the statistical approaches and assumptions involved in the LoA analysis. An overview of these variants is provided in Table 12.

Table 12. Intended use of several variants to the LoA analysis.

LoA analysis variant	Intended use
Classic [13]	Assess agreement in single pair of measurements per subject.
Repeated measurements [16,64]	Assess agreement in multiple measurements per subject.
Mixed-effect [14]	Assess agreement based on the mixed-effect LoA analysis, allowing to correct, for example, multiple measurements per subject or systematic relationship between the difference and mean.
Regression of difference [16]	Assess agreement in a single measurement per subject, with a linear relationship between difference and mean for bias and/or LoA.

Classic LoA analysis

The classic LoA analysis can be applied when evaluating the agreement between a single pair of measurements per subject [13]. This method quantifies the accuracy and precision of the differences by the bias (*B*), standard deviation (*SD*) and 95 per cent of the LoA (*LoA*$_{95}$). When the differences follow a normal distribution, the *SD* is multiplied by a z-score of 1.96 to represent that 95% of data points in a normal distribution fall within 1.96 times the *SD*. The *B*, *SD* and *LoA95* are calculated as,

Equation 6:
$$B = \frac{1}{n}\sum_{i=1}^{n} d_i,$$

Equation 7:
$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(d_i - B)^2},$$

Equation 8:
$$LoA_{95} = B \pm 1.96SD,$$

where d_i is the mean of differences of every paired measurement.

Repeated measurements LoA analysis

The dataset is clustered when multiple measurements on the same subject are recorded. Measurements depend on one other, leading to too narrow 95% LoA estimates when utilising the classic LoA analysis. By building the total variation from two components, the correction for clustering is applied (regarding 'subjects' as cluster variable):

- Within-cluster variation: Differences between multiple observations in the same subject.
- Between-cluster variation: Differences between the averages of the two methods across subjects.

Using a one-way analysis of variance (ANOVA), the between-cluster and within-cluster variance can be estimated (the model in R: difference ~ subject). The *SD* is calculated as,

Equation 9:

$$SD = \sqrt{MS_{residual} + \frac{(MS_{cluster} - MS_{residual})}{divisor}},$$

where *MS*

where $MS_{residual}$ is the mean square of the residuals (beween-cluster variance), $MS_{cluster}$ is the mean square of the clusters, which are both extracted from the ANOVA model. The divisor corrects for the number of measurements and clusters, and is calculated as,

Equation 10:

$$divisor = \frac{(\sum m_i)^2 - \sum m_i{}^2}{(n-1)\sum m_i},$$

where *m* is the number of measurements per cluster, and *n* is the number of clusters to weights the observations correctly [64]. Using Equation 9, the bias and 95% LoA can be calculated using Equation 6 and Equation 8. An elaborate explanation of the repeated measurements LoA analysis can be found in [16,64].

In the ANOVA method, the subjects are regarded as having a fixed effect. We treat them as consisting of the entire population of interest and do not describe them as coming from a distribution of a wider population [14].

Mixed-effect LoA analysis

Repeated measurements and the mixed-effect LoA analysis allow for correction for clustering in the data structure. However, the mixed-effect LoA analysis literature is less well-developed [13,16] than the repeated measurements LoA analysis. There are several reasons why the mixed-effect LoA analysis of Parker et al. [14] is preferred above other methodologies.

1. We can consider subjects as a random effect, as in the repeated measurements LoA method, subjects are regarded as having a fixed effect. When treating them as a fixed effect, we must assume that they comprise the entire population of interest and that the included subjects do not come from a wider population. To generalise the results to the actual population of interest (e.g. all COPD subjects), we should consider subjects as a random effect [14].
2. In the repeated measurements LoA analysis, we must assume that the dependent variable is continuous and the independent variables are categorial. Continuous predictors, such as time or mean in our study, cannot be implemented in the ANOVA model, reducing the flexibility of model building [97].
3. In the mixed-effect LoA analysis, we can correct for multiple confounding factors. In the example of Parker, he uses data with multiple observations per subject, performing multiple activities. To correct the variation due to different activities, Parker included activity as a fixed effect, next to subjects as a random effect.

The mixed-effect LoA analysis estimates the bias and the 95% LoA. The bias (B) is calculated as,

Equation 11:

$$B = \beta_0 + \beta_1 r + \alpha_0 + \varepsilon_i,$$

$$\varepsilon \sim N(0, \sigma_\varepsilon{}^2), \alpha_0 \sim N(0, \sigma_\alpha{}^2),$$

where *r* is the random effect with β_0 as intercept and β_1 as slope of the random effect. α_0 is the constant fixed effect. The total standard deviation (SD_{tot}) is calculated as,

Equation 12:

$$SD_{tot} = \sqrt{\sigma_b + \sigma_w}$$

where σ_b is the between-subject variance, σ_w is the within-subject variance when α_0 indicates the subjects. The between- and within-variance are extracted and summed to estimate the 95% LoA (implementing Equation 12 in Equation 8).

Regression of difference LoA analysis

The independence assumption cannot be met when there is a relationship between the mean and the *SD* throughout the measurement range. Most commonly, an increase in variability is observed as the mean increases. Ignoring these systematic relationships will give a limit of agreement that is too large for small measurements and too small for large measurements, as seen in Figure 2. The differences may be positive linear related to the mean that the 95% LoA will fail to capture accurately. According to Bland and Altman [16], there are two ways to adjust for this relationship: (I) Logarithmic transformation of both measurements, although complicating the clinical interpretation. (II) Regression of difference to model the *SD*. This section focuses on the latter approach since the clinical interpretation remains better to understand.

A simple linear regression is sufficient to model the bias by a first-order formula (the model in R: difference ~ mean). The bias (*B*) is calculated as,

Equation 13:

$$B = \alpha_0 + \alpha_1 m,$$

$$\varepsilon \sim N(0, {\sigma_\varepsilon}^2),$$

where *m* is the mean of the paired measurements, α_0 is the fixed intercept and α_1 is the fixed slope of the mean. If the slope is not significant, the formula can be simplified to Equation 6. The 95% LoA can be determined through two methods, depending on whether the residuals (variability) around the bias is constant, so there is no relationship between the residuals and the mean. If the variability is constant, the *SD* of the residuals from Equation 13 can be used to calculate the 95% LoA can be calculated using Equation 8.

On the other hand, if the variability is not constant, the absolute residuals of Equation 13 are calculated as,

Equation 14:

$$R_{abs} = |\delta - (\gamma_0 + \gamma_1 m)|,$$

where γ_0 is the predicted intercept and γ_1 is the predicted slope of the difference from Equation 13, and δ is the actual difference of the paired measurements. It is essential to use absolute residuals, as using raw residuals would result in the model is adjusted to zero since the mean of residuals is zero. The absolute residuals follow a half-normal distribution $\sqrt{(2/\pi)}\sigma$, so to obtain the *SD*, there should be multiplicated by $\sqrt{(\pi/2)}$ as the *SD* follows a half-normal distribution. *Rabs* is used to modelled to estimate the *SD* around the bias and can be used to calculate the 95% LoA (*LoA95*) as,

Equation 15:

$$LoA_{95} = B \pm 1.96\sqrt{\pi/2}\, R_{abs}.$$

LoA analysis assumptions and testing

Statistical assumptions need to be checked to guarantee valid results of the LoA analysis. Tools are provided to test these assumptions. Table 13 explains the statistical assumptions of the several LoA analyses variants, after which the tools are further explained.

Table 13. Statistical assumption valid for the four LoA analyses variants.

Assumption	Explanation	Testing
Normal distribution of the difference[1,2,3,4]	The precision is correctly described as 95% of the data points fall within the 95% LoA. When the differences are highly skewed (see Figure 14), it may be appropriate to consider non-parametric 95% LoA, as described by Frey et al. [98].	**Histogram**: if the data is normally distributed, the histogram will resemble a bell curve with a symmetrical shape. **Q-Q plot**: if the data is normally distributed, the points will form a straight line.
Constant agreement over the measurement range[1,2,3]	The variability of differences is not dependent on the mean. If the variability of differences increases with an increase in the mean, it can result in too wide 95% LoA for low values and too small 95% LoA for high values [13] (see Figure 2). Use the *regression of difference* LoA analysis [16] to correct for non-constant agreement over the measurement range.	**Scatterplot**: if the data has constant agreement over the measurement range, there should be no systematic relationship between the difference and mean.
Independent observations[1,4]	Measurements are independent of one another. For example, when multiple measurements are recorded per subject, the measurements become dependent and violate the independence assumption. Failing to correct for this dependence can result in 95% LoA that are too narrow and do not accurately represent the precision of the measurements. The *repeated measurement*s or *mixed-effect* LoA analyses can correct this violation by incorporating the within-cluster-SD.	Check the data structure.
Within-cluster-SD independent of cluster-mean[2,3]	The within-cluster standard deviation (SD) of the difference should be independent of the cluster-mean. A constant variability over the multiple measurements is assumed to represent the precision of the measurements (see Figure 15). Violation of non-constant variability may be corrected using a logarithmic transformation [13,15,16].	**Within-cluster-SD plot**: The SD within the cluster on the y-axis should be constant across the cluster-mean on the x-axis.
Normal distribution of residuals[3,4]	The residuals represent the difference between the actual data points and the values predicted by the model. The residuals should be normally distributed with a mean of zero to satisfy the linear regression model's assumptions. When the residuals are not normally distributed, it can lead to biased estimates and inaccurate predictions. Correcting the SD may be appropriate when the residuals are non-uniform distributed, as Maas and Hox described [99].	**Histogram**: if the data is normally distributed, the histogram will resemble a bell curve with a symmetrical shape. **Q-Q plot**: if the data is normally distributed, the points will form a straight line.

