Aus dem Institut für Translationale Physiologie
der Medizinischen Fakultät Charité – Universitätsmedizin Berlin

DISSERTATION

The B-Score: Evaluation of blood pressure measurement systems by a novel score for the determination of the true, relative measurement performance

Der B-Score: Ein neuer Score zur Evaluierung der wahren, relativen Leistungsfähigkeit von Blutdruckmesssystemen

zur Erlangung des akademischen Grades
Medical Doctor - Doctor of Philosophy (MD/PhD)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Tomas Lucca Bothe

Datum der Promotion: 29. November 2024

Table of contents

List of figures

List of abbreviations

BP	blood pressure
DBP	diastolic blood pressure
HT	arterial hypertension
RMSE	root mean squared error
B1 RMSE	root mean squared error comparing all measurements to the dataset mean
B2 RMSE	root mean squared error comparing all measurements to the patient’s first measurement
M	standardized (Deep Learning) model
M RMSE	root mean squared error comparing all measurements to the Deep Learning model’s prediction

Abstract

Background: Arterial hypertension (increased arterial blood pressure) is one of the most important predictors of adverse cardiovascular events and frequent cause for medical intervention. There is a steadily increasing number blood pressure monitors available, based on conventional or novel technical approaches. However, the large heterogeneity of validation studies and lack of an easily understandable metric of true measurement performance poses a grave issue for the reliable evaluation of blood pressure monitors.
Objective: It was our goal to create a novel, easily interpretable, and accessible metric for the true measurement performance of blood pressure monitors: The B-Score.
Methods: We designed the B-Score to compare the absolute performance of a blood pressure monitor with the difficulty (e.g., variability) of the dataset it was tested upon. This creates a metric of relative performance, directly comparably to B-Scores calculated on other devices. Following its design, we tested the B-Score on a variety of simulated and real-world datasets to assess it for its mathematical properties, as well as interpretability and real-world applicability.
Results: The B-Score proved mathematically predictable behaviour and strong discrimination between different performing blood pressure measurement systems when tested on simulated data. Further, we were able to show that the B-Score can be easily calculated for challenging real-world data and provides important and intuitively understandable insights.
Conclusion: The B-Score is a novel, powerful tool for the evaluation of blood pressure measurement systems. It allows the direct comparison of different blood pressure monitors, even if tested on heterogenous data.

Zusammenfassung

Hintergrund: Die arterielle Hypertension (erhöhter arterieller Blutdruck) ist einer der bedeutsamsten Prädiktoren für adverse kardiovaskuläre Ereignisse und ist einer der häufigsten Initialgeber für eine pharmakologische Intervention. Die Menge der Verfügbaren Blutdruckmesssysteme wächst stetig – basierend auf sowohl konventionellen als auch neuen Messtechniken. Dabei stellt die große Heterogenität zwischen den Validierungsstudien und das Fehlen einer einfachen Metrik für die wahre Messgenauigkeit ein großes Hindernis für die Bewertung der vorhandenen Systeme dar.

Zielsetzung: Es war unser Ziel eine neuartige, einfach zu interpretierende und zugängliche Metrik zu entwickeln, um die wahre Messgenauigkeit von Blutdruckmessgenauigkeiten abzubilden: Den B-Score.

Methoden: Der B-Score vergleicht die absolute Messungenauigkeit eines Geräts mit der Schwierigkeit (z.B. Variabilität) der Daten, gegen die es getestet wurde. Das Ergebnis ist eine Metrik, die die relative Genauigkeit eines Geräts angibt und direkt mit den Ergebnissen anderer Geräte vergleichbar ist. Im Anschluss an die Entwicklung haben wir den B-Score an einer Vielzahl simulierter und echter Datensätze getestet, um das mathematische Verhalten, die Interpretierbarkeit und die Anwendbarkeit in der echten Welt zu testen.

Ergebnisse: Unsere Tests des B-Scores mit simulierten Daten zeigten ein mathematisch erwartbares Verhalten sowie eine starke Unterscheidung zwischen unterschiedlich genauen Messsystemen. Weiterhin konnten wir zeigen, dass der B-Score auch für die Berechnung mit Echtweltdaten geeignet ist und dabei wichtige und intuitiv zu interpretierende Ergebnisse liefert.

Schlussfolgerung: Der B-Score ist eine neue, leistungsstarke Metrik für die Bewertung von Blutdruckmesssystemen. Er ermöglicht den direkten Vergleich verschiedener Systeme, selbst wenn mit unterschiedlichen Daten getestet wurden.

1 Introduction

1.1 Modern medicine and the importance of arterial hypertension

Modern medicine is diverse and highly specialized. Advances in almost all fields of medicine over the last decades have led to the emergence of ever more mature, effective, and increasingly personalized diagnoses and treatments.[1–3] At the same time, in clinical practice, many diagnoses and treatment decisions are based upon well-known, tried-and-trusted examinations.

Arterial hypertension (HT) is the medical condition leading to the most premature deaths worldwide.[4] The estimated prevalence of HT, defined as a pathologically increased arterial blood pressure (BP), is between 30-45% of the adult western population.[4,5] Subsequently, an antihypertensive pill is the most prescribed medication in the United States, followed by five more agents for mitigating the risk of HT and the associated metabolic syndrome.[6] Followingly, assessing patients' BP levels correctly is not only a frequent but extremely important task in everyday medicine.

1.2 Current hypertension diagnostics and treatment management

The American and European Societies for Hypertension have been publishing and updating guidelines for the correct assessment auf HT. These guidelines uniformly recommend the measurement of brachial BP with a validated cuff-based device, either via auscultation or automated. The gold standard for HT diagnostics and treatment monitoring is the 24-hour, ambulatory BP measurement, performed via an automated, commonly oscillometric device.[7–9] This is reasonable, as the 24-hour and especially the nocturnal BP have been identified as the most predictive marker for cardiovascular events and mortality.[10–12]

1.3 The disadvantages of automated, cuff-based devices for blood pressure measurement

Unfortunately, there are drawbacks to cuff-based BP measurement devices. The cuff-based measurement paradigm is dependent on the intermitted in- and subsequent deflation of the cuff to determine the BP, which leads to multiple disadvantages:

1. Discontinuity: Cuff-based BP measurement can only provide one measurement per inflation, typically ever 15-30 minutes. Therefore, short-term alterations of the BP are only detected by chance and can most likely not be adequately interpreted.[5]
2. Patient discomfort and sleep impairment: The repeated cuff inflations can be perceived as disturbing and sometimes painful. Especially during the night, this can lead to arousal reactions which themselves influence the BP level. This limits the reliability of nocturnal BP measurements.[13–15]
3. Measurement artefacts: Cuff-based BP measurement is prone to measurement artefacts. Movements during the measurement process or arrhythmic events can lead to errors in the BP determination and therefore greatly influence the measurement results.[16–18]
4. Insufficient validation: Cuff-based devices have been used for 24-hour, ambulatory BP measurement for decades. However, these devices are validated in a short-term, laboratory setting in seated subjects at total rest. There is no widely accepted 24-hour validation protocol, nor a reliable estimation of the measurement accuracy of cuff-based devices.[19] Worryingly, there are investigations showing the limited reproducibility of ambulatory BP measurement results.[20,21]

1.4 Cuff-less blood pressure measurement as an alternative?

The described limitations of the cuff-based technique have led to a growing interest in alternative methodologies for BP measurement. These new devices, most commonly based on the correlation between the BP and surrogate parameters of vessel stiffness (e.g., pulse-wave-velocity), are designed to measure the BP non-invasively, continuously and without the drawbacks of repeated cuff inflations.[5] However, as for the cuff-based devices, there is no agreed upon validation protocol. Further, unlike for the cuff-based devices, there is no decade-long clinical experience, which is leading to a fair bit of scepticism towards these new devices. Consequently, the European Society of Hypertension has stated that cuff-less BP measurement devices are a promising development but there is as of now not enough clinical evidence to support its broad clinical application.[9,22]

1.5 The issue of evaluating blood pressure measurement devices

As of now, evaluating the performance of BP measurement devices is difficult. There is a plethora of devices in development, most of which by scientist. In an ideal world, there would be clinical outcome studies for all these devices. However, such studies consist of thousands of patients, have years of follow-up time and cost millions.

As clinical outcome studies are unfeasible, researchers had to find alternatives. Therefore, proposed devices are tested under very heterogenous conditions, with inconsistent cohort characteristics and in a multitude of scenarios. The results of such studies are mostly reported as measurement error between the proposed and a reference device (e.g., a cuff-less (test) and a cuff-based (reference) device), mostly in form of the mean deviation or mean absolute error. This has led to a situation in which there is a multitude of studies which are impossible to compare:

A study might show a given measurement error for a device (A) tested in young subjects who were put under physical load on a bike ergometer – a highly dynamic scenario with large BP fluctuations. Another device (B) might show the same measurement error but was tested on middle-aged subjects during short-term study at total rest (e.g., validation protocol for cuff-based devices). Device (C) may yet again show the same mean absolute error but was tested in a 24-hour ambulatory BP measurement setting in children and adolescents. Device (D) could have been tested on an intensive care unit, with an intraarterial BP measurement as reference. Even though devices (A) – (D) might show the same absolute measurement error, it would be a fallacy to conclude that all devices are equally good at measuring BP.

As of now there is no way of comparing the true performance of BP measurement devices which is why it is impossible to provide the evidence demanded by the European Society of Hypertension to promote continuous BP measurement into broad clinical application.

1.6 The B-Score: Evaluating the true measurement performance of blood pressure measurement devices

My work focusses on solving the described problem and developing a way of enabling the comparison of true measurement performances across different devices and studies. To do so, it was the goal of our working group and me personally to develop a score

which depicts the true performance of a given device and which any researcher can easily calculate and report for their own data.

For this, we created a measure which sets the absolute measurement error of a device (measurement error between tested and reference device) in contrast to dataset characteristics, such as inter- and intraindividual dataset variability and overall predictive difficulty (assessed via a standardized Deep Leaning application) of the dataset: The B-Score.[23]

Here, I describe the rationale behind the B-Score, its development, mathematical and computational properties, its testing on simulated and real-world data and its possible future applications.[23]

2 Methods

2.1 Conceptualizing the B-Score

2.1.1 A measure of relative blood pressure measurement performance

Quantifying the true performance of BP measurement devices is difficult. Given that devices can be based on very different physiological and/or mathematical approaches and that they are developed and tested by various research groups, there is no chance of directly comparing all interesting devices in one large investigation.
Comparing the reported measures of absolute measurement errors (e.g., mean absolute error, standard deviation) is equally futile, as the devices are tested on very different datasets.
To solve this problem, we decided to develop a measure of relative measurement performance. By setting the absolute measurement errors of any given device in relation to the difficulty the dataset is tested upon, we were able to make devices comparable, even if tested on very different datasets. The B-Score is designed to be calculated independently for systolic and diastolic BP values and therefore allow separate interpretation of a device's systolic and diastolic measurement performance.[23]

2.1.2 The root mean squared error as the foundation of the B-Score

Relative scores are created by setting absolute measures in contrast to each other. Therefore, choosing the right metric of absolute measurement error was fundamental for the B-Score's reliability and meaningfulness.
We decided on the root mean squared error (RMSE) as our fundamental measure of absolute measurement performance. The RMSE is a well-established and widely adopted metric, which possesses properties especially desirable for evaluating BP measurement devices.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(prediction - reference\ value)^2}$$

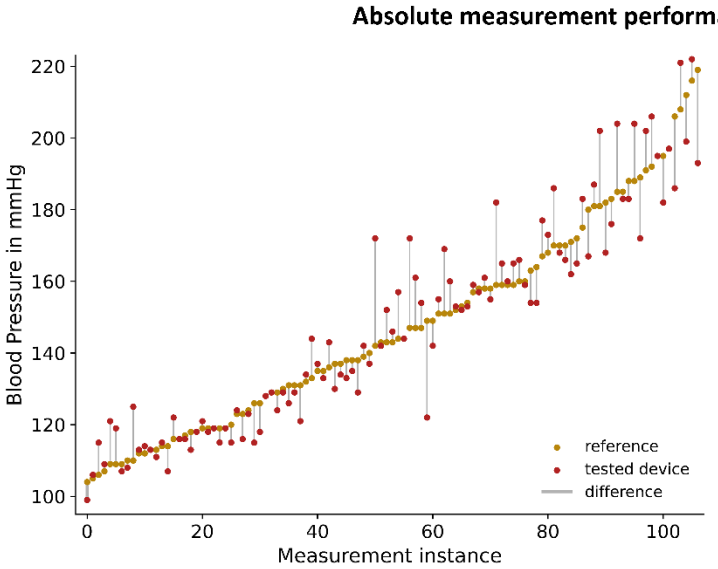
n = number of samples in the dataset.