Homogeneity of residuals[3,4]	The homogeneity of residuals ensures that the variance of the residuals is constant across the differences. Heteroscedasticity violates the assumption of independent data and suggests that some grouping is present in the dataset [100]. As a result, this can lead to a biased estimation of the regression coefficients, impacting the accuracy and precision estimates.	**Residual plot**: A constant variance across every level of the mean must be seen to conclude that the variance of the residuals is independent.
Exogeneity[3]	Exogeneity ensures that the predictor variables (such as 'subjects') are not affected by the errors in the model. If the exogeneity assumption is violated, it can lead to influenced estimates of the regression coefficients, impacting the accuracy and precision estimates.	**Covariance**: The covariance between the fixed effect and the residuals, as well as between the fixed effect and random effects, should be zero.

Superscripts indicate the statistical assumption valid for the (1) classic, (2) repeated measurements, (3) mixed-effect, and (4) regression of difference LoA analyses variant.

Histogram: Graphical representation of the distribution of numerical data. It is a type of bar chart that shows the frequency or number of values within a range of values. Histograms are commonly used in statistics and data analysis to visualise the distribution of a dataset. Several factors are considered to determine whether a histogram is normally distributed: (I) Symmetrical shape with mean, median and mode all being equal. (II) Bell shape with the highest frequency of observations at the mean and the frequency of observations decreasing moving away from the mean.

Q-Q plot: Graphical tool used to assess whether a data set follows a particular distribution. If the data follows a normal distribution, the Q-Q plot should be linear along the 45°-line. Other distributions will result in deviation from this line.

Scatter plot: Graphical method to display values of two variables for a data set. The data points can also be colour-coded by a third variable. Scatter plots are used to observe relationships between variables.

Residual plot: Displays the predicted values of the model on the horizontal axis and the residuals on the vertical axis. Residuals are the observed values minus the predicted values. Residual plots are a helpful tool for assessing the goodness of fit of a regression model. If the points in a residual plot are randomly dispersed around the horizontal axis, then a linear regression model is appropriate for the data. If there is a pattern in the residual plot, it suggests that the model is not a good fit for the data.

Within-cluster SD plot: Graphical tool that shows the within-cluster SD against the group mean. The within-cluster SD measures the variation of responses within a cluster (such as 'subjects'). An example is given in Figure 15. The within-cluster-SD is displayed on the vertical axis and calculated by (I) squaring the difference in a cluster, (II) taking the average, (III) calculating the square root. The cluster-mean is displayed on the horizontal axis. The within-cluster-SD can assess the consistency of the responses.

Covariance: Covariance is a measure of the relationship between two random variables. The covariance is positive if the two variables increase or decrease together. If one variable tends to increase while the other tends to decrease, the covariance is negative. The covariance is zero if there is no relationship between the two variables. The covariance (Cov) is calculated as,

Equation 16:

$$Cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

where X and Y are the two variables, $\bar{X}$ and $\bar{Y}$ are the means of these variables, and n is the number of observations.

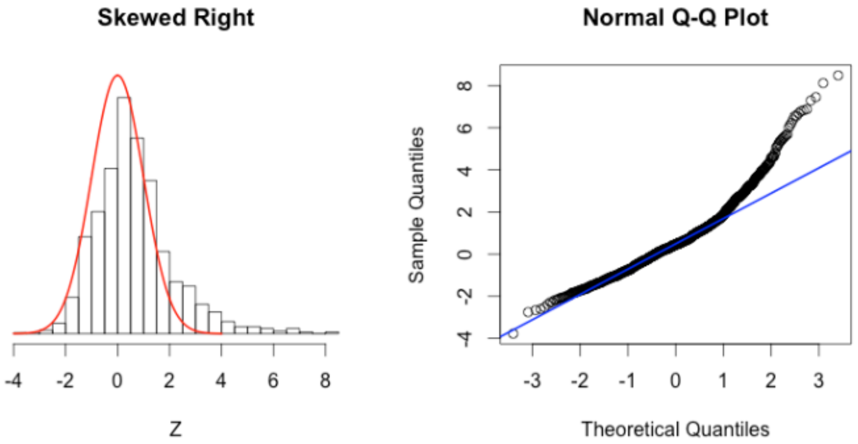


Figure 14. Example of a positive (right) skewed distribution in a histogram and Q-Q plot.

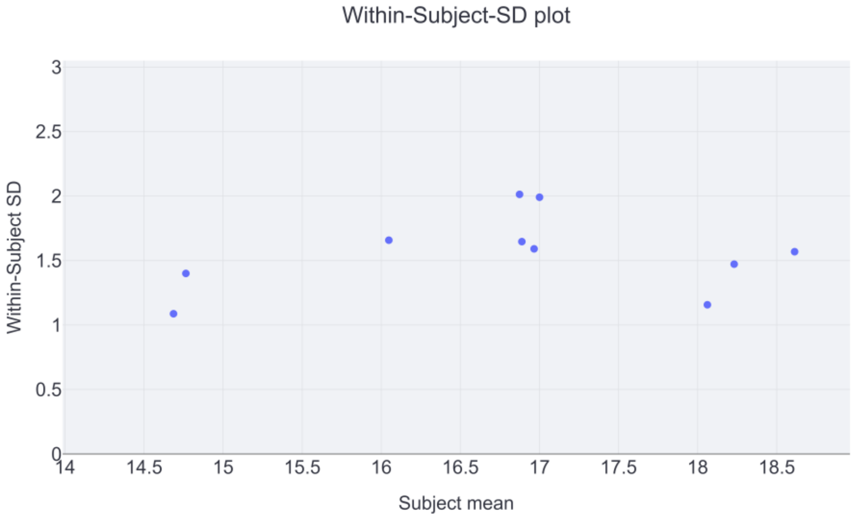


Figure 15. Example of within-cluster standard deviation plot, with 'Subjects' as cluster variable. The within-cluster standard deviation (SD) is plotted on the y-axis, and the cluster-mean on the x-axis.

B. Introduction to regression modelling

This section gives a background about linear regression and linear mixed-effect models. Moreover, we visualise the differences between the two methods.

Linear regression models

Regression analysis is a statistical process for estimating the relationship among variables. A linear regression model represents the functional relationship between the dependent variable and one or more independent variables. The dependent variable is also called the response variable. The ith observation of the dependent variable (y) is calculated as,

Equation 17:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \varepsilon_i,$$

$$\varepsilon \sim N(0, {\sigma_\varepsilon}^2),$$

where x_i refers to the ith observation of the independent variable, α_0 is the fixed intercept of the model, α_1 the fixed slope of the model, ε_i is the random error or residual of the model, reflecting the vertical variation in dependent variable. Independent variables are also called explanatory or predictor variables. Continuous predictor variables are also called covariates, and categorial predictor variables are called factors. With linear regression, the ordinary least squares method can draw a line that minimises the sum of squared differences between the actual data and the line. The vertical spaces between measurements and the line are minimised.

Linear mixed-effect models

A mixed-effect model is a statistical model containing both fixed and random effects. The mixed-effect model is an extension of the linear regression model, consisting of fixed effects only. Fixed effects represent population-level effects that persist in a particular experiment (e.g., age, sex, or ethnicity). Fixed effects do not change over time and are assumed to be measured without error. Random effects are assumed to be values drawn from a larger population of values representing the population (e.g., subjects or items). Random effects are included in a mixed-effect model to account for subjects' behaviour that may differ from the average trend. The ith observation of the dependent variable (y) is calculated as,

Equation 18:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \beta_0 + \beta_1 x_{i1} + \varepsilon_i,$$

$$\varepsilon \sim N(0, {\sigma_\varepsilon}^2), \alpha \sim N(0, {\sigma_\alpha}^2),$$

where x_i refers to the ith observation of the independent variable, α_0 is the fixed intercept of the model, α_1 the fixed slope of the model, β_0 is the random intercept of the model, β_1 the random slope of the model, and ε_i is the random error or residual of the model, reflecting the vertical variation in the dependent variable.

Including random effects for subjects resolves the nonindependence problem of fixed effect-based models (such as linear regression): Some subjects respond differently than others. The random deviations from the mean of the population are called random intercepts. These random deviations in the dependent variable X, are implemented as random slopes.

In contrast to linear regression, linear mixed-effect models have individuals' random intercepts and/or random slopes for the random effect, lowering the residuals. In mixed-effect models, the fixed-intercept estimate represents the average intercept, and the random intercepts allow each random effect (subjects in our case) to deviate from this average.

Visualisation of linear regression and linear mixed-effect modelling

The concept of random intercepts and slopes is visualised in this section, based on the article by Brown et al. [97]. First, we consider a linear regression model (fixed effect only) where to calculate the response time as,

Equation 19:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \varepsilon_i$$

$$\varepsilon \sim N(0, {\sigma_\varepsilon}^2), \alpha \sim N(0, {\sigma_\alpha}^2),$$

where *x* is the word difficulty, α_0 is the fixed intercept of the model, α_1 the fixed slope of the model, and ε_i is the random error or residual of the model. The ε is indicated by the vertical lines in Figure 16A. If we consider a mixed-effect model with random intercepts for subjects, the response time is calculated as,

Equation 20:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \beta_0 + \varepsilon_i$$

$$\varepsilon \sim N(0, {\sigma_\varepsilon}^2), \alpha \sim N(0, {\sigma_\alpha}^2),$$

where *x* is the word difficulty, α_0 is the fixed intercept of the model, α_1 the fixed slope of the model, β_0 is the random intercept of the model, and ε_i is the random error or residual of the model. This model considers that some subjects have different responses than others, reflected by different grey lines. As seen in Figure 16B, the residuals are decreased when the random intercept is included compared to Figure 16A. If we consider a model with random intercept and random slopes included, the response time is calculated as,

Equation 21:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \beta_0 + \beta_1 x_{i1} + \varepsilon_i,$$

$$\varepsilon \sim N(0, {\sigma_\varepsilon}^2), \alpha \sim N(0, {\sigma_\alpha}^2),$$

where x_i refers to the i^th^ observation of the independent variable, α_0 is the fixed intercept of the model, α_1 the fixed slope of the model, β_0 is the random intercept of the model, β_1 the random slope of the model, and ε_i is the random error or residual of the model. This model allows the relationship between y and x to vary across subjects. The residuals decrease even more when the random intercept and random slope are included, as seen in Figure 16C.

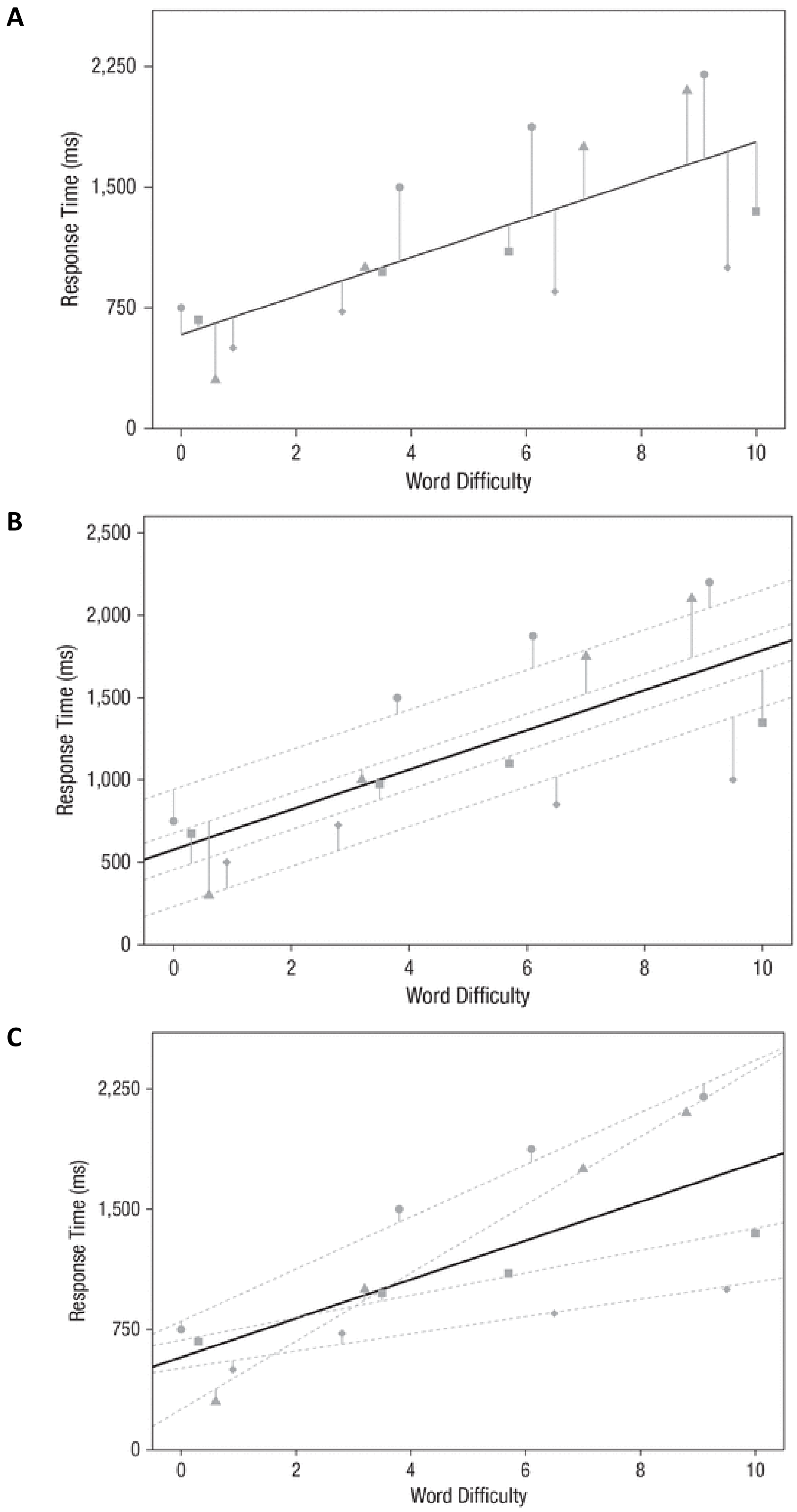


Figure 16. A) Fixed-effect regression line. B) Mixed-effect regression line with random intercepts. C) Mixed-effect regression line with random intercepts and random slopes. Vertical lines represent the deviation from the best line of fit, also called residual error). Grey lines depict model predictions for single subjects. The solid line depicts the estimate for the average fixed effects. Different shapes represent subjects. Figure derived from [97].

C. Hemodynamic monitoring

Blood pressure (BP) monitoring is available on the market in different types: non-invasive, minimally invasive, and invasive BP monitoring. The different types are explained in this section, and the accuracy is discussed.

Oscillometer blood pressure monitoring

Oscillometer BP monitoring is a non-invasive intermittent method. It uses an air-filled pressure cuff placed on the upper arm. With every arterial pulse wave, there is a pulsating volume change in the brachial artery. A transducer detects the increase and decrease of the cuff pressure. Rising pressure in the cuff will stop the arterial blood flow, and therefore pulsation ceases. The cuff pressure is slowly released. Detection of pulsation indicates the SBP. After the cuff is further deflated, there is a point where the pulsations are not detected anymore, indicating DBP.

The accuracy of BP estimation depends mainly on the proper cuff size used and the location on the upper arm. Moreover, the oscillometer techniques are less accurate than the gold standard of arterial cannulation. Oscillometer BP devices underestimate BP in hypertension and overestimate BP in hypotension. Furthermore, it is widely appreciated that the oscillometer (and auscultation) method tends to underestimate the actual systolic and overestimate the DBP [35,71–74].

Auscultatory blood pressure monitoring

Auscultatory BP monitoring is similar to the oscillometer method. Instead of measuring the BP in the cuff, the auscultatory method uses a stethoscope to detect Korotkoff sounds in the artery. The first sound corresponds to the SBP, and the latter to the DBP [71].

Most of the accuracy problems of the oscillometer method also apply to the auscultatory method. Landgraf et al. [101] found discrepancies between the two methods, with higher BP in the auscultatory method. Discrepancies are increased in patients above 65 years.

The inaccuracy of the cuff-based BP measurements is illustrated by Sharman and Marwick [35]. On average, the cuff-BP underestimates BP, underestimates intra-arterial brachial SBP by 5.7 mmHg and overestimates DBP by 5.5 mmHg. Only 33% of the cuff BP were within the ± 5 mmHg from intra-arterial values. Therefore, we conclude that the commonly used auscultatory method is inaccurate.

Arterial line

An arterial line is one of the direct and continuous methods to measure the systolic and diastolic pressure in the arteries, indicated for high-risk surgical and critically ill patients. An arterial line is inserted in the radial or brachial artery. A saline-filled, non-compressible tube between the arterial line and the pressure transducer is placed. The pressure transducer is placed at the heart level, measuring the arterial waveforms on which the systolic and diastolic pressure can be calculated [71].

Inaccuracies in this method can occur, for example, due to an obstructed cannula, blood clotting in the arterial line, or incorrect levelling and zeroing of the pressure transducer. Training and education about this type of BP monitoring are essential to guarantee accurate measurements [102].

D. Checkpoint Cardio System

The Nightingale project was initiated to address the requirement for improved wireless monitoring of vital signs and identifying high-risk patients. The project employs Checkpoint Cardio (CPC), a monitoring device that enables remote wireless monitoring without physically attaching stationary bedside monitoring systems. Figure 17 displays the sensor measuring heart rate (HR), respiratory rate (RR), systolic blood pressure (SBP), diastolic blood pressure (DBP), peripheral capillary oxygen saturation (SpO2), body temperature, and activity.

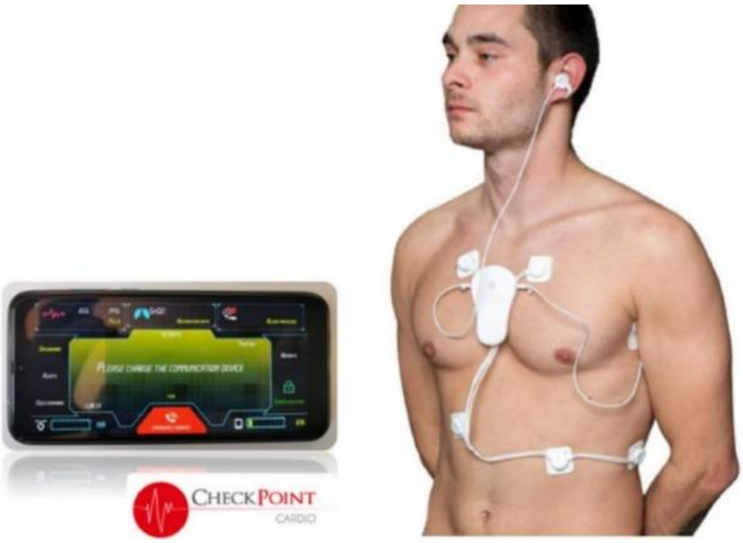


Figure 17. Continuous monitoring device from Checkpoint Cardio. Figure derived from [103].

Sensor specifications

The CPC sensor comprises various sensor modalities that cater to the specific sensing requirements of each vital sign. The ECG module determines the HR and uses a 1-lead, 3-lead, or 12-lead ECG. The RR is derived through a transthoracic impedance. Temperature is measured using a separate module that includes a thermistor. The accelerometer in the main body of the sensor is used to estimate patient activity and body position. The peripheral capillary oxygen saturation (SpO2) is measured by analysing the pulse wave peaks from the PPG sensor.

To measure SBP and DBP, multiple sensor components are required. These measurements are obtained from the pulse plethysmography (PPG) signal, measured using the ear sensor, and the ECG signal. The pulse transit time (PTT) is derived from these signals, explained in the subsequent section. The SBP and DBP measurements can be obtained using a model based on the PPG signal or a stethoscope model. The first model combines the R-peak in the ECG wave with the pulse wave in the PPG sensor, built into an ear lobe or finger sensor module. The second model employs a stethoscope signal measuring the second heart tone, calculating the BP. In both BP models, calibration must be manually performed using a BP cuff [104].