We chose the RMSE because in BP measurement there is no linear relationship between a measurement error and the gravity of the mistake. A measurement error of 10 mmHg

is not only twice a measurement error of 5 mmHg but much more harmful, as enlarged measurement errors can and will likely lead to bad treatment decisions and subsequently unnecessary harm done to patients. The RMSE reflects this consideration, as it penalizes larger measurement errors more rigorously than other metrics such as the mean absolute error.

2.1.3 Base performances for dataset characterization

After choosing the RMSE as fundamental metric for the B-Score calculation, we had to design dataset dependent RMSE values to which we could set in relation to the absolute model performance of the tested model as RMSE (T-RMSE). (Figure 1)



$$difference = (BP_{reference} - BP_{tested\ device})$$

$$T\ RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(difference_{\ in\ mmHg})^2}$$

Figure 1: Absolute measurement performance (T-RMSE): The calculation of the T-RMSE is graphically displayed. On the left-hand side of the figure shows the measurement differences between a reference and a tested device for systolic BP values. The datapoints are ordered by increasing reference The grey, vertical lines indicate the disagreement for each individual measurement. On the right-hand side of the figure displays the mathematical formulas for calculating the T-RMSE. The figure depicts original data from the “dobutamine dataset” as published in the original B-Score article.[23] Source: Own illustration.

We named these dataset dependent RMSE values “*base performances”* of the tested dataset. The *base performances* are designed to reflect how difficult it is for a tested device to retrieve a high absolute measurement accuracy on the used dataset. To achieve this, the base performances are reflective of the inter- and intraindividual BP variability in the dataset as well as the overall dataset difficulty, which we assessed by a standardized Deep Learning application. (Figure 2)

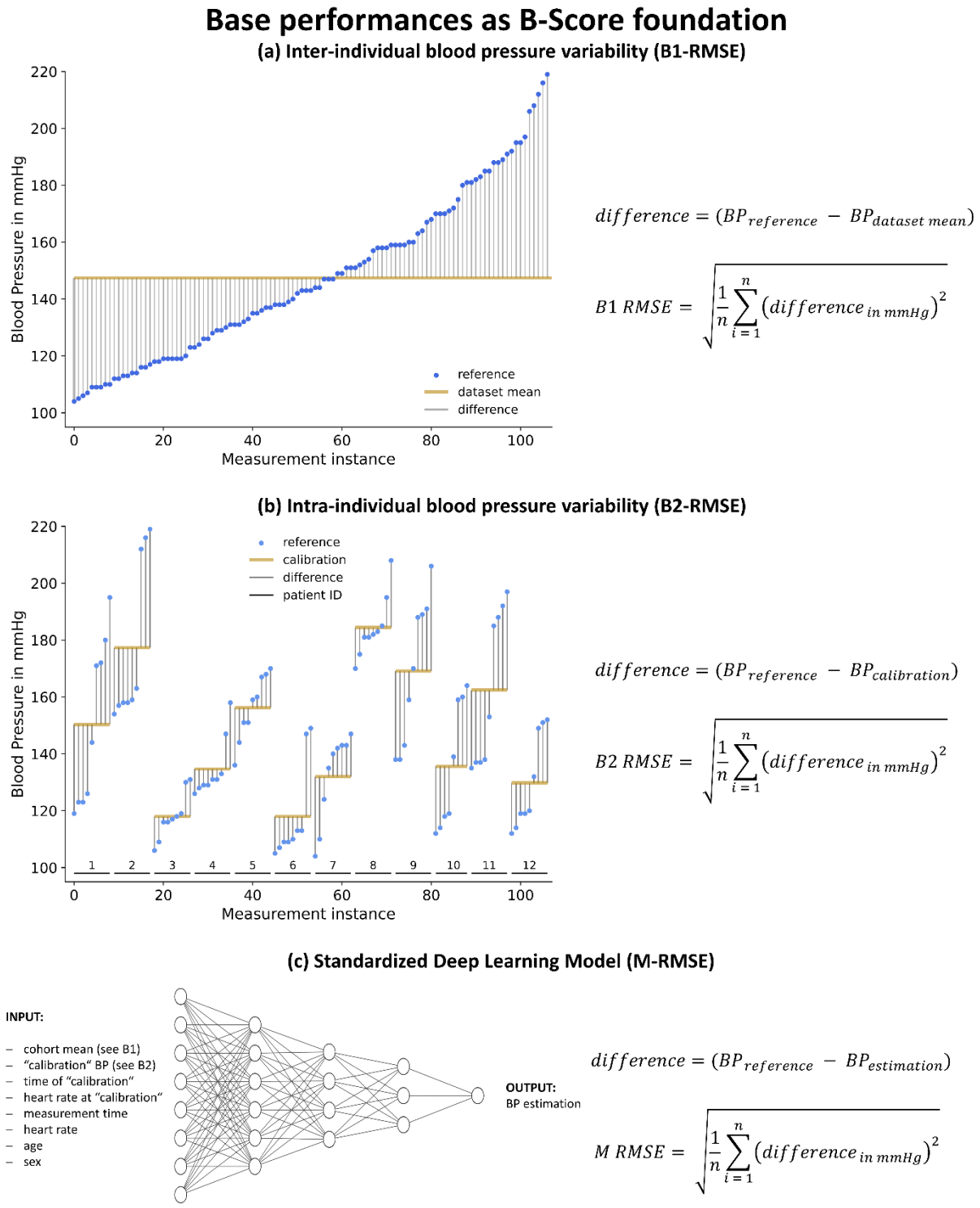


Figure 2: Base performances as B-Score foundation: The figure displays the calculation for all three *base performances*. Panel (a) depicts the calculation of the B1-RMSE, a measure of inter-individual BP variability. The shown data is ordered by increasing reference BP values. Panel (b) depicts the calculation of the B2-RMSE, a measure of the intra-individual BP variability. The data is ordered by subjects and increasing reference values. The grey, vertical lines indicate the disagreement for each individual measurement in (a) and (b). Panel (c) depicts the calculation of the

M-RMSE, a measure of the overall difficulty of creating a BP estimation model for the given dataset. A graphical representation of the standardized Deep Learning model is shown on the left-hand side. All panels (a) – (c) provide the mathematical formulas for calculating the respective *base performance.* Panels (a) and (b) depict original data from the “dobutamine dataset” as published in the original B-Score article.[23] Panel (c) shows an adaptation of a figure first published in the original B-Score article (modified from Bothe et al., 2022).[23] Source: Own illustration.

In total, we created three *base performance* measures, each reflecting one of the described dataset properties:

1. B1-RMSE: A measure of interindividual BP variability. The B1-RMSE is the RMSE value derived from comparing each reference value to the mean BP value of the whole dataset. This equates to a RMSE value a BP measurement device would achieve if it estimated the population mean at every measurement instance.
2. B2-RMSE: A measure of intraindividual BP variability. The B2-RMSE is the RMSE value derived from comparing each patients’ reference value to the first reference BP value of the given patient in the dataset. This equates to a RMSE value a BP measurement device would achieve if it estimated a patient specific calibration value (e.g., office measurement) at every measurement instance.
3. M-RMSE: A measure of how difficult it is to derive a BP model for the given dataset. The M-RMSE is the RMSE value derived from comparing each reference value to the BP output value of a standardized Deep Leaning (model (M)) application. This Deep Leaning model is designed to intake information available in any BP measurement study (heart rate, time of measurement, age, sex and a calibration (first measurement) time, heart rate and BP). This equates to a RMSE value a simplified BP estimation model (e.g., integrable in a smartwatch) would achieve.

These *base performance* RMSE values each calculated and serve as the building blocks for the subsequent B-Score calculation. (Figure 2)

2.1.4 B-Score calculation and characteristics

The B-Score is calculated by setting the absolute measurement performance of a tested device (T-RMSE) in contrast to the *base performances* which provide deep information about the dataset's structure.

We designed the B-Score in a way so that an increased B-Score represents increased predictive performance. Therefore, the B-Score increases with increased *base performances* and a reduced T-RMSE.

$$B\ Score = \log_{10}\left(\sqrt{\left(\frac{B1\ RMSE \cdot M\ RMSE}{T\ RMSE^2}\right) \cdot \left(\frac{B2\ RMSE \cdot M\ RMSE}{T\ RMSE^2}\right)}\right)$$

Per our definition, the B-Score can only be calculated if the T-RMSE is smaller than all three *base performances* (T-RMSE < min(*base performances*)). In any other case, the B-Score should be reported as "B-Score < 0.00".[23]

To ensure reliable and reproducible results, we decided to integrate a process of repeated *base performance* calculations per dataset with re-shuffled and re-sampled versions of the dataset. Afterwards, the results are merged and averaged to guarantee minimal calculation insecurity.

2.2 Testing the B-Score with simulated blood pressure datasets

2.2.1 The rationale for using simulated datasets

It was important to us to show that the B-Score does have the desired mathematical properties before testing it on possibly noisy real-world data. Therefore, we decided to simulate three distinct datasets (each consisting of systolic and diastolic BP values). We did this to "stress-test" the B-Score calculation and provided code with large datasets. Further, as we controlled the datasets, we were able to test whether our assumptions about the B-Score are correct (the B-Score increased with increasing model performance).

2.2.2 Simulated blood pressure datasets

We created the simulated datasets with a self-developed, simplified BP model. Our model comprised the most important BP fluctuations during a 24-hour cycle. Consequently, we were able to model a close approximation of the circadian (24 hours)[24], the Traube-Hering-Meyer rhythms (about 7-10 seconds)[25], as well as overall BP variability (as

added deviation from the mean BP value) and a modelled, device specific measurement uncertainty. (Figure 3)