Pulse Transit Time

The CPC systems use a PPG sensor to detect changes in microvascular blood volume. Another parameter, PTT, is determined by the time it takes for a pulse to travel between two arteries (as shown in Figure 18A) or between the R-peak in the ECG signal and the pulse wave peak in the CPC system [105]. The PTT is affected by arterial wall properties, such as vasoconstriction of smooth muscle cells or arterial wall stiffening due to ageing [106], and is inversely related to BP (as demonstrated in Figure 18B). The PTT depends on blood flow and arterial wall characteristics and increases with decreased cardiac output and vascular tone. Consequently, it can reflect changes in BP.

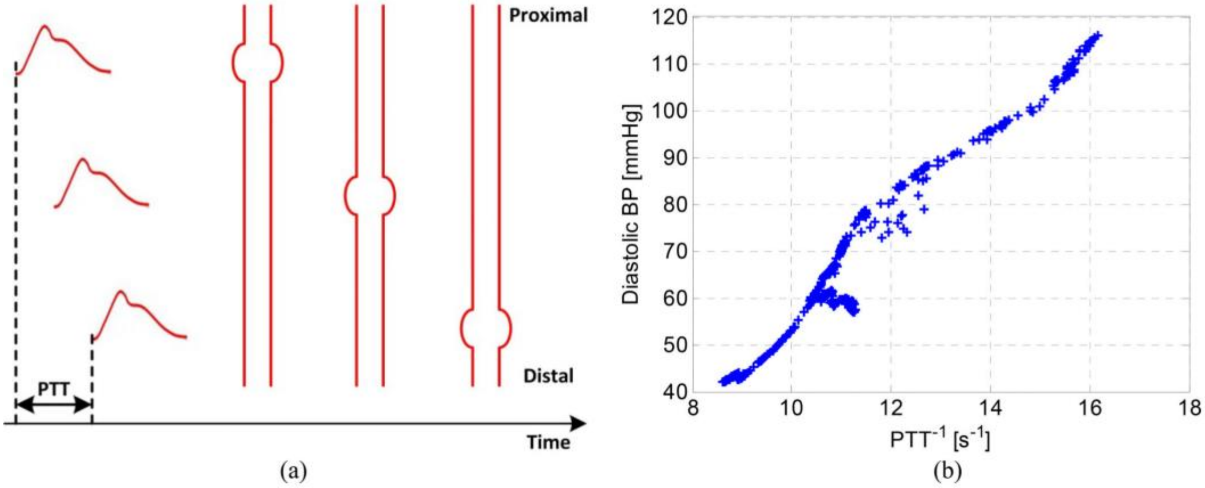


Figure 18. Pulse Transit Time (PTT) for Blood Pressure (BP) monitoring. A) The time delay for the pressure wave to travel between two artery sites determines the PTT. B) Inversely relationship between PTT and BP. Figure derived from [96].

E. Early Warning Score and Rapid Response Teams

Detection of clinical deterioration was attempted to improve by introducing early warning scores and rapid response teams. In this appendix, the results are discussed.

Early Warning Score & Rapid Response Teams

Early recognition and adequate therapy for deteriorating patients were attempted to improve through the worldwide introduction of early warning scores (EWS) and rapid response teams (RRT). In EWS, an aggregated score is calculated, based on the degree of deviation from normal physiology for multiple parameters, such as respiratory rate (RR), heart rate (HR), systolic blood pressure (SBP) and oxygen saturation (SpO2). Higher scores indicate vital instability in patients. In vital unstable patients, the RRT is alarmed, in which trained healthcare professionals are consulted. RRT assesses the patient, optimises diagnostic work-up, and starts necessary clinical progression interventions. An effective response after patient deterioration depends on detecting patient deterioration (afferent limb) and prompting therapeutic interventions (efferent limb), see Figure 19. Maharaj et al. (2015) found that RRT implementation results in a reduction of hospital mortality (RR 0.87, 95% CI 0.81-0.95) and cardiopulmonary arrest (RR 0.65, 95% CI 0.61-0.70) [107].

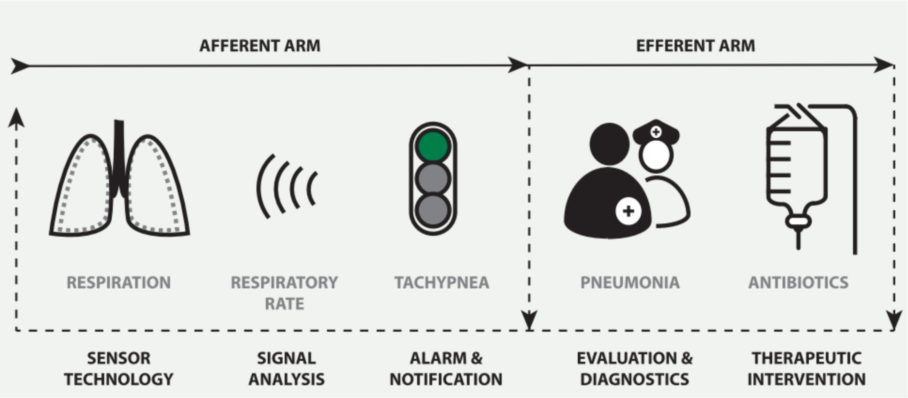


Figure 19. Monitoring system. The primary goal is for the afferent limb to detect physiological parameters and recognise vital abnormalities. When these are present, the efferent limb activates health care professionals to evaluate the patient and prompt therapeutic interventions. Figure derived from [108].

Failure-to-rescue remains

However, failure-to-rescue events continue with the EWS and RRT in place [9], also known as a failure of the afferent limb [109,110]. Intermittent vital sign monitoring might be insufficient to detect clinical deterioration; therefore, increasing the frequency of measurements is desirable [55]. Implementing continuous monitoring is preferred above increasing the intensity of vital sign checks by nurses due to time limitations and budget constraints. Moreover, intermittent manually collected vital signs are less accurate in reflecting the patients’ actual physiological state than continuous real-time monitoring [111–113].

F. Additional figures

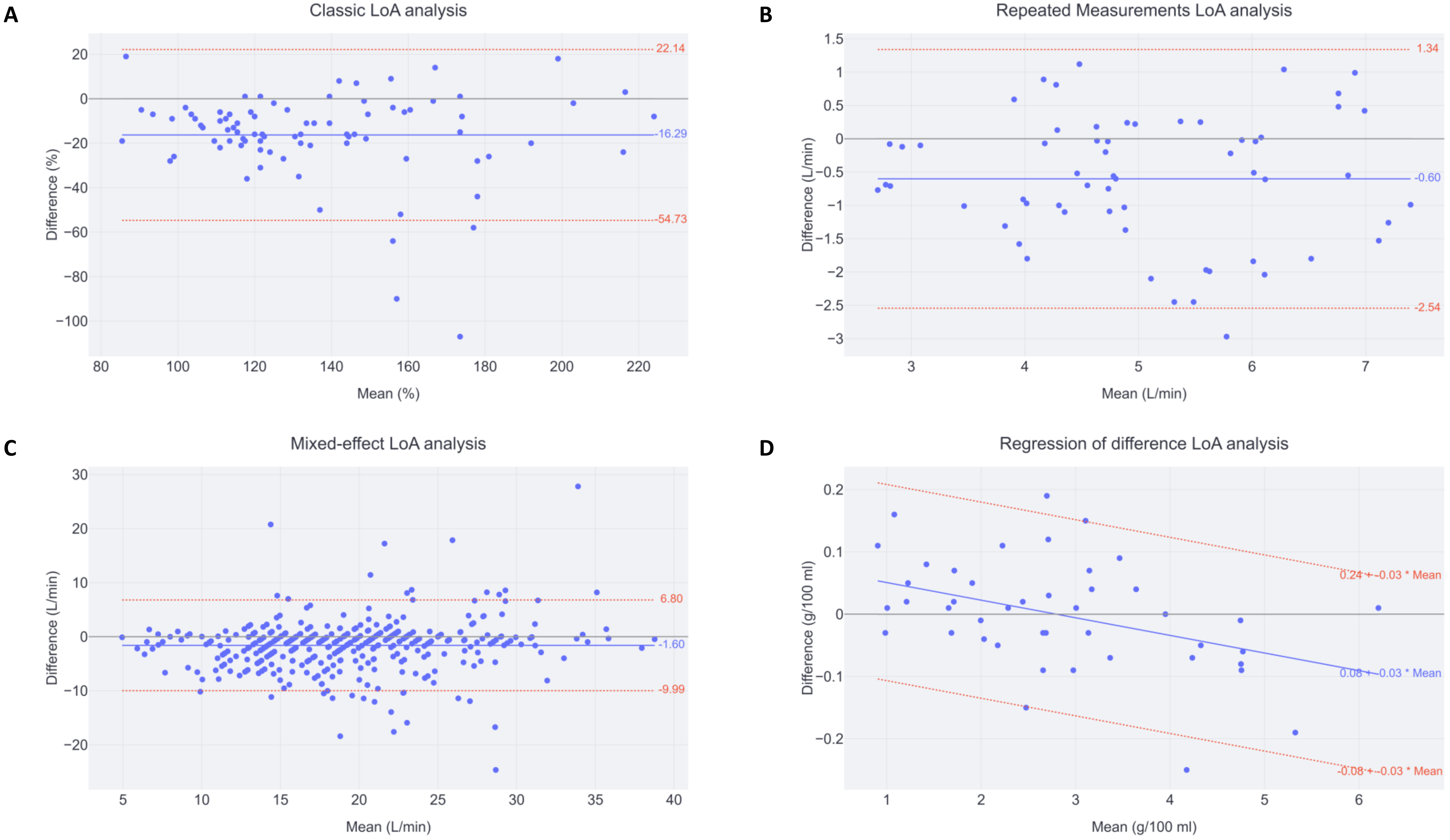


Figure 20. Bland-Altman plots to verify correct implementation based on original datasets of the four LoA analyses: A) Classic, B) Repeated measurements, C) Mixed-effect, D) Regression of difference.

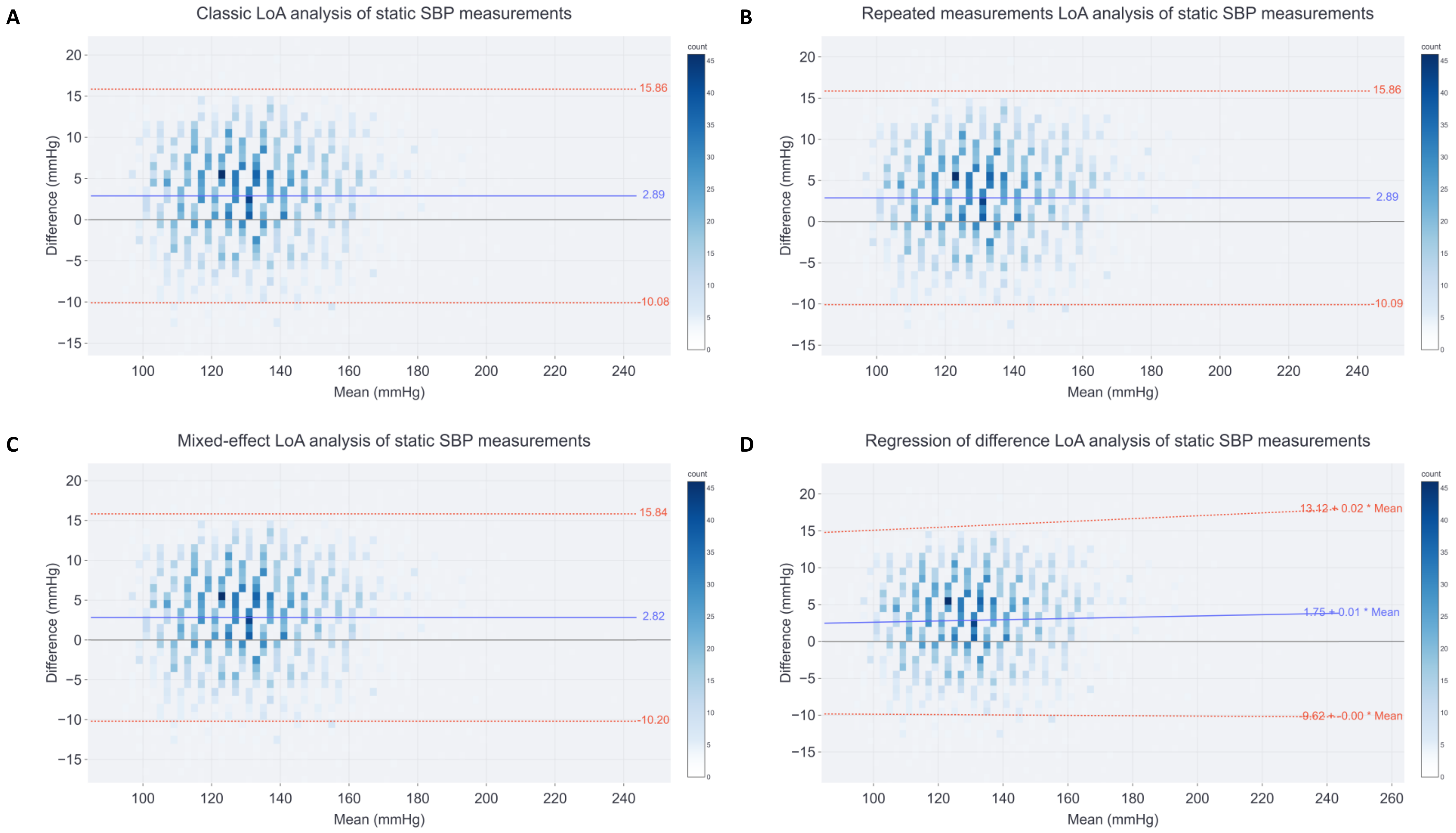
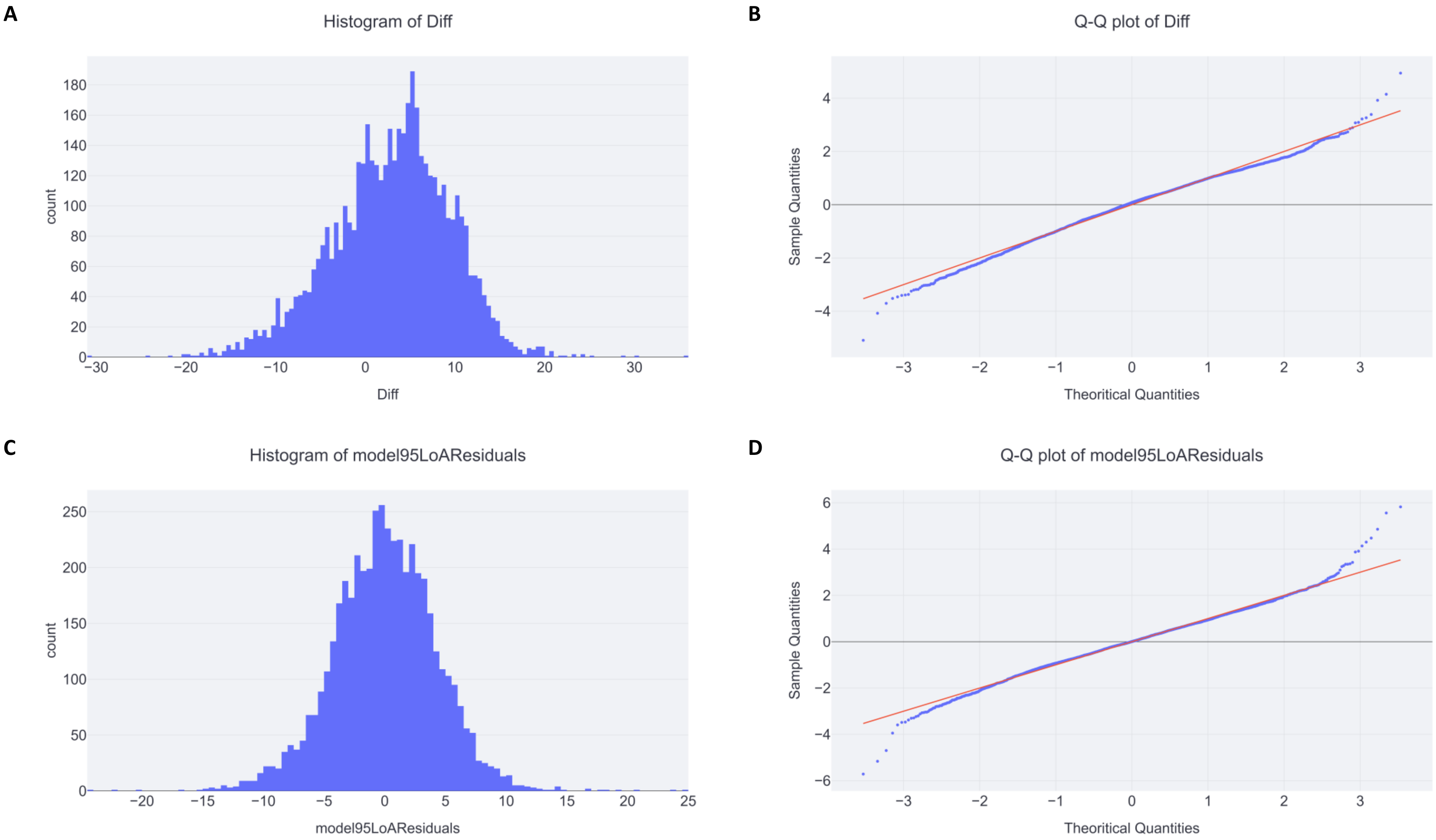


Figure 21. Bland-Altman plots to compare the four LoA analyses based on the static systolic blood pressure reading of the dataset outlined in section 4.2.4 and 5.2.1. A) Classic, B) Repeated measurements, C) Mixed-effect, D) Regression of difference LoA analyses. Heatmap (in blue) to show the density of measurements in different regions to prevent overplotting.


A
Histogram of Diff
count
180
160
140
120
100
80
60
40
20
0
−30
−20
−10
0
10
20
30
Diff
B
Q-Q plot of Diff
Sample Quantities
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Theoritical Quantities
C
Histogram of model95LoAResiduals
count
250
200
150
100
50
0
−20
−15
−10
−5
0
5
10
15
20
25
model95LoAResiduals
D
Q-Q plot of model95LoAResiduals
Sample Quantities
6
4
2
0
−2
−4
−6
−3
−2
−1
0
1
2
3
Theoritical Quantities