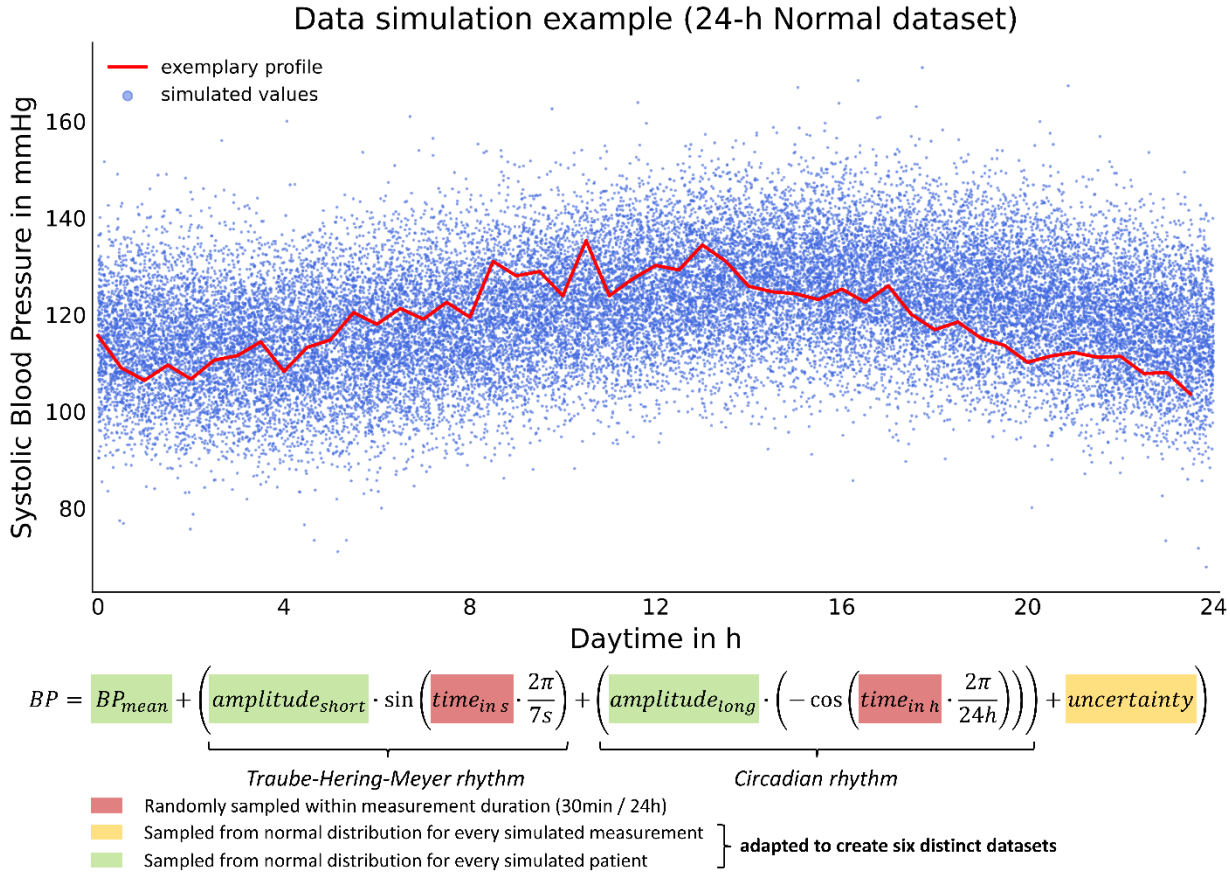


Figure 3: Data simulation example: The top of the figure depicts a subset (30,000 samples) of one of the datasets simulated for testing the B-Score. It shows an adaptation of a figure first published in the original B-Score article (modified from Bothe et al., 2022).[23] The blue points indicate individual BP values over a simulated 24-hour measurement regime. The red line indicates an exemplary BP profile of a single, simulated patient. The bottom of the figure shows the mathematical formula used for simulating the BP datasets. The green, yellow, and red boxes indicate which parameters have been adapted at what stage to simulate the dataset. Source: Own Illustration.

In this way, we were able to create datasets (systolic and diastolic each) modelling three distinct real-word application. We modelled a short-term (30 minutes, ‘Lab’) dataset, representative of a laboratory study and two 24-hour datasets with increasing difficulty (’24-h Normal’, ’24-h Hard’). We modelled the dataset in a way to represent an increasing BP prediction difficulty from ‘Lab’ to ’24-h Normal’ to ’24-h Hard’. Each dataset consisted of 500,000 samples (10,000 patients x 50 BP samples).

We further created additional datasets (50,000 samples) with strictly increasing standard deviations of BP values to test for the predictability of the *base performances* under modulated BP variability.

2.2.3 Testing the B-Score for desired properties

Following the creation of the datasets, we calculated the B-Score for all six datasets to assess whether the B-Score would show the desired behaviour. To do so, we chose an arbitrary T-RMSE value of 4.0 mmHg and validated that the resulting B-Scores increased with increasing dataset difficulty in the 'Lab', '24-Normal' and '24-Hard' datasets.
Further, we analysed the *base performances'* predictability under increasing BP variability. Concludingly, we also analysed the B-Score's behaviour with varying T-RMSE values to confirm a predictable and reliable increase of the B-Score under improving T-RMSE values.

2.3 Testing the B-Score in an extreme real-world environment

2.3.1 Real-world datasets for testing the B-Score

To test the B-Score in a real-world scenario, we decided to use the B-Score in its intended purpose. We therefore chose an already published study of a continuous BP measurement device and calculated the B-Score for it. Subsequently, we calculated what performance an alternative BP measurement device would need to achieve on an already present dataset to be able to claim coequal measurement performance.
The B-Score allows this analysis by rearranging the B-Score equation[23]:

$$T\ RMSE = \sqrt[4]{\frac{B1\ RMSE \cdot B2\ RMSE \cdot M\ RMSE^2}{10^{(2 \cdot B\ Score)}}}$$

Since we aimed at testing the B-Score in the most rigorous way possible, we made the decision to compare two very different datasets, both at the very extreme spectrum of available datasets.
As published validation study, we chose a study consisting of only 107 individual BP measurements (twelve subjects), comparing a continuous BP measurement device with an intraarterial (invasive) BP measurement during a dobutamine stress test.[26] The dataset was generated in a short-term and highly dynamic BP environment and called the *"dobutamine"* dataset in the original B-Score publication.[23]

To allow the most extreme comparison possible, we used the *MIMIC IV* clinical dataset, which consists of millions of datapoints from ICU patients, gathered in North America.[27,28]. The *MIMIC IV* dataset is the, to our knowledge, largest available dataset for real-world, clinical BP data. After pre-processing and cleaning the dataset, we retrieved a systolic and diastolic dataset with more than 2.3 million individual BP entries each.

2.3.2 Testing the B-Score on real-world data

We calculated the B-Score for the *dobutamine* dataset, by calculating the *base performances* and the T-RMSE value for the tested device from the original data. Subsequently, we interpreted the results from the tested device and analysed the *base performance* to gain further insights into the device's performance.
To complete our real-world B-Score test, we calculated the *base performances* for the *MIMIC IV* dataset. Followingly, we used the rearranged B-Score formula to calculate the T-RMSE a fictional BP estimation device or model would need to reach on the *MIMIC IV* database to be considered of coequal predictive performance to the device tested on the *dobutamine* dataset.

2.3.3 Analysing the time complexity of the B-Score calculation

As it was of utmost importance to us that the B-Score is not only providing an intuitively interpretable measure of relative model performance but that it is easily and quickly for researchers. Followingly, we assessed the B-Scores computational performance by measuring the time needed for calculation on increasingly larger subsets of the *MIMIC IV* dataset.

2.3.4 Code availability and programming architecture of the B-Score

We wrote the B-Score's code in Python 3, using a multitude of data scientifical, statistical, graphical and Deep Learning libraries.[29–33] To develop the B-Score as a tool available to all interested researchers, it was at all timepoint throughout the design prospect our goal to make the B-Score code publicly available. Therefore, we streamlined the code into an easily applicable script which is available under a GNU General Public Licence (v3.0) on GitHub.[23]

3. Results

3.1 The B-Score's mathematical behavior tested with simulated datasets

3.1.1. Characteristics under increasing standard deviation of blood pressure

Testing the *base performances* as well as the B-Score with subsets of the *MIMIC IV* dataset revealed a high stability and interpretability under increasing standard deviations of BP. Accordingly, all three *base performances* increased under increasing BP variability. In addition to that, we were able to show that the B-Score increases with increasing *base performance* RMSE values. Finalizing the analysis of the B-Score's mathematical predictability, we were able to show that the B-Score changes with changing T-RMSE values. The B-Score rose for smaller and fell for larger T-RMSE values. (Figure 4)

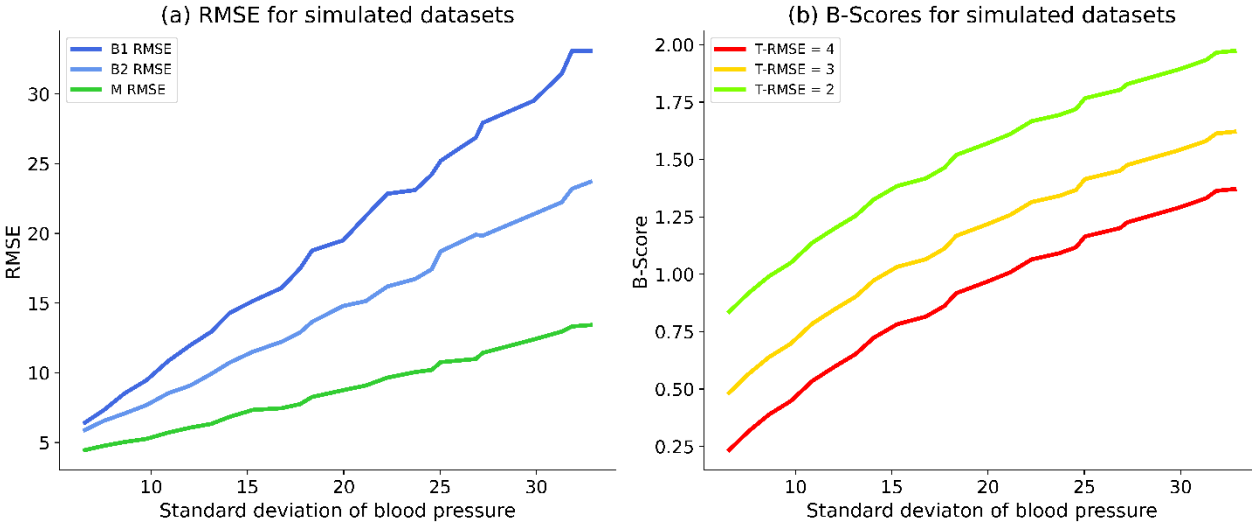


Figure 4: Predictability of the B-Score's behavior: Panel (a) shows the behavior of the three *base performance* RMSE values when calculated for simulated datasets with increasing standard deviation. All three RMSE values increase strictly under increasing blood pressure variability. Panel (b) shows the B-Score under increasing blood pressure variability. Following the *base performance* measures, the B-Score increases with increasing blood pressure variability (dataset difficulty). Further, three different T-RMSE values are depicted. The B-Score increases with a decreasing T-RMSE value. Source: Modified from Bothe et al. (2022).[23]

3.1.2 Testing the assessment of true measurement performance on simulated datasets

After assessing the B-Score for its desired properties on the smaller, simulated datasets, we continued to test the B-Score on the six large, simulated datasets. We calculated the *base performances* of all six datasets and calculated the B-Score with our arbitrary T-

RMSE value of 4.0 mmHg. The B-Score discriminated between the datasets, increasing with increasing *base performances* and modelled dataset difficulty. (Figure 5)

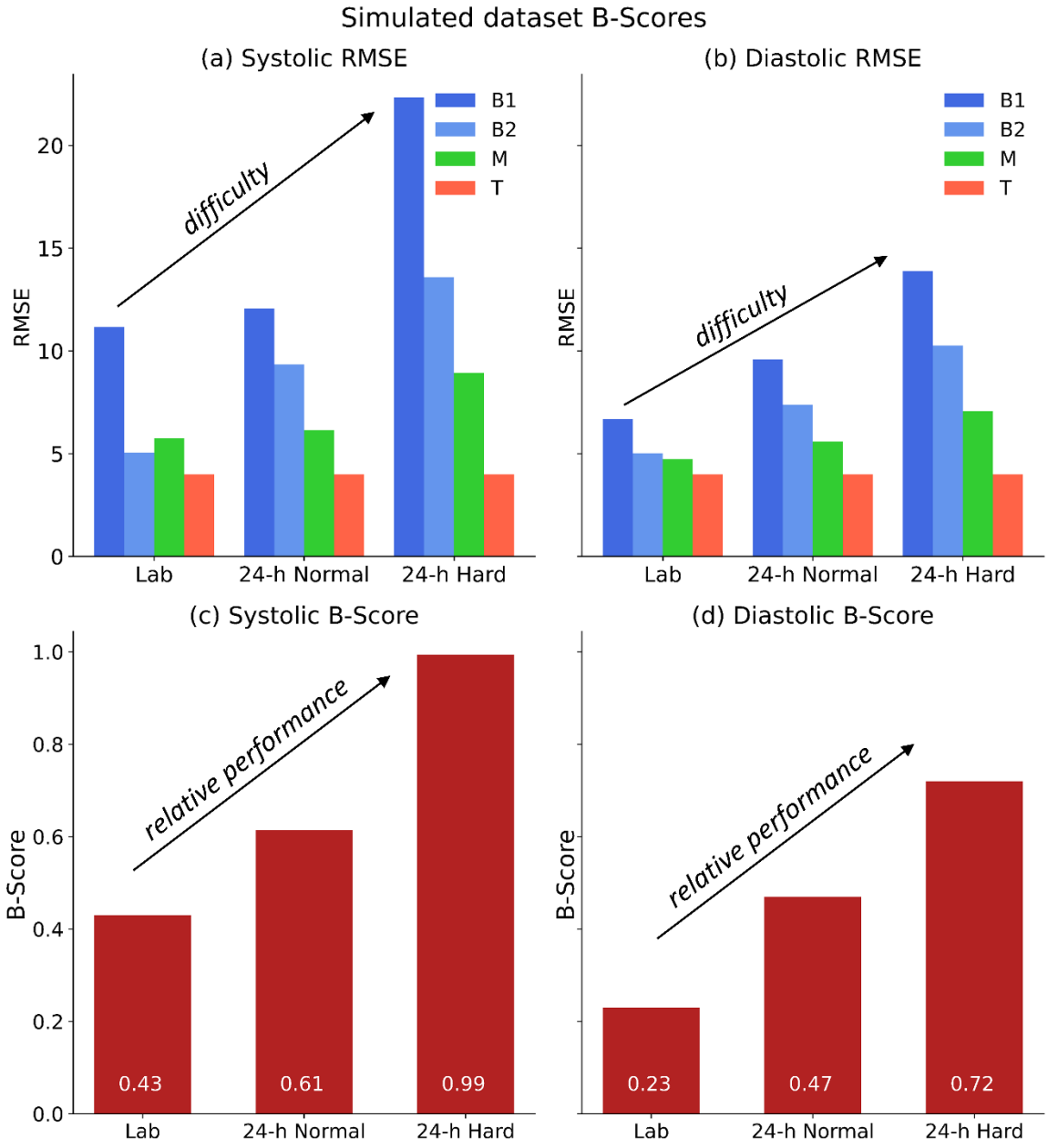


Figure 5: Simulated dataset B-Scores: Panel (a) and (b) show the calculated *base performance* RMSE values for all six simulated datasets. The black arrow indicates the increasing dataset difficulty, depicted by the increasing *base performance* measures. In the second row, panels (c) and (d) depict the increasing B-Score resulting from the *base performances'* reaction to the increased dataset difficulty. The resulting B-Score increase is a direct measure of increased measurement performance, as indicated by the black arrow. Source: Modified from Bothe et al (2022).[23]

3.2 Testing the B-Score in a real-world scenario

3.2.1 Assessing the measurement performance on a small, real-world dataset

We were able to calculate the B-Score for the dobutamine dataset for both systolic and diastolic BP values. For systolic values, the tested device greatly outperformed the calculated base performances, leading to a B-Score of 0.94. However, the tested device did not outperform the base performances for diastolic BP measurements, leading to a B-Score of < 0.0 as per the B-Score's definition.[23] (Figure 6)

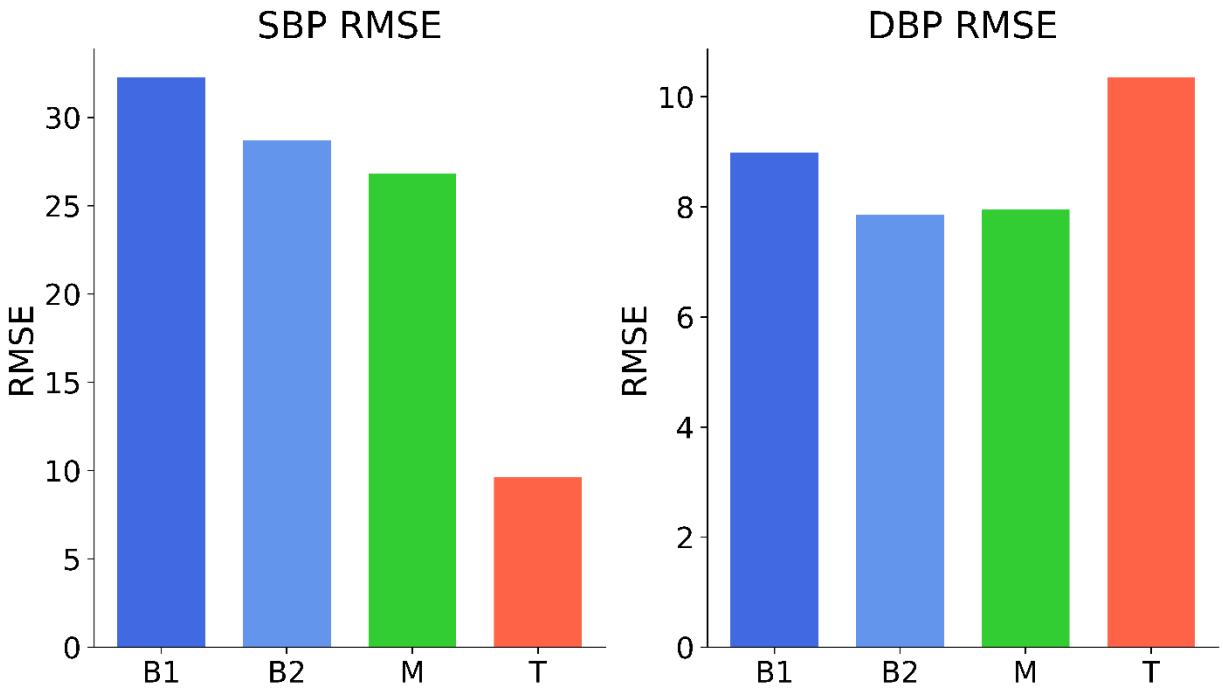


Figure 6: *Dobutamine* dataset *base performances*: The figure shows the calculated *base performances* for both systolic (left) and *diastolic* (right) BP values of the *dobutamine* dataset. The *base performances* are displayed in blue (B1-RMSE, B2-RMSE) and green (M-RMSE). The absolute measurement inaccuracy of the tested device is displayed in red (T-RMSE). The resulting B-Scores are 0.94 (systolic) and < 0.0 (diastolic). Source: From Bothe et *al.* (2022)[23]

3.2.2 Comparing the results to the largest available dataset

As for the *dobutamine* dataset, we calculated the *base performances* for the *MIMIC IV* dataset. Similarly, we found decreasing *base performance* measures (from B1-RMSE to M-RMSE) for systolic but this time also for diastolic BP measurements. (Figure 7)

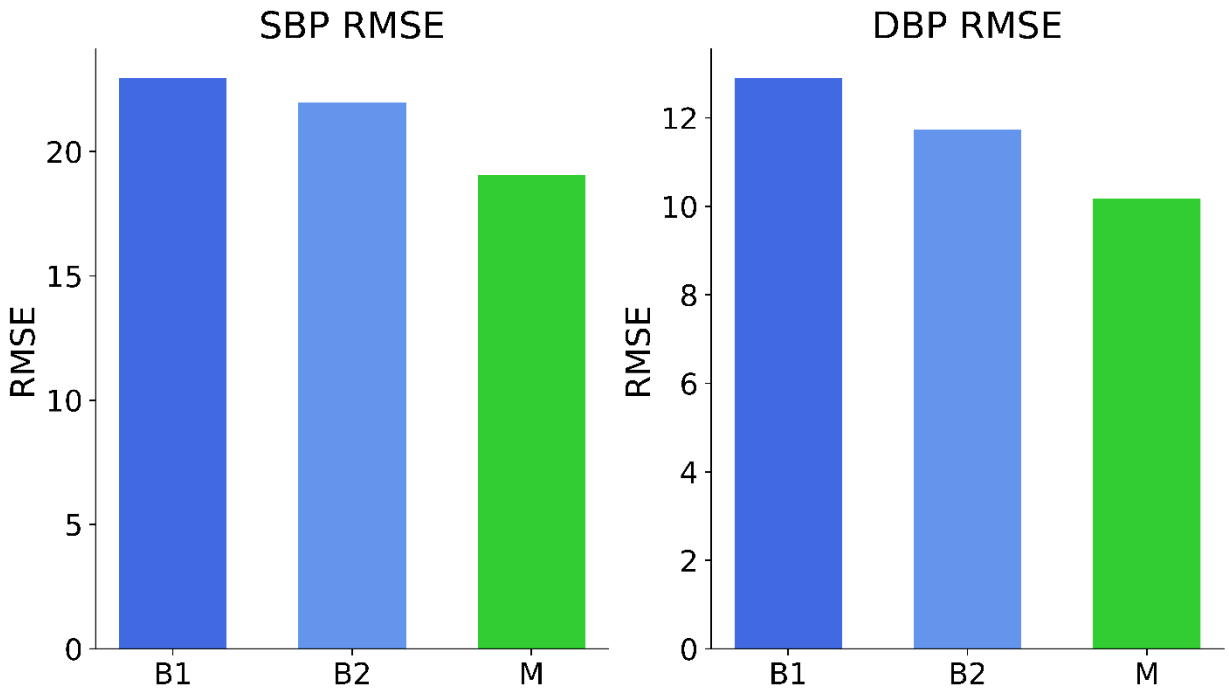


Figure 7: *Base performances* for the *MIMIC IV* dataset: The figure shows the *base performances* calculated for the *MIMIC IV* dataset. The *base performances* are displayed in blue (B1-RMSE, B2-RMSE) and green (M-RMSE). Source: From Bothe et *al.* (2022)[23]

Subsequently, we were able to use the derived *base performances* to calculate the T-RMSE value needed for any device (or BP estimation model) tested on the *MIMIC IV* dataset to claim coequal measurement performance to the device tested on the *dobutamine* dataset. We identified a T-RMSE of 6.98 mmHg as the point of coequal measurement performance. Smaller T-RMSE values would indicate a superior performance and higher T-RMSE values would vice versa result in a lower true measurement performance. (Figure 8) As the device tested on the *dobutamine* dataset did not reach a B-Score of over zero for diastolic values, the resulting goal for the *MIMIC IV* dataset would be to outperform the M-RMSE *base performance* of 10.18 mmHg. (Figure 8)

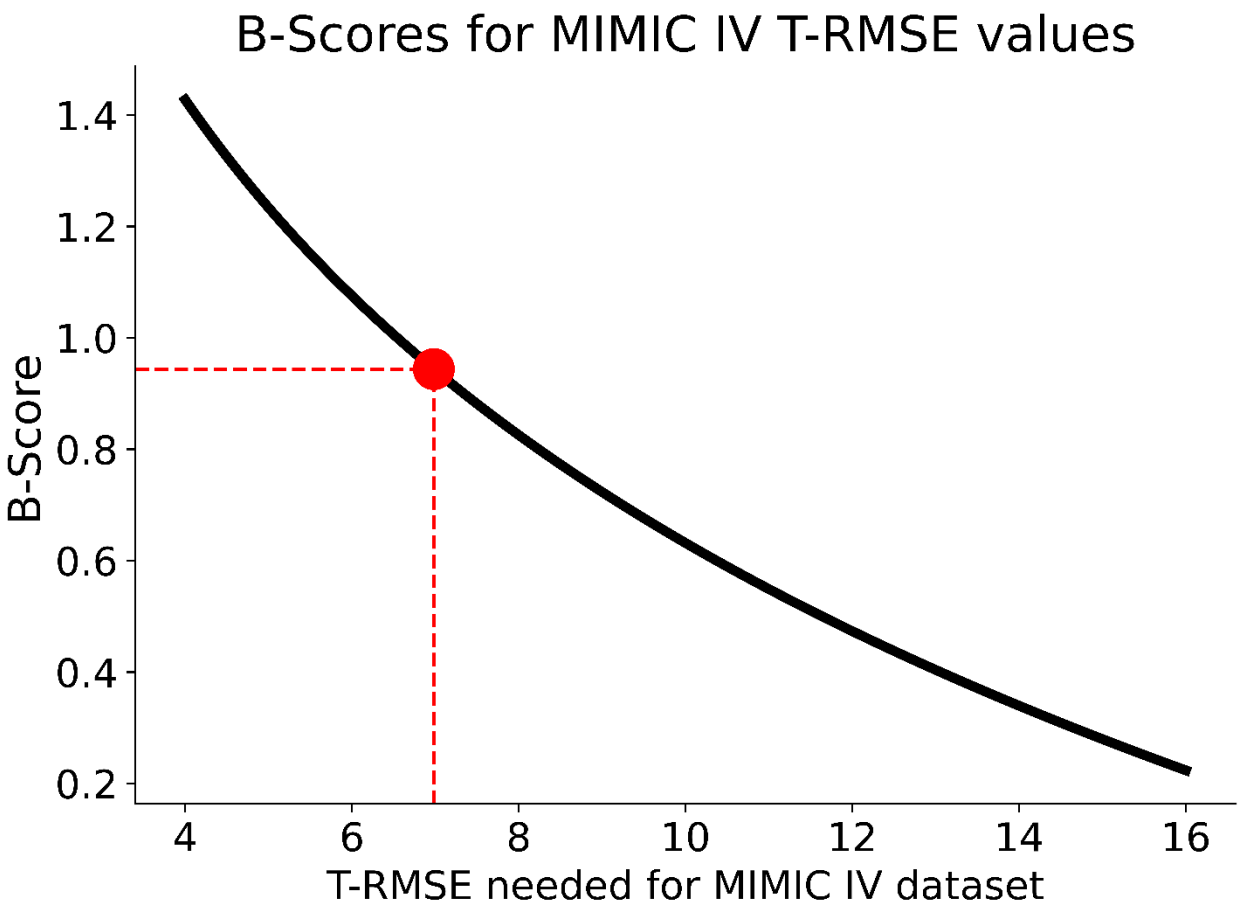


Figure 8: B-Scores for MIMIC IV T-RMSE values: The figure shows the B-Scores calculated for the *MIMIC* IV dataset for various systolic T-RMSE values. The red dot is defined as the point of coequal measurement performance to the *dobutamine dataset* (same B-Score, horizontal red line) and retrieves the T-RMSE needed to reach the same relative performance on the *MIMIC IV* dataset (vertical red line). Source: From Bothe et *al.* (2022)[23]