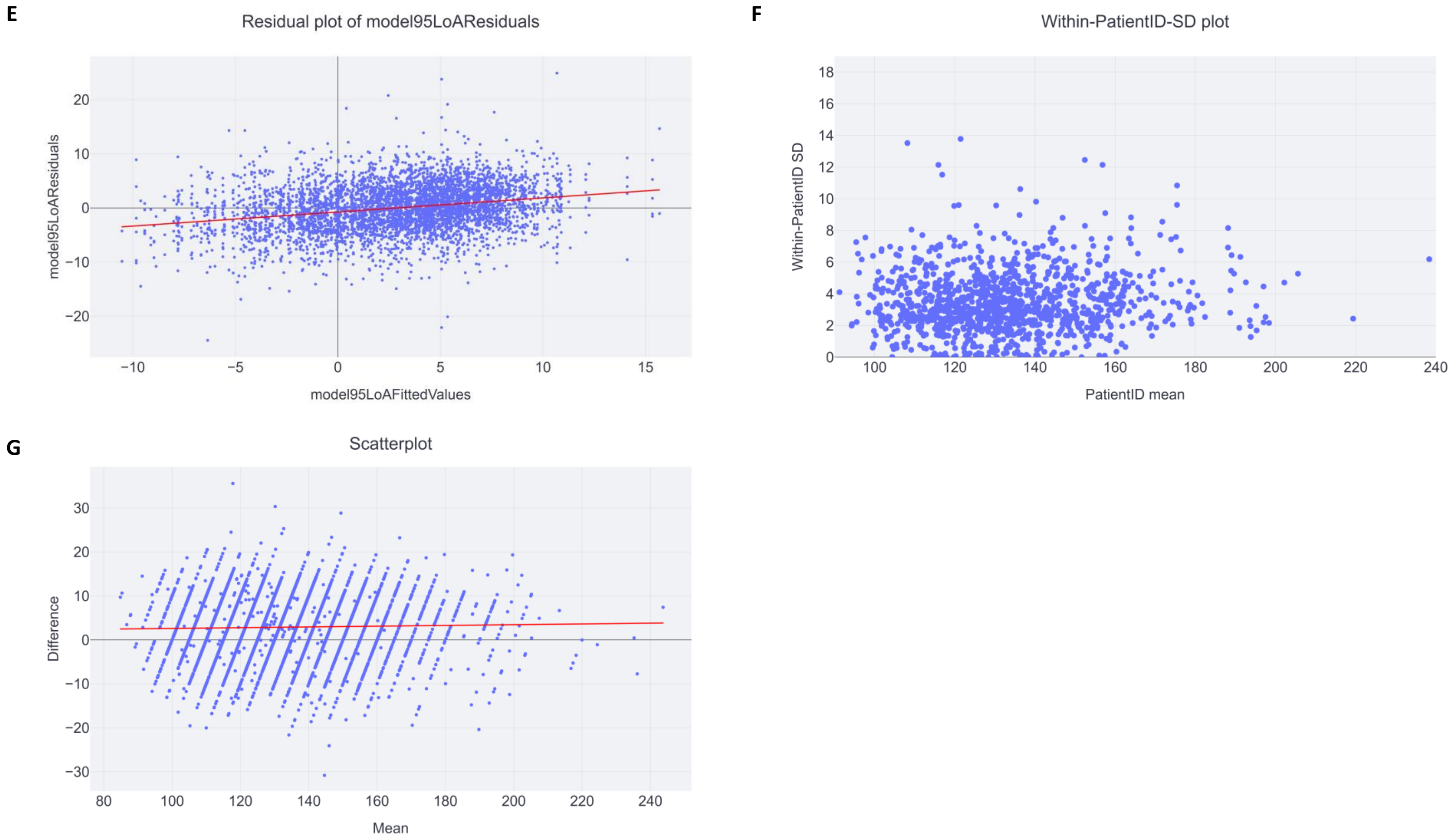
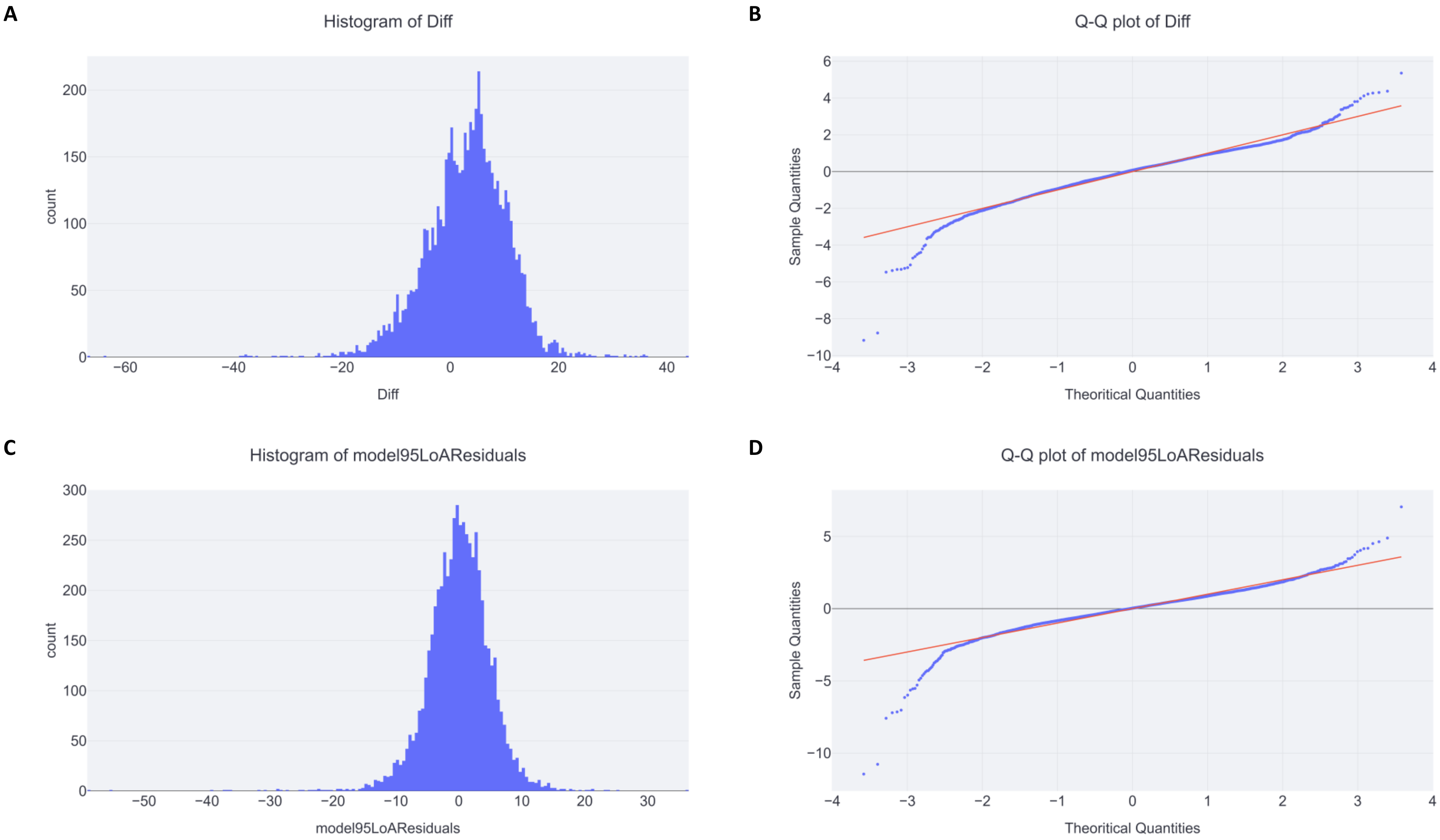


Figure 22. Statistical assumptions of the mixed-effect LoA analyses in the static systolic blood pressure measurements of the dataset outlined in section 4.2.4 and 5.2.1. A) Histogram of difference. B) Q-Q plot of the difference. C) Histogram of residuals of the bias model. D) Q-Q plot of the residuals of the bias model. E) Residual plot of the 95% LoA-model. F) Within-subject standard deviation plot. G) Scatterplot.


A
Histogram of Diff
count
200
150
100
50
0
−60
−40
−20
0
20
40
Diff
B
Q-Q plot of Diff
Sample Quantities
6
4
2
0
−2
−4
−6
−8
−10
−4
−3
−2
−1
0
1
2
3
4
Theoritical Quantities
C
Histogram of model95LoAResiduals
count
300
250
200
150
100
50
0
−50
−40
−30
−20
−10
0
10
20
30
model95LoAResiduals
D
Q-Q plot of model95LoAResiduals
Sample Quantities
5
0
−5
−10
−4
−3
−2
−1
0
1
2
3
4
Theoritical Quantities

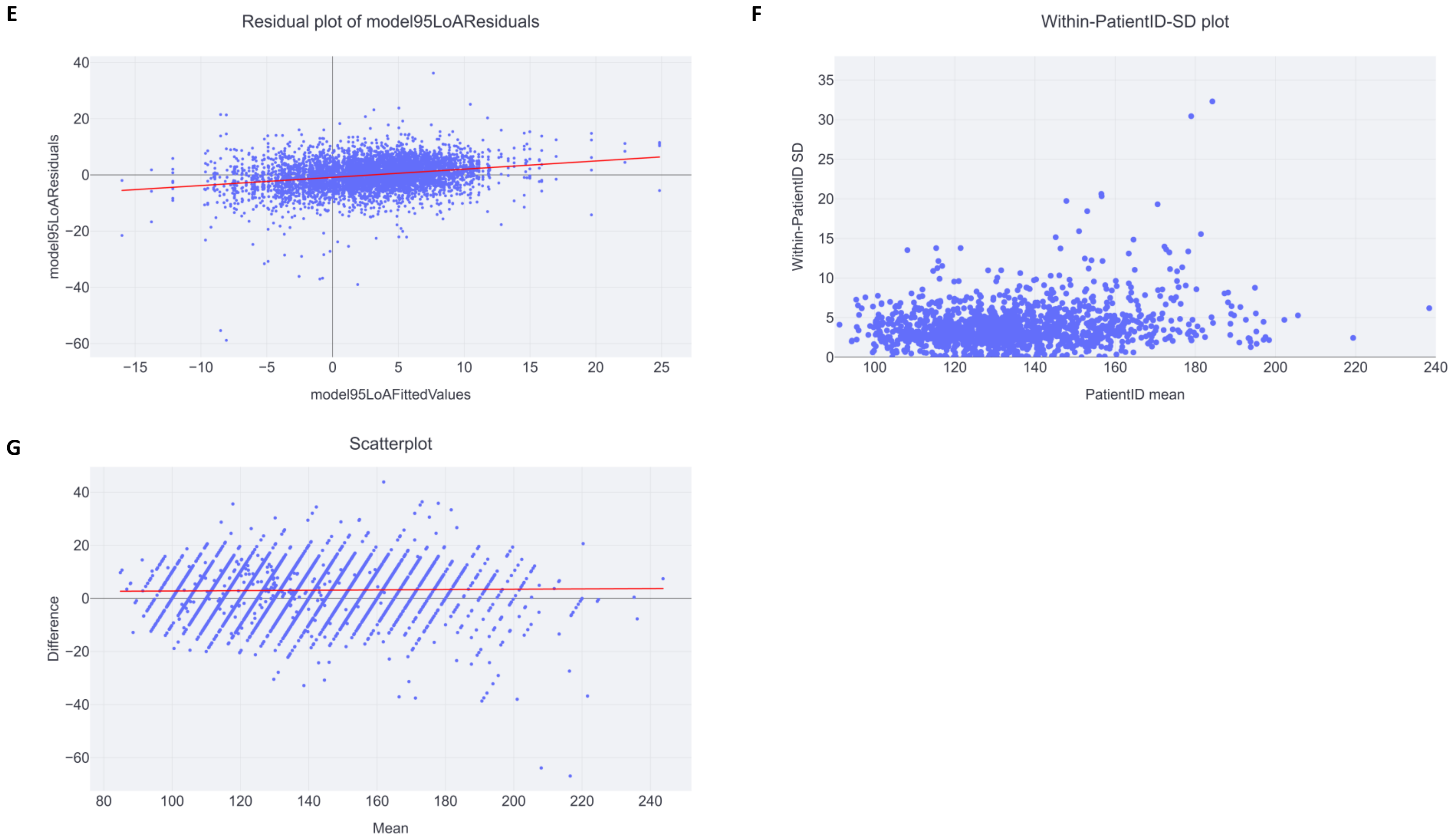


Figure 23. Statistical assumptions of the mixed-effect LoA analysis in the induced systolic blood pressure measurements of the dataset outlined in section 5.2.1. A) Histogram of difference. B) Q-Q plot of the difference. C) Histogram of residuals of the bias model. D) Q-Q plot of the residuals of the bias model. E) Residual plot of the 95% LoA-model. F) Within-subject standard deviation plot. G) Scatterplot.