3.3 Analysing the time needed for calculating the B-Score

Calculating the B-Score for increasingly larger subsets of the MIMIC IV dataset allowed us to assess the time needed for B-Score. We derived a U-shaped time complexity with a minimum calculation time of 3 minutes for medium sized datasets (around 50,000 BP values). The U-shape is a result of the B-Score code conducting more re-calculations for smaller datasets to ensure reliable results. On the other side, very large datasets are more computationally demanding per dataset and therefore as well lead to an increased calculation time. (Figure 9)

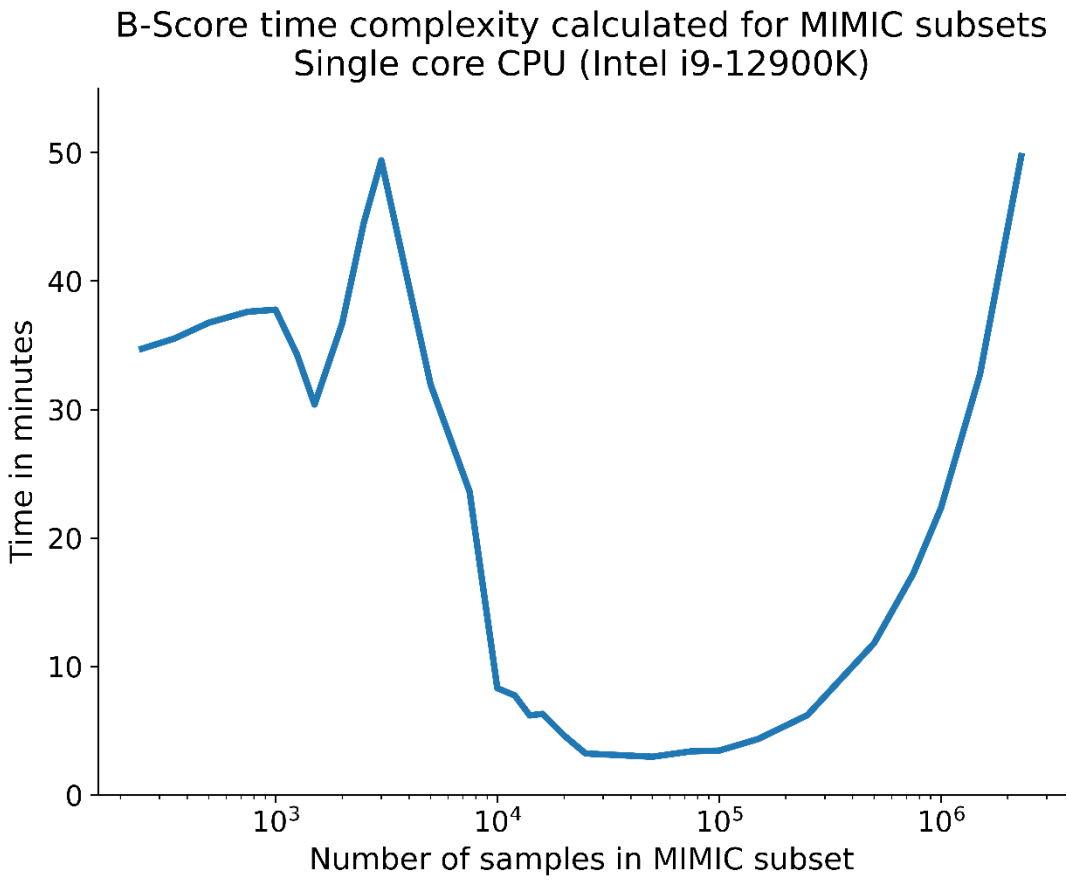


Figure 9: B-Score time complexity calculated for MIMIC subsets: The figure shows the B-Score's time complexity for calculations conducted on subsets of the *MIMIC IV* dataset. The U-shaped form is a consequence of the increased demand or re-calculations for smaller datasets (to ensure reliable results) and the increased computational load per calculation for larger datasets. Source: From Bothe et *al.* (2022)[23]

By conducting this analysis, we were further able to show that there is a M-RMSE insecurity of less than 3 mmHg for all and less than 1 mmHg for *MIMIC IV* subsets larger than 1,250 samples.[23]

4. Discussion

4.1 Summary and interpretation

In our analysis, the B-Score showed the desired mathematical properties, both when tested on simulated as well as on real-world datasets.

4.1.1 Interpretation of the results from the simulated dataset analyses

The B-Score's *base performances* increased under increasing BP variability in our analysis with simulated datasets. Similarly, the B-Score increased under increasing BP variability (with constant T-RMSE) and was modulated as expected by varying the T-RMSE value (increasing B-Score with decreasing T-RMSE).
Our subsequent test with the six large, simulated dataset yielded equally promising results: The *base performance* metrics showed an increased dataset difficulty (B1-, B2-, and M-RMSE) for increasingly difficult datasets. Following from this, we were able to calculate strongly differing B-Scores for the six datasets with the highest B-Score achieved on the most difficult dataset ('24-Hard'). This was the hoped-for result, as we calculated the B-Scores for a constant T-RMSE value.

4.1.2 Interpretation of the results from the real-world dataset analyses

After we conducted on the real-world datasets after confirming the B-Score's desired mathematical and practical properties with simulated data, we wanted to provide a real-world application example. To test the B-Score in the most extreme environment possible, we selected a very small dataset, generated in a highly dynamic environment, and compared it to an extremely large database of intensive care unit BP measurements.
The results for the *dobutamine* dataset are a picture-perfect example for why it is not only possible to calculate the B-Score on real-world data but moreover extremely important. When comparing the absolute measurement error for the systolic and diastolic BP values for the proposed device tested on the *dobutamine* dataset, there seems to be little difference (both around 10 mmHg). However, after calculating the *base performance* it became evident that, while the systolic T-RMSE greatly outperforms the *base performances*, the diastolic T-RMSE cannot even reach any of the *base performance* metrics. This led to a high B-Score for systolic but a B-Score below zero for diastolic values.

With traditional metrics for interpreting BP measurement performance, this divergence between systolic and diastolic performance would not have been detected. The B-Score easily allowed the discrimination between the well-performing systolic BP measurement and the very limited diastolic measurement performance.

Comparing these results with the *base performances* calculated for the *MIMIC IV* dataset retrieved the T-RMSE needed to claim coequal measurement performance to the device proposed in the *dobutamine* publication. The calculation is straightforwardly conducted by inversing the B-Score formula. In this way, we created a tool for researchers to analyse their datasets, retrieve the respective *base performances* and successively retrieve a T-RMSE value which they can set as a goal for their BP measurement models. This will be especially helpful for scientists working in the growing field of data-driven, continuous BP measurement.[5]

Finally, we were able to ensure that the B-Score can be computed quickly and with ease by showing the B-Score's time complexity and providing a user-friendly script which allows researchers to compute their own B-Scores, even with no or very limited programming expertise. Our time complexity analysis showed that the B-Score can be computed in under an hour for any available dataset size when run on a single core of a capable CPU. As the B-Score script can be run simultaneously for systolic and diastolic values on multi-core CPUs (nowadays the standard even for low-budget devices), we can confidently conclude that the B-Score can be calculated in one working day by any interested researcher on any moderately modern computer.

4.2 Strengths of the B-Score

In the B-Score, we created a metric suited for comparing BP measurement systems and models across a wide range of paradigmatic approaches and validation scenarios. To our knowledge, the B-Score is the first attempt to achieve a broadly applicable and easily interpretable metric for the evaluation of BP measurement devices.

4.2.1 Insights provided by the B-Score

In our work, we were able to develop the B-Score into a mathematical reliable metric which showed expected outcomes for a wide range of simulated data. Further, we were

able to provide an extreme example of a real-world use case for the B-Score by comparing to vastly different datasets. Moreover, we were not only capable of demonstrating the B-Score's readiness for real-world application but were able to highlight the value added by calculating the B-Score by revealing large performance differences for systolic and diastolic BP measurement in the *dobutamine* dataset. Notably, these differences would have stayed unnoticed without the B-Score as a measure of true, relative model performance.

4.2.2 A tool freely available to researchers around the world

The B-Score is designed as a tool for comparing BP measurement devices across publications and will only unlock its full potential if it gets widely adopted by the scientific community. Keeping this in mind, it was our focus during the development of the B-Score to provide a metric which can be quickly and easily calculated by researchers with all kinds (including none) of programming experience and with access to a wide range of computational hardware. Achieving this was a challenge, as we wanted to include a standardized Deep Learning model (as M-RMSE reference) into the B-Score.
The resulting B-Score can be easily calculated on any modern computational hardware within one working day – this is true for systolic and diastolic BP values at the same time when using an industry-standard multi-core CPU. Further, we published an easy-to-understand, which can be freely accessed by researchers and used to calculate their own B-Scores with little to no programming experience needed.[23]
We consider the level of transparency and user-oriented design as one of the main achievements of the B-Score – equally important to the invaluable insights it provides.

4.2.3 The chance of transforming hypertension diagnostics

When taking all of it into account, it becomes evident that the B-Score has the potential to revolutionize the way researchers and clinicians think about BP measurement – and therefore ultimately greatly affect the way patients are diagnosed and subsequently treated.
We created a tool which allows researchers to quickly evaluate their devices' and models' true measurement performances and to report them in an easily understandable and most importantly straightforwardly comparable way. Following from this, incorporating the B-Score allows scientists, clinicians and ultimately patients alike to evaluate immediately which device has the highest predictive value. This is an invaluable leap forward from the

situation today – in which those devices show the lowest absolute measurement error which have been tested on the easiest dataset.
In the future, scientists will be able to analyse and optimize their own devices and models in direct comparison to already published approaches, even when working with a vastly different patients and/or datasets. Beyond that, scientists will easily be able to conduct secondary analyses of already published approaches to BP measurement and identify promising trends in the scientific literature (e.g., a measurement paradigm achieving above-average B-Scores in multiple different publications).
This could lead the scientific community to invest more energy into truly promising approaches and reduce the time spent on underperforming paradigms. In particular, the highly diverse field of continuous BP measurement could benefit from separating the chaff from the wheat and focussing on the most promising approaches.[5] Subsequently, the B-Score provide the push needed for alternative BP measurement approaches to reach the evidence of measurement performance as demanded by the European Society of Hypertension [22] – which would constitute the biggest paradigmatic advance in hypertension diagnostics since the adoption of automated BP monitors.

4.3 Limitations of the B-Score and this work

However convinced we are to have solved an issue present in hypertension research for decades by developing tool for actually comparing true BP measurement performance, both the B-Score and this work have clear limitations.

4.3.1 The B-Score’s limitations

The B-Score is a metric for evaluating the true performance of BP measurement systems. As any metric, the B-Score has limitations, stemming from its conceptual design, mathematical properties, and data it is applied upon.

We designed the B-Score to compare a tested system (device / model) with a reference method, such as a validated BP monitor, auscultatory or intraarterial measurement. Unfortunately, therein lies a fundamental issue: Any reference method itself has a build in, hardly determinable measurement error. This is especially true for automated, cuff-based BP monitors, very frequently used as comparison for novel measurement approaches.[5,15–17]. Consequently, the B-Score might be ideally suited to evaluate the

error between a tested and a reference device but cannot discriminate whether the observed measurement error is due to a flawed proposed design or errors of the reference method.
However, choosing a bad reference method will in all likelihood lead to a deteriorated B-Score results. Therefore, the B-Score serves as an incentive to use the most precise reference method available. While this does not completely eradicate the issue at hand it likely is the most any metric can do.

Further, while the B-Score is the only metric which enables the comparison of very different approaches and datasets in the realm of BP measurement, there are some caveats with its interpretation. If the any given device shows differing B-Scores in different studies, it would be too easy to conclude that the B-Score is not working properly. To the contrary, it is plausible that one given BP measurement device does indeed perform differently well under different circumstances and in different patient collectives. A device developed on healthy, athletic young adults will most likely perform better in this subpopulation than in frail, elderly, multi-morbid patients. Any single study can therefore only be an indication of true BP measurement performance, even when enhanced by the B-Score. However, if a device can score above-average B-Scores in different studies, possibly conducted on different patient collectives, additional confidence in the devices general validity for BP measurement is warranted.

Lastly, the B-Score has a built-in feature, which is a mathematical limitation by design. Because of the issue of inherent measurement errors in every reference method there is a lower bound for the T-RMSE value - even if the tested device would measure BP perfectly accurately, there still would be a disagreement with the reference method. Consequently, for very easy datasets there is a natural limit to the achievable B-Score. This is the case, because for very easy dataset most BP measurement devices will score similar (in the order of magnitude of the reference method’s error) T-RMSE values, even if their true maximum performance might be very different.
As a solution, researchers are forced to test their datasets in dynamic measurement circumstances, ideally in a heterogenous patient collective, to increase overall BP variability and dataset difficulty and ultimately increase the datasets *base performances*. This in turn would allow promising BP measurement devices to outperform competitors an achieve higher B-Scores.

We wilfully accepted this circumstance during the B-Score's design, as we consider an incentive for developing and testing novel approaches for BP measurement on difficult and heterogenous datasets a feature. Devices which perform well on such data are more likely to perform well under a wide variety of circumstances and warrant increased confidence in the study's results.

4.3.2 Limitations of this work

Our work up to this point and especially this work and the original B-Score publication [23] are clearly limited by their methodological character. We have developed and thoroughly tested a novel score for the evaluation of the true performance of BP measurement systems. While we consider this an important step towards the advancement of BP measurement, we were not yet able to provide deep insights by applying the B-Score on a wide range of already existing systems.

There are two reasons why we refrained from trying to incorporate a larger, real-world analysis using the B-Score at this stage of our research.

Firstly, while the B-Score is designed to be easily interpretable, its internal concepts and mathematical underpinning is complex. We made a great effort to present our results in a comprehensive and understandable manner but are aware that understanding all details of the B-Score's design can be challenging. Adding a full-scale analysis of multiple, additional datasets would have further increased the complexity of the B-Score publication and would have been unsuited for a first publication, aimed at presenting, and explain the B-Score's features and use-cases.

Secondly, researchers are rightfully sceptical to share their data with other researchers, if they do not understand what kind of analysis is planned with their data. Therefore, accessing third-party data is a challenge of its own, greatly aggravated when proposing a novel, unknown and *unpublished* score. Consequently, in accordance with the reasons stated above, we decided to first publish a methodological paper with a limited, real-world example. Naturally, it was and still is our ambition to build on that foundation and provide larger analyses of multiple datasets and approaches based on the B-Score.

4.4 A perspective on the future of the B-Score and our work

Time will tell whether the B-Score is able to reach its full potential and become an integral part of how research on BP measurement is conducted and discussed. We are aware

that the B-Score's utility is strongly connected to its rate of adoption. The more researchers use the B-Score and publish their results, the more points-of-reference are out in the scientific literature.
To propel the B-Score's adoption as much as possible, we as a group and especially me personally are taking actions to integrate the B-Score in projects and discussions with researchers around the world. Lastly, we want to make further efforts to reduce the barriers for calculating the B-Score. In this sense, we are working on providing a downloadable application ("B-Score App") which will greatly further simplify the process of calculating the B-Score.

Besides applying the B-Score directly, we are engaged in providing further insights into issues arising in the field of BP measurement in general. We are greatly concerned by the unknown extent of hidden measurement error in reference devices (devices in everyday clinical use). Consequently, we are conducting large-scale experiments to quantify those measurement errors and therefore provide even more insights into how BP measurement accuracy should be assessed could be improved.

5. Conclusion

The B-Score is a novel tool for evaluating the BP measurement devices and comparing their true measurement performance across a wide range of datasets. We developed the B-Score to set a device's performance in contrast to the difficulty of the dataset it was tested upon. By doing so, we are able to derive a single metric depicting the device's true BP measurement performance.

We tested the B-Score on a variety of simulated dataset of different difficulty and size and were able to confirm the B-Score's mathematical predictability and desired properties. Further, we tested the B-Score in an extreme real-world scenario and highlighted the additional insights provided by the B-Score over conventional metrics.

In the future, we want to build on the published results and establish the B-Score as a tool for evaluating BP measurement systems and want to provide further insights into trends in the latest scientific literature. Further, we are working on various other projects in the realm of improving BP measurement both in our working group and with researchers from around the globe.

References

1. Fröhlich H, Balling R, Beerenwinkel N, Kohlbacher O, Kumar S, Lengauer T, Maathuis MH, Moreau Y, Murphy SA, Przytycka TM, Rebhan M, Röst H, Schuppert A, Schwab M, Spang R, Stekhoven D, Sun J, Weber A, Ziemek D, Zupan B. From hype to reality: Data science enabling personalized medicine. BMC Med. 2018 Aug 27;16(1):1–15.
2. Cirillo D, Valencia A. Big data analytics for personalized medicine. Curr Opin Biotechnol [Internet]. 2019 Aug 1 [cited 2022 Oct 13];58:161–7.
3. Hamburg MA, Collins FS. The Path to Personalized Medicine. New England Journal of Medicine. 2010 Jul 22;363(4):301–4.
4. Mills KT, Stefanescu A, He J. The global epidemiology of hypertension. Nature Reviews Nephrology 2020 16:4. 2020 Feb 5;16(4):223–37.
5. Pilz N, Patzak A, Bothe TL. Continuous cuffless and non-invasive measurement of arterial blood pressure—concepts and future perspectives. https://doi.org/101080/0803705120222128716. 2022 Dec 31;31(1):254–69.
6. Fuentes A v., Pineda MD, Venkata KCN. Comprehension of Top 200 Prescribed Drugs in the US as a Resource for Pharmacy Teaching, Training and Practice. Pharmacy 2018, Vol 6, Page 43. 2018 May 14;6(2):43.
7. Williams B, Mancia G, Spiering W, Rosei EA, Azizi M, Burnier M, Clement DL, Coca A, de Simone G, Dominiczak A, Kahan T, Mahfoud F, Redon J, Ruilope L, Zanchetti A, Kerins M, Kjeldsen SE, Kreutz R, Laurent S, Lip GYH, McManus R, Narkiewicz K, Ruschitzka F, Schmieder RE, Shlyakhto E, Tsioufis C, Aboyans V, Desormais I, de Backer G, Heagerty AM, Agewall S, Bochud M, Borghi C, Boutouyrie P, Brguljan J, Bueno H, Caiani EG, Carlberg B, Chapman N, Cífková R, Cleland JGF, Collet JP, Coman IM, de Leeuw PW, Delgado V, Dendale P, Diener HC, Dorobantu M, Fagard R, Farsang C, Ferrini M, Graham IM, Grassi G, Haller H, Hobbs FDR, Jelakovic B, Jennings C, Katus HA, Kroon AA, Leclercq C, Lovic D, Lurbe E, Manolis AJ, McDonagh TA, Messerli F, Muiesan ML, Nixdorff U, Olsen MH, Parati G, Perk J, Piepoli MF, Polonia J, Ponikowski P, Richter DJ, Rimoldi SF, Roffi M, Sattar N, Seferovic PM, Simpson IA, Sousa-Uva M, Stanton A v., van de Borne P, Vardas P, Volpe M, Wassmann S, Windecker S, Zamorano JL. 2018 ESC/ESH Guidelines for the management of arterial hypertension. Eur Heart J. 2018 Sep 1;39(33):3021–104.

8. Whelton PK, Carey RM, Aronow WS, Casey DE, Collins KJ, Himmelfarb CD, DePalma SM, Gidding S, Jamerson KA, Jones DW, MacLaughlin EJ, Muntner P, Ovbiagele B, Smith SC, Spencer CC, Stafford RS, Taler SJ, Thomas RJ, Williams KA, Williamson JD, Wright JT, Levine GN, O'Gara PT, Halperin JL, Past I, Al SM, Beckman JA, Birtcher KK, Bozkurt B, Brindis RG, Cigarroa JE, Curtis LH, Deswal A, Fleisher LA, Gentile F, Goldberger ZD, Hlatky MA, Ikonomidis J, Joglar JA, Mauri L, Pressler SJ, Riegel B, Wijeysundera DN, Walsh MN, Jacobovitz S, Oetgen WJ, Elma MA, Scholtz A, Sheehan KA, Abdullah AR, Tahir N, Warner JJ, Brown N, Robertson RM, Whitman GR, Hundley J. 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Hypertension. 2018 Jun 1;71(6):E13–115.
9. Stergiou GS, Palatini P, Parati G, O'Brien E, Januszewicz A, Lurbe E, Persu A, Mancia G, Kreutz R. 2021 European Society of Hypertension practice guidelines for office and out-of-office blood pressure measurement. J Hypertens. 2021 Jul 1;39(7):1293–302.
10. Mancia G, Sega R, Bravi C, de Vito G, Valagussa F, Cesana G, Zanchetti A. Ambulatory blood pressure normality: Results from the PAMELA study. J Hypertens [Internet]. 1995 [cited 2020 Apr 14];13(12 I):1377–90.
11. Sega R, Facchetti R, Bombelli M, Cesana G, Corrao G, Grassi G, Mancia G. Prognostic value of ambulatory and home blood pressures compared with office blood pressure in the general population: Follow-up results from the Pressioni Arteriose Monitorate e Loro Associazioni (PAMELA) study. Circulation [Internet]. 2005;111(14):1777–83.
12. Gavriilaki M, Anyfanti P, Nikolaidou B, Lazaridis A, Gavriilaki E, Douma S, Gkaliagkousi E. Nighttime dipping status and risk of cardiovascular events in patients with untreated hypertension: A systematic review and meta-analysis. The Journal of Clinical Hypertension. 2020 Nov 8;22(11):1951–9.
13. Agarwal R, Light RP. The effect of measuring ambulatory blood pressure on nighttime sleep and daytime activity - Implications for dipping. Clinical Journal of the American Society of Nephrology. 2010;5(2):281–5.