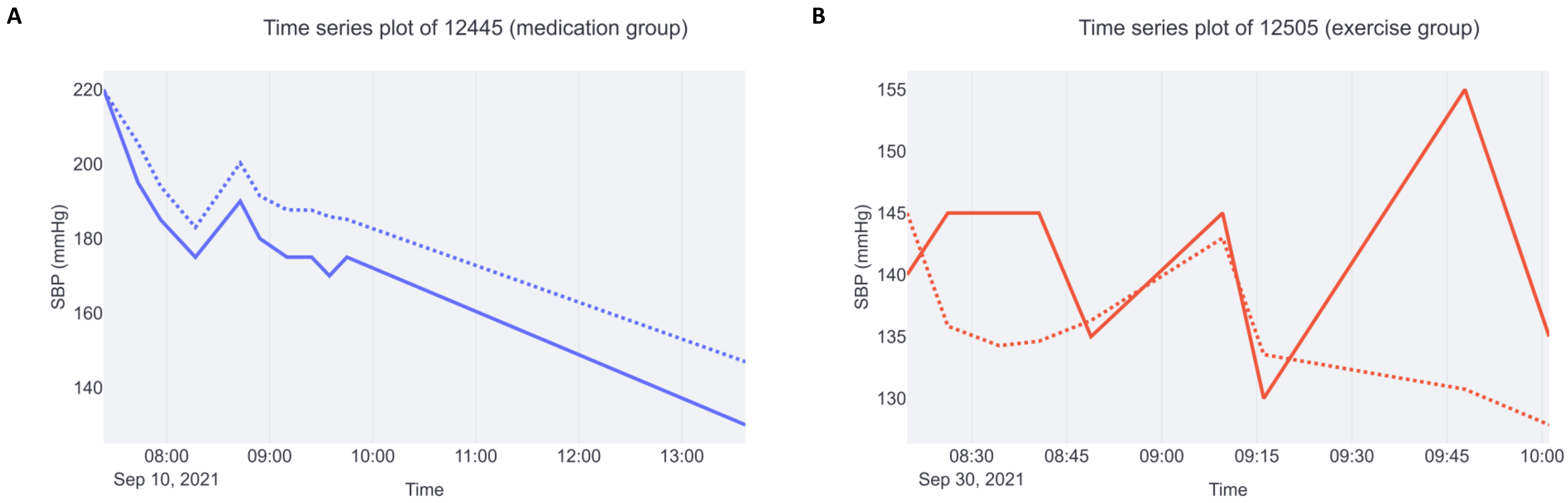


Figure 24. Time series of the induced systolic blood pressure measurements. The CPC device (dotted line) fails to detect the increasing inaccuracy over time compared to the reference device (solid line). The device was calibrated around the start of the measurements. A) subject 12445 of the medication group, B) subject 12505 of the exercise group.

G. VALIDSENSE PYTHON PACKAGE

We developed the *ValidSense.py* package, which is available (https://github.com/petervtooster/ValidSense). This Python package is used to build up the toolbox in the streamlit interface but can also be used for automatic implementation in the user's software. The functions in the Python package are divided into three parts:

1. **Loading**: functions to load (multiple files) to the Streamlit GUI.
2. **Preprocessing**: functions to preprocess the data by standardisation of variable names, conversion to datetime, removal of missing value, and calculation of the difference and mean between the paired measurements.
3. **Analysis**: functions to calculate the four existing LoA analysis statistics (classic, repeated measurements, mixed-effect, regression of difference), longitudinal analysis statistics, visualisations of the statistics (Bland-Altman plot, agreement plot and time series plots), and assessment of the statistical assumptions (histogram, Q-Q plot, residual plot, scatterplot and within-cluster-SD plot).

In addition, we implement the scripts for performing the *IEEE Standard for Wearable, Cuffless Blood Pressure Measuring Devices* [1,2] to the ValidSense toolbox in the *IEEEcufflessBP* folder. The IEEE script is not implemented in the GUI. The script exports the results to an Excel file. Note that there are some deviations from the original protocol.

Table 14. Loading functions in *ValidSense.load*.

Function	Description	Argument	Return	Warnings
upload list to dict	Function to convert streamlit file uploader file to dictionary. Multiple CSV/XLSX files are allowed. Multiple sheets in XLSX are separated.	**upload list**: (list) streamlit file uploader input with accept multiple files=True. **sep**: (str = ';') delimiter to use for pandas read CSV.	(dict) dict with all uploaded files in pd.DataFrame format.	**upload list** is empty list. sep is not str.
add name column to dict	Function to add name as column in dictionary.	**data dict**: (dict) dict with all loaded files in pandas DataFrame format.	(dict) dict with all loaded files in pandas DataFrame format with added name column of file.	**data dict** is not dict.
merge dict to df	Function to merge filtered files in dict to pandas DataFrame.	**data dict**: (dict) all loaded files in dict containing dataframe. **file filter**: (list or None = None) list of names of dataframe to filter.	(pandas DataFrame) combined dataframe with filtered files.	**data dict** is not dict. **file filter** is not list or NoneType.

Table 15. Preprocessing functions in *ValidSense.pre*.

Function	Description	Argument	Return	Warnings
df rename col	Function to rename column name in dataframe from column name old to column name new.	**df**: (pandas DataFrame) dataframe. **column name old**: (str) old column name. **column name new**: (str) new column name.	(pandas DataFrame) dataframe with changed column name.	**df** is not pandas DataFrame. **column name old** is not str. **column name new** is not str. **column name old** exist not in df.
df to datetime	Function to convert column in dataframe to datetime64[ns] format Date and time could be in separate columns or in one column. Column will be renamed to 'Datetime' Format of datetime input can be changed.	**df**: (pandas DataFrame) dataframe to be converted to datetime. **separate datetime**: (bool) True when datetime in separate column. False if datetime is in one column. **datetime**: (str = None) column containing both date and time. **time**: (str = None) column containing time. **date**: (str = None) column containing date. **format strftime**: (str = None) change format input (https://docs.python.org/3/library/datetime.html#strftime-and-strptime-behavior). **datetime unit**: (str = None) unit of datetime (D,s,ms,us,ns) after UNIX epoch start (January 1, 1970, at 00:00:00 UTC").	(pandas DataFrame) dataframe with colum 'Datetime' in format datetime64[ns].	**df** is not pandas DataFrame. **separate datetime** is not bool. **datetime** is not str. **date** is not str. **time** is not str. **format strftime** is not str or NoneType. **datetime unit** is not str or NoneType.
missing	Function to delete rows with missing values (nan), only in the subset of columns. Specific values can also be set to nan and can therefore be deleted.	**df**: (pandas DataFrame) dataframe with missing values. subset col: (list, default None) subset of columns of dataframe where rows with missing values are deleted.	([pandas Dataframe, pandas Dataframe]) returns dataframes with information about missing values.	**df** is not pandas DataFrame. **subset col** is not str or NoneType.
df diff mean	Function to calculate the difference (test - reference) and mean between two devices.	**df**: (pandas DataFrame) dataframe with reference and test device. **test device**: (str = 'Dev1') column name of Test device. **ref device**: (str = 'Dev2') column name of Reference device.	(pandas DataFrame) dataframe with difference and mean added as column.	**df** is not pandas DataFrame. **ref device** is not str. **test device** is not str. **ref device** is not in df.columns. **test device** is not in df.columns. **ref device** contains missing values. **test device** contains missing values.

Table 16. Analysis functions in *ValidSense.analysis*.