14. Sherwood A, Hill LK, Blumenthal JA, Hinderliter AL. The Effects of Ambulatory Blood Pressure Monitoring on Sleep Quality in Men and Women with Hypertension: Dipper vs. Nondipper and Race Differences. Am J Hypertens. 2019 Jan 1;32(1):54–60.
15. Davies RJO, Jenkins NE, Stradling JR. Effect of measuring ambulatory blood pressure on sleep and on blood pressure during sleep. BMJ : British Medical Journal. 1994 Mar 26;308(6932):820.
16. Zheng D, Giovannini R, Murray A. Effect of respiration, talking and small body movements on blood pressure measurement. Journal of Human Hypertension 2012 26:7. 2011 May 26;26(7):458–62.
17. Stergiou GS, Kyriakoulis KG, Stambolliu E, Destounis A, Karpettas N, Kalogeropoulos P, Kollias A. Blood pressure measurement in atrial fibrillation: review and meta-analysis of evidence on accuracy and clinical relevance. J Hypertens. 2019 Dec 1;37(12):2430–41.
18. Bothe TL, Bilo G, Parati G, Haberl R, Pilz N, Patzak A. Impact of oscillometric measurement artefacts in ambulatory blood pressure monitoring on estimates of average blood pressure and of its variability: a pilot study. J Hypertens [Internet]. 2022 Oct 21;
19. Stergiou GS, Alpert BS, Mieke S, Wang J, O'Brien E. Validation protocols for blood pressure measuring devices in the 21st century. J Clin Hypertens (Greenwich). 2018 Jul 1;20(7):1096–9.
20. Mochizuki Y, Okutani M, Donfeng Y, Iwasaki H, Takusagawa M, Kohno I, Mochizuki S, Umetani K, Ishii H, Ijiri H, Komori S, Tamura K. Limited Reproducibility of Circadian Variation in Blood Pressure Dippers and Nondippers. Am J Hypertens. 1998 Apr 1;11(4):403–9.
21. Stergiou GS, Baibas NM, Gantzarou AP, Skeva II, Kalkana CB, Roussias LG, Mountokalakis TD. Reproducibility of home, ambulatory, and clinic blood pressure: implications for the design of trials for the assessment of antihypertensive drug efficacy. Am J Hypertens. 2002 Feb 1;15(2):101–4.
22. Stergiou GS, Mukkamala R, Avolio A, Kyriakoulis KG, Mieke S, Murray A, Parati G, Schutte AE, Sharman JE, Asmar R, McManus RJ, Asayama K, de La Sierra A, Head G, Kario K, Kollias A, Myers M, Niiranen T, Ohkubo T, Wang J, Wuerzner G, O'Brien E, Kreutz R, Palatini P. Cuffless blood pressure measuring devices: review and statement by the European Society of Hypertension Working Group on Blood

Pressure Monitoring and Cardiovascular Variability. J Hypertens. 2022 Aug 17;40(8).

23. Bothe TL, Patzak A, Pilz N. The B-Score is a novel metric for measuring the true performance of blood pressure estimation models. Scientific Reports 2022 12:1. 2022 Jul 16;12(1):1–10.
24. Douma LG, Gumz ML. Circadian clock-mediated regulation of blood pressure [Internet]. Vol. 119, Free Radical Biology and Medicine. Elsevier Inc.; 2018. p. 108–14. Available from: https://pubmed.ncbi.nlm.nih.gov/29198725/
25. Julien C. The enigma of Mayer waves: Facts and models. Cardiovasc Res. 2006 Apr 1;70(1):12–21.
26. Patzak A, Mendoza Y, Gesche H, Konermann M. Continuous blood pressure measurement using the pulse transit time: Comparison to intra-arterial measurement. Blood Press [Internet]. 2015 Aug 1;24(4):217–21.
27. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK, Stanley HE. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. Circulation. 2000 Jun 13;101(23).
28. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. MIMIC-IV (version 1.0). PhysioNet. 2021.
29. Mckinney W. Data Structures for Statistical Computing in Python. In: Stefan van der Walt, Jarrod Millman, editors. PROC OF THE 9th PYTHON IN SCIENCE CONF [Internet]. 2010 [cited 2021 May 7]. p. 56–61.
30. Pedregosa F, Michel V, Grisel O, Blondel M, Prettenhofer P, Weiss R, Vanderplas J, Cournapeau D, Pedregosa F, Varoquaux G, Gramfort A, Thirion B, Grisel O, Dubourg V, Passos A, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python [Internet]. Vol. 12, Journal of Machine Learning Research. 2011. Available from: http://scikit-learn.sourceforge.net.
31. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, Vijaykumar A, Bardelli A pietro, Rothberg A, Hilboll A, Kloeckner A, Scopatz A, Lee A, Rokem A, Woods CN, Fulton C, Masson C, Häggström C, Fitzgerald C,

Nicholson DA, Hagen DR, Pasechnik D v., Olivetti E, Martin E, Wieser E, Silva F, Lenders F, Wilhelm F, Young G, Price GA, Ingold GL, Allen GE, Lee GR, Audren H, Probst I, Dietrich JP, Silterra J, Webber JT, Slavič J, Nothman J, Buchner J, Kulick J, Schönberger JL, de Miranda Cardoso JV, Reimer J, Harrington J, Rodríguez JLC, Nunez-Iglesias J, Kuczynski J, Tritz K, Thoma M, Newville M, Kümmerer M, Bolingbroke M, Tartre M, Pak M, Smith NJ, Nowaczyk N, Shebanov N, Pavlyk O, Brodtkorb PA, Lee P, McGibbon RT, Feldbauer R, Lewis S, Tygier S, Sievert S, Vigna S, Peterson S, More S, Pudlik T, Oshima T, Pingel TJ, Robitaille TP, Spura T, Jones TR, Cera T, Leslie T, Zito T, Krauss T, Upadhyay U, Halchenko YO, Vázquez-Baeza Y. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods 2020 17:3. 2020 Feb 3;17(3):261–72.

32. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Rafal Jozefowicz, Yangqing Jia, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Mike Schuster, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, Xiaoqiang Zheng. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems [Internet]. 2015. Available from: https://www.tensorflow.org/
33. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng [Internet]. 2007;9(3):90–5.
34. Bothe TL, Pilz N, Dippel LJ. The compass of biomedicine. Acta Physiologica. 2022 Sep 1;236(1):e13856.
35. Bothe TL, Dippel LJ, Pilz N. The Art of Planning-How many samples are enough? Acta Physiol (Oxf). 2022 Feb 1;234(2).
36. Bothe TL, Patzak A, Schubert R, Pilz N. Getting it right matters! Covid-19 pandemic analogies to everyday life in medical sciences. Acta Physiologica [Internet]. 2021 Sep 1;233(1).
37. Bothe TL, Patzak A. Significant significance? [Internet]. Acta Physiologica. Blackwell Publishing Ltd; 2021. p. e13665. Available from: https://doi.org/10.1111/apha.13665

Statutory Declaration

“I, Tomas Lucca Bothe, by personally signing this document in lieu of an oath, hereby affirm that I prepared the submitted dissertation on the topic *“The B-Score: Evaluation of blood pressure measurement systems by a novel score for the determination of the true, relative measurement performance” (“Der B-Score: Ein neuer Score zur Evaluierung der wahren, relativen Leistungsfähigkeit von Blutdruckmesssystemen”)*, independently and without the support of third parties, and that I used no other sources and aids than those stated.
All parts which are based on the publications or presentations of other authors, either in letter or in spirit, are specified as such in accordance with the citing guidelines. The sections on methodology (in particular regarding practical work, laboratory regulations, statistical processing) and results (in particular regarding figures, charts and tables) are exclusively my responsibility.

Furthermore, I declare that I have correctly marked all of the data, the analyses, and the conclusions generated from data obtained in collaboration with other persons, and that I have correctly marked my own contribution and the contributions of other persons (cf. declaration of contribution). I have correctly marked all texts or parts of texts that were generated in collaboration with other persons.

My contributions to any publications to this dissertation correspond to those stated in the below joint declaration made together with the supervisor. All publications created within the scope of the dissertation comply with the guidelines of the ICMJE (International Committee of Medical Journal Editors; http://www.icmje.org) on authorship. In addition, I declare that I shall comply with the regulations of Charité – Universitätsmedizin Berlin on ensuring good scientific practice.

I declare that I have not yet submitted this dissertation in identical or similar form to another Faculty.

The significance of this statutory declaration and the consequences of a false statutory declaration under criminal law (Sections 156, 161 of the German Criminal Code) are known to me.”

Date Signature

Declaration of your own contribution to the publications

Tomas Lucca Bothe contributed the following to the below listed publications:

Publication 1: Bothe TL, Patzak A, Pilz N. The B-Score is a novel metric for measuring the true performance of blood pressure estimation models. Scientific Reports 2022

Contribution: I designed the B-Score and was the main contributor in all phases of its creation. Hence, I came up with the idea to develop a metric for evaluating the true performance of blood pressure measurement systems. Followingly, I devised a project plan, including the testing of the B-Score on simulated and real-world data.
Together with Mr. Pilz, I developed and iteratively improved the mathematical underpinning of the B-Score. Further, I wrote the code for the simulated datasets, preparation of the real-world datasets and subsequent testing. Mr. Pilz and Prof. Patzak supported my efforts with continued, invaluable discussions. Additionally, Mr. Pilz reviewed the code for plausibility.
All eight figures and their results were created by me, as well as the user-friendly B-Score script, which was published as a supplement and was thoroughly reviewed by Mr. Pilz. Finally, I wrote the first manuscript draft and finalized it with Prof. Patzak's and Mr. Pilz's continued support.

_____

Signature, date and stamp of first supervising university professor / lecturer

_____

Signature of doctoral candidate

Excerpt from Journal Summary List

Journal Data Filtered By: **Selected JCR Year: 2020** Selected Editions: SCIE,SSCI
Selected Categories: **"MULTIDISCIPLINARY SCIENCES"** Selected Category Scheme: WoS
Gesamtanzahl: 73 Journale

Rank	Full Journal Title	Total Cites	Journal Impact Factor	Eigenfactor Score
1	NATURE	915,925	49.962	1.089400
2	SCIENCE	814,971	47.728	0.895760
3	National Science Review	5,889	17.275	0.011400
4	Nature Communications	453,215	14.919	1.238540
5	Science Advances	65,205	14.136	0.218640
6	Nature Human Behaviour	5,549	13.663	0.023120
7	Science Bulletin	8,832	11.780	0.016400
8	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA	799,058	11.205	0.806620
9	Journal of Advanced Research	5,927	10.479	0.006800
10	GigaScience	5,876	6.524	0.018630
11	Scientific Data	10,617	6.444	0.034470
12	Frontiers in Bioengineering and Biotechnology	7,470	5.890	0.011340
13	ANNALS OF THE NEW YORK ACADEMY OF SCIENCES	52,619	5.691	0.021430
14	iScience	5,235	5.458	0.012300
15	Research Synthesis Methods	3,926	5.273	0.007520
16	NPJ Microgravity	594	4.415	0.001790
17	Scientific Reports	541,615	4.379	1.232500
18	PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A-MATHEMATICAL PHYSICAL AND ENGINEERING SCIENCES	24,950	4.226	0.025400

Printing copy of the publication

www.nature.com/scientificreports

scientific reports

Check for updates

OPEN

The B-Score is a novel metric for measuring the true performance of blood pressure estimation models

Tomas L. Bothe✉, Andreas Patzak & Niklas Pilz

We aimed to develop and test a novel metric for the relative performance of blood pressure estimation systems (B-Score). The B-Score sets absolute blood pressure estimation model performance in contrast to the dataset the model is tested upon. We calculate the B-Score based on inter- and intrapersonal variabilities within the dataset. To test the B-Score for reliable results and desired properties, we designed generic datasets with differing inter- and intrapersonal blood pressure variability. We then tested the B-Score's real-world functionality with a small, published dataset and the largest available blood pressure dataset (MIMIC IV). The B-Score demonstrated reliable and desired properties. The real-world test provided allowed the direct comparison of different datasets and revealed insights hidden from absolute performance measures. The B-Score is a functional, novel, and easy to interpret measure of relative blood pressure estimation system performance. It is easily calculated for any dataset and enables the direct comparison of various systems tested on different datasets. We created a metric for direct blood pressure estimation system performance. The B-Score allows researchers to detect promising trends quickly and reliably in the scientific literature. It further allows researchers and engineers to quickly assess and compare performances of various systems and algorithms, even when tested on different datasets.

High arterial blood pressure (BP) levels lead to higher numbers of cardiovascular events and all-cause mortality[1,2]. Cuff-based BP measurement devices have dominated the field of arterial BP determination for over one century. There has been substantial recent interest in alternative, cuff-less and continuous BP measurement devices[1–8]. Cuff-less BP measurement has the advantage of being possibly less disturbing (night BP) and providing beat-to-beat BP data. More data points benefit secondary parameter calculation, such as the BP variability[9–14]. The growing interest manifests itself in a multitude of research papers, proposing various options for cuff-less BP estimation[15–20].