Function	Description	Argument	Return	Warnings
loa classic	Function to calculate the bias and limits of agreement statistics according to the classic limits of agreement analysis, see https://pubmed.ncbi.nlm.nih.gov/2868172/.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference.	([pandas DataFrame, str]) dataframe with classic limits of agreement analysis statistics and their assumptions.	**df** is not pandas DataFrame. **'Diff'** is not in df.columns **'Mean'** is not in df.columns. **'Diff'** contains missing values. **'Mean'** contains missing values.
loa repeated measurements	Function to calculate the bias and limits of agreement statistics according to the repeated measurements (multiple observations per subject) limits of agreement analysis. This subtype corrects for multiple observations per subject, see https://pubmed.ncbi.nlm.nih.gov/10501650/ (section 5.2) and https://pubmed.ncbi.nlm.nih.gov/17613642/ (section 3).	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **group by**: (str= 'Sub') column in dataframe where multiple subjects are grouped by.	([pandas DataFrame, str, bioinfokit analys stat]) dataframe with repeated (multiple observations per subject) limits of agreement analysis statistics, assumptions and model.	**df** is not pandas DataFrame. **group by** not str. **'Diff'** is not in df.columns. **'Mean'** is not in df.columns. **group by** not in df.columns. **'Diff'** contains missing values. **'Mean'** contains missing values. **group by** contains missing values.
loa regression of difference	Function to calculate the bias and limits of agreement statistics according to the regression of difference limits of agreement analysis. This subtype corrects for systematic relationship between difference and mean, see https://pubmed.ncbi.nlm.nih.gov/10501650/ (section 3.2).	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **bias order**: (int = 0) order of equation for bias. 0 is horizontal bias, 1 is linear bias. **loa order**: (int = 0) order of equation for limits of agreement. 0 is horizontal limits of agreement, 1 is linear limits of agreement.	([pandas DataFrame, str, statsmodels regression linear model RegressionResultsWrapper, statsmodels regression linear model RegressionResultsWrapper]) dataframe with regression of difference limits of agreement analysis statistics, assumptions and model properties (when bias order, respectively loa order, is set to 1).	**df** is not pandas DataFrame. **bias order** is not int. **loa order** is not int. **'Diff'** is not in df.columns. **'Mean'** is not in df.columns. **'Diff'** contains missing values. **'Mean'** contains missing values.
loa mixed-effect model	Function to calculate the bias, limits of agreement and standard deviation statistics according to the mixed-effect model limits of agreement analysis. This subtype corrects for different fixed and random effects in both bias and limits of agreement, see https://pubmed.ncbi.nlm.nih.gov/27973556/.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **bias fixed variable**: (list) list with fixed effects for bias **bias random variable**: (list) list with random effects for bias **loa fixed variable**: (list) list with fixed effects for loa **loa random variable**: (list) list with random effects for loa	([pandas DataFrame, str, statsmodels regression mixed linear model MixedLMResultsWrapper, statsmodels regression mixed linear model MixedLMResultsWrapper]) dataframe with mixed effect model limits of agreement analysis statistics, their assumptions, and model properties of bias and 95% LoA.	**df** is not pandas DataFrame. **bias fixed variable** is not list. **bias random variable** is not list. **'Diff'** is not in df.columns. **'Mean'** is not in df.columns. **'Diff'** contains missing values. **'Mean'** contains missing values. **bias random variable** is empty. **loa random variable** is empty.
longitudinal analysis	Function to calculate the bias and 95% LoA over time. For every step in window unit in the column col datetime, the bias and 95% LoA are calculated. Rows with time windows where no data is available are dropped. Similar for rows when the max of df[col datetime] exceeds the window size.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. window unit: (str) window unit in days (D), hours (h) or minutes(m). **window size**: (int) window size. **col datetime**: (str = 'Datetime') column containing both date and time. **loa subtype**: (str = 'Classic') subtype of the limits of agreement analysis for the time series analysis. **rep group by**: (str = None) if subtype is 'Repeated Measurements': column in dataframe where multiple subjects are grouped by. **mem bias fixed var**: (list = None) if subtype is 'Mixed-effect Model': list with fixed effects for bias. **mem bias random var**: (list = None) if subtype is 'Mixed-effect Model': list with random effects for bias. **mem loa fixed var**: (list = None) if subtype is 'Mixed-effect Model': list with fixed effects for loa.	([pandas DataFrame, str, statsmodels regression mixed linear model, statsmodels regression mixed linear model, bioinfokit analys stat]) dataframe with limits of agreement variant statistics, assumptions and model.	**df** is not pandas DataFrame. **window unit** is not str. **window size** is not int. **col datetime** is not str. **loa subtype** is not str. **col datetime** is not in df.columns. **'Diff'** is not in df.columns. **'Mean'** is not in df.columns. **Loa subtype** is not in ['Classic', 'Repeated Measurements', 'Mixed-effect Model']. **'Diff'** contains missing values. **'Mean'** contains missing values. **window size** is not positive number. **rep group by** is None. **rep group by** is not str.

		mem loa random var: (list = None) if subtype is 'Mixed-effect Model': list with random effects for loa		**rep group by** contains missing values. **mem bias fixed var** is not list. **mem bias random var** is not list. **mem loa fixed var** is not list. **mem loa random var** is not list. **mem bias random var** is zero. **mem loa random var** is zero.
df add model fits residuals	Function to add model fits and residuals as column to df, with 'name' added in columnname.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **model**: (statsmodels regression mixed linear model MixedLMResultsWrapper) Statsmodel. name: (str) name of the model.	(pandas DataFrame) dataframe with added fittedvalues and residuals columns.	**df** is not pandas DataFrame. **name** is not str. **model** is None.
extract df bias loa	Function to extract bias and 95% LoA from df bias loa time according to time start. Moreover, filter df based on time start column in df.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **df bias loa time**: (pandas DataFrame) dataframe bias and limits of agreement for every step. **time start**: (pandas Timestamp) timestamp to extract.	([pandas DataFrame, str]) dataframe with statistics of the Longitudinal Analysis.	**df** is not pandas DataFrame. **df bias loa** is not pandas DataFrame. **time start** is not pandas Timestamp. **df bias loa time** is empty.
fig bland altman plot	Function to make the Bland-Altman plot, based on the statistics in df or df bias loa.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **df bias loa**: (pandas DataFrame) dataframe with bias and limits of agreement statistics. **x**: (str = 'Mean') x-axis. **y**: (str = 'Diff') y-axis. **group color**: (str = None) column in df to group data by. **title text**: (str = 'Bland-Altman plot') title of the Bland-Altman plot. **heatmap**: (bool = False) if true, heatmap is showed, otherwise the scatterplot. **heatmap nbins** (int = None) number of bins in the heatmap. **marginal** (str = None) marginal subplots of horizontal and vertical axis if heatmap is False. **unit** (str = 'mmHg) unit for x and y label.	(plotly graph objs figure Figure) Bland-Altman plot figure.	**df** is not pandas DataFrame. **df bias loa** is not pandas DataFrame. **x** is not str. **y** is not str. **group color** is not str or NoneType. heatmap is not bool. **heatmap bins** is not int or NoneType. **marginal** is not str or NoneType. **unit** is not str. **x** is not in df.columns. **y** is not in df.columns. **group color** is not None and not in df.columns. **heatmap bins** is not None and heatmap bins is not positive number. **marginal** is not in [None, ‘rug’, ‘box’, ‘violin’, ‘histogram’].
fig agreement plot	Function to make the agreement plot.	**df bias loa time**: (pandas DataFrame) dataframe with bias and limits of agreement statistics. **title text**: (str = ‘Agreement plot) title of the Agreement plot.	(plotly graph objs figure Figure) agreement plot figure.	**df bias loa time** is not pandas DataFrame
fig time series plot	Function to make a time series plot scatterplot with trendlines	**df**: (pandas DataFrame) dataframe with all measurements **x**: (str) x-axis indicating time. **y1**: (str = 'Diff') y-axis Dev1. **y2**: (str = 'Diff') y-axis Dev2. **group color**: (str = None) column in df to group data by. **title text**: (str = ‘Time series individual subjects') title of figure. **show dev1**: (bool = True) show scatter of dev1. **show dev1**: (bool = True) show scatter of dev1.	(plotly graph objs figure Figure) time series figure of individual subjects with moving (median) average.	**df** is not pandas DataFrame. **x** is not str. **y1** is not str. **y2** is not str. **group color** is not str. unit is not str. **window size trendline** is int.

		show dev1 trend: (bool = True) show trendline of dev1. **show dev1 trend**: (bool = True) show trendline of dev1. **window size trendline**: (int = 30) window size of trendline moving (median) average		**window size trendline** is not positive number. **show dev1** is not bool. **show dev2** is not bool. **show dev1 trend** is not bool. **show dev2 trend** is not bool. **x** is not in df.columns. **y1** is not in df.columns. **y2** is not in df.columns. **group color** is not in df.columns. **entry bp** is not in df.columns.
fig histogram	Function to visualize the distribution of column in a histogram figure.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **column**: (str = 'Diff') extract the distribution of this column. **number bins**: (int = 0) number of bins of histogram, if 0, plotly.express.histogram automatically define the number of bins.	(plotly graph objs figure Figure) histogram figure.	**df** is not pandas DataFrame. **column** is not str. **number bins** is not int. **column** is not in df.columns. **number bins** is negative number. **number bins** exceed len(df).
fig qq plot	Function to create a probability distributions by plotting their quantiles against each other.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **column**: (str = 'Diff') extract the distribution of this column. **line**: (str or None = 's') options for the reference line to which the data is compared: “45” - 45-degree line, “s” - standardized line, the expected order statistics are scaled by the standard deviation of the given sample and have the mean added to them, “r” - A regression line is fit, “q” - A line is fit through the quartiles, None - by default no reference line is added to the plot. **fit**: (bool = True) the quantiles are formed from the standardized data.	(plotly graph objs figure Figure) Q-Q plot figure.	**df** is not pandas DataFrame. **column** is not str. **line** is not str or NoneType. **fit** is not bool. **column** is not in df.columns. **line** is not in [None, ’45 ‘, ‘s’, ‘r’, ‘q’].
fig residual plot	Function to create residual plot, with fitted values on the x-axis, residuals on the y-axis.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **fits**: (str) fitted values of the regression model or mixed-effect model. **residuals**: (str) residuals of the regression model or mixed-effect model.	(plotly graph objs figure Figure) Residual plot with ordinary least squares trendline.	**df** is not pandas DataFrame. **fits** is not str. **residuals** is not str. **fits** is not in df.columns. **residuals** is not in df.columns.
fig scatter plot	Function to create a scatter plot.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **x**: (str) x-axis **y**: (str) y-axis **group color**: (str = None) column in df to group data by. unit (str = 'mmHg) unit for x and y label.	(plotly graph objs figure.Figure) Scatter plot figure.	**df** is not pandas DataFrame. **x** is not str. **y** is not str. **group color** is not str or NoneType. **unit** is not str. **x** is not in df.columns. **y** is not in df.columns. group color is not None and not in df.columns.
fig within cluster SD plot	Function to create the within-cluster standard deviation plot.	**df**: (pandas DataFrame) dataframe with column 'mean' and 'diff', representing the mean and difference. **group**: (str) column in df to group by.	(plotly graph objs figure.Figure) Within-cluster SD plot figure.	**df** is not pandas DataFrame. **group** is not str. **‘Diff’** is not in df.columns. **‘Mean’** is not in df.columns. **group** is not in df.columns. **‘Diff’** contains missing values. **‘Mean’** contains missing values. **group** contains missing values.