Model performance has been assessed in a multitude of ways. Unfortunately, the performance of proposed BP estimation models is only evaluated by absolute metrics. Used absolute metrics include the mean value deviation, standard deviation, mean absolute error, root mean squared error (RMSE), and many more. Regrettably, these absolute values depend not only on the BP model's sophistication but also on what dataset it is used on. Datasets can be very different: Compare 24 h ambulatory BP measurements of clinical patients with measurements taken at rest in a laboratory setting amongst the young and healthy. Clinical patients are a heterogeneous group and therefore BP values differ more between individual patients than between young and healthy subjects (inter-personal variability). Further, 24 h measurements are less stable than measurements taken at rest within single patients (intrapersonal variability). This results in a critical problem: A given absolute value (e.g., mean value deviation) does not provide information about the true model performance. In reverse, model performances cannot be compared when not tested on the same dataset. Absolute metrics will show better results for "easier" (lower variability) datasets. Two options remain: Testing every model on the same dataset or creating a metric able to depict dataset-adjusted performance. Option one seems to be nearly impossible due to various reasons (e.g., proposed models use very different input data)[7,21–23]. In consequence, a new metric of dataset-adjusted (relative) model performance is needed. Such metric would allow easy and reliable comparison between models even when tested on different datasets.

Charité – Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Institute of Translational Physiology, Chariteplatz 1, 10117 Berlin, Germany. ✉email: tomas-lucca.bothe@charite.de

After realizing this lack of comparability between different BP estimation model performances, we decided to develop a metric of relative model performance, the Base-Score (B-Score). The B-Score is designed to be intuitively understandable. It allows direct and easy evaluation of model performance, as higher B-Scores equal better dataset-adjusted (relative) model performance.

We tested the B-Score on generic datasets and further provide a real-world application example of comparing two very different datasets in this work. We used a published model tested on a small dataset and set it in contrast to the MIMIC IV clinical database, the to our knowledge largest dataset available[24,25].

Methods

The B-Score is calculated by comparing a proposed model's absolute performance to dataset specific base-performances (B1, B2). These base performances are: The B1-performance, measuring the interpersonal variability, the B2-performance, measuring the intrapersonal variability, and the M-performance, measuring the performance of a minimalistic BP estimation model (M). Setting the absolute performance of the researcher's model (T) into contrast with the three base performances (B1, B2 and M) results in a novel measure of relative performance. The B-Score for systolic and diastolic estimations are calculated separately, allowing a differentiated analysis of systolic and diastolic performance.

B-Score calculation. *Root mean squared error (RMSE).* The B-Score to evaluates relative model performance based on absolute performance measures. The metric of absolute model performance on which the B-Score is based on is the root mean squared error (RMSE). We chose the RMSE for a specific reason: It is a measure of average differences between a researcher's values and a reference method but also takes the reliability of those differences into account. This becomes evident when comparing the RMSE to another measure of absolute performance (e.g., the mean absolute error). Being off 4 mmHg in every measurement leads to a mean absolute error of 4.0. If every second measurement is perfectly accurate but every other measurement is off by 8 mmHg the mean absolute error stays 4.0. In scenario one the RMSE equates to 4.0 as well but changes to 5.66 in the second case. The RMSE provides additional information about the measurement consistency which is important for any BP estimation. Further, it also provides information about the absolute error, combining the advantages of pure absolute measures (e.g., mean absolute error) and pure measures of proportionality (e.g., correlation coefficient)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(prediction - reference\ value\right)^2},$$

n = number of samples.

Test-RMSE. The Test-RMSE (T-RMSE) is the measure of absolute performance which we designed the B-Score to base upon. It is the RMSE between the BP estimations a researcher's model derived values and the reference values provided in the study.

$$TRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(researcher'model\ prediction - reference\ value\right)^2},$$

n = number of samples.

B1-RSME. The B1-RMSE is the measure of *inter*personal variability in the researcher's dataset. The B1-RMSE is calculated as the RMSE between the mean of all reference BP values and every single reference value. This is closely connected to the dataset's standard deviation but not entirely equal. The reason for choosing the RMSE are explained above.

$$B1RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(cohort\ mean - reference\ value\right)^2},$$

n = number of samples.

B2-RMSE. The B2-RMSE is the measure of *intra*personal variability in the researcher dataset. The first measurement for every subject is defined as their personal "calibration value". The B2-RMSE is calculated as the RMSE between the calibration value (subject specific) and every single reference value. This equates to estimating a subjects first measurement result for every upcoming measurement.

$$B2RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\text{“}calibration\text{”}\ value - reference\ value\right)^2},$$

n = number of samples.

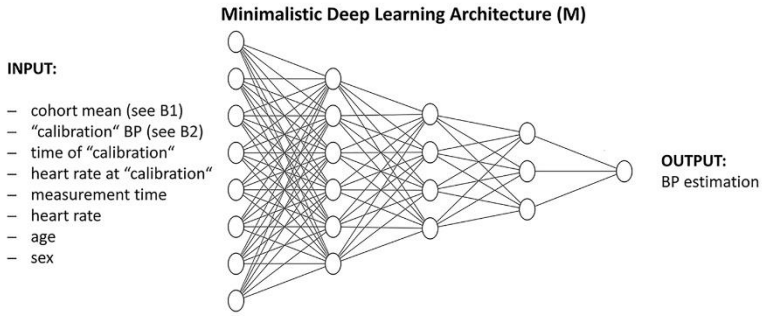


Figure 1. Minimalistic Deep Learning architecture used for M-RMSE calculation. It is a five-layer feed-forward Neural Network with Dropout and L2-Regression for ensuring reliable results. The system must be retrained for every dataset and provides BP estimations which are needed to M-RMSE calculation. For further information please see Supplementary Appendix 1/2.

M-RMSE. The M-RMSE is the measure of how easy it is to derive a well-performing BP estimation model for the researcher's dataset. To assess this, we created a minimalistic Deep Learning architecture. This architecture does not change but must be retrained for every given dataset (as well as for systolic and diastolic values). We designed the architecture to ensure reliable results and therefore applicability for almost all datasets.

This minimalistic tool intakes parameters present in every BP dataset: The subjects age, sex, the time of measurement, the heart rate, a calibration BP (see B2-RMSE), the time of calibration, the calibration heart rate, and the mean of all reference values (see B1-RMSE). The system then estimates BP values (Fig. 1).

The M-RMSE is calculated as the RMSE between the system's estimations and every single reference value. Detailed insight into the Deep Learning architecture can be found in the Supplementary Appendix and the provided code (Supplementary Appendix 1/2).

$$MRMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(minimalistic\ model\ predictions - reference\ value\right)^2},$$

n = number of samples.

As the M-RMSE is only based on a single BP calibration and subsequent monitoring of heart rate and time, it would be easily implemented using a singular cuff measurement and subsequent "smartwatch" monitoring. The M-RMSE is therefore a reasonable minimal standard for any BP estimation model to beat.

B-Score. To retrieve a measure of relative performance the B-Score sets the T-RMSE (absolute performance of researcher's model) in relation to the three presented, dataset-specific RMSE values (B1, B2, M). The B-Score is designed to increase with increasing model performance. To achieve this, it rises the more the tested model (T-RMSE) outperforms the base performances (B1-, B2- and M-RMSE):

$$BScore = \log_{10}\left(\sqrt{\left(\frac{B1RMSE \cdot M\ RMSE}{T\ RMSE^2}\right) \cdot \left(\frac{B2\ RMSE \cdot M\ RMSE}{T\ RMSE^2}\right)}\right).$$

We defined the B-Score for all instances in which the T-RMSE is smaller than the M-RMSE. If this is not the case the proposed model performs worse than the minimalistic Deep Learning architecture and possibly worse than B1 or B2. We recommend to report "B-Score < 0.00" for every case that the T-RMSE is greater than any base performance (B1-, B2-, M-RMSE). The B-Score should be reported rounded to two decimal places.

A minimum of three patients with at least three measurements per patient are needed for B-Score calculation. We do not recommend calculating the B-Score for datasets containing less than 100 distinct BP measurements to ensure reliable results.

Reliable B-Score results. In line with the best practices of Machine Learning the M-RMSE is calculated via k-fold evaluation. To ensure reliable results, the calculation is repeated, resampled, and averaged, depending on the datasets size. In some cases, the M-RMSE might be larger than other base performances because of measures taken to ensure generalized model behaviour (e.g., Dropout, L2-regularization). Specific information is available in the provided code (Supplementary Appendix 2).

Testing the B-Score for desired properties. We tested the B-Score to confirm reliable results and desired properties of identifying superior BP estimation systems on generic datasets.

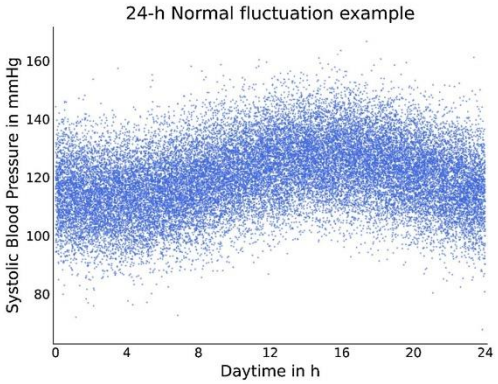


Figure 2. Plot of 30,000 samples from the systolic "24-h Normal" dataset. The normally distributed circadian rhythm is displayed.

Dataset generation. We created three datasets (3 × systolic + diastolic) to test the B-Score. The datasets display basic characteristics of BP profiles. Short-time and long-time fluctuation rhythms primarily influence BP fluctuations, which themselves are based on the Traube-Hering-Mayer- and circadian rhythms[26–28]. Based on these rhythms we modelled two (2 × systolic + diastolic) 24-h BP datasets. The datasets differ in inter- and intrapersonal variability and are therefore named "Normal 24-h" and "Hard 24-h" datasets. We further modelled one dataset (systolic + diastolic) to be a 30-min dataset, replicating a laboratory measurement setting. It is named the "Lab" dataset. Additional information about the dataset generation is presented in the Supplementary Appendix 3.

For every dataset, 10,000 subjects with 50 measurements each were simulated, resulting in datasets with 500,000 single measurement entries (Fig. 2).

We set a fictional T-RMSE value of 4 mmHg and calculated the B-Score accordingly for all six datasets. The B-Score can be considered functional if calculation is trouble-free and the retrieved B-Scores clearly rank the datasets from lowest (easiest, "Lab") to highest (hardest, "24-h Hard").

B-Score under increasing standard deviation. We further simulated smaller (50,000 BP values) generic datasets with increasing BP standard deviation. Followingly, we calculated the base performances (B1-, B2- and M-RMSE) and B-Scores for all datasets and plotted the results against the BP standard deviation.

Published small dataset. We used a small dataset published by Patzak et al. in 2015 to test the B-Score at the lower boundary of dataset size. The dataset consists of 12 patients with a total of 107 (each systolic and diastolic) measurements. The authors validated a pulse-wave-velocity-based BP estimation device against intraarterial measurements taken during dobutamine induced BP increases[20]. We calculated the systolic and diastolic B-Score for the proposed device and dataset.

MIMIC IV dataset. The MIMIC IV clinical dataset is the latest iteration of the to our knowledge largest available clinical dataset providing BP data. Data are descended from mainly ICU patients[24,25]. We used the MIMIC IV dataset to stress test the B-Score for applicability for the largest available dataset.

Data cleaning. We pre-processed the dataset to only hold data point suitable for the B-Score. Specifically, we kept data points which provided BP information as well as information about the additional input parameters (time of measurement, heart rate, sex, age) the M-RMSE requires. We kept BP data points within 3 standard deviations of the mean (= 99.8%) to mitigate the effect of stray-bullet measurements. We reduced the MIMIC IV dataset from nearly 330 million data points to a systolic and diastolic dataset (> 2.3 million entries each).

B-Score interpretation. We calculated the B1-, B2- and M-RMSE for both the systolic and diastolic MIMIC IV dataset. We were then able to interpret the results from the dobutamine-dataset in respect to the MIMIC IV dataset. More specifically, we were able to calculate a T-RMSE value of equal B-Scores. Reaching this T-RMSE value (on the MIMIC IV dataset) coequals the system performance proposed by Patzak et al.[20]. The calculation is easily obtained by transposing the B-Score equation. It is available in the Supplementary Appendix and directly computed in the provided code (Supplementary Appendix 2/4).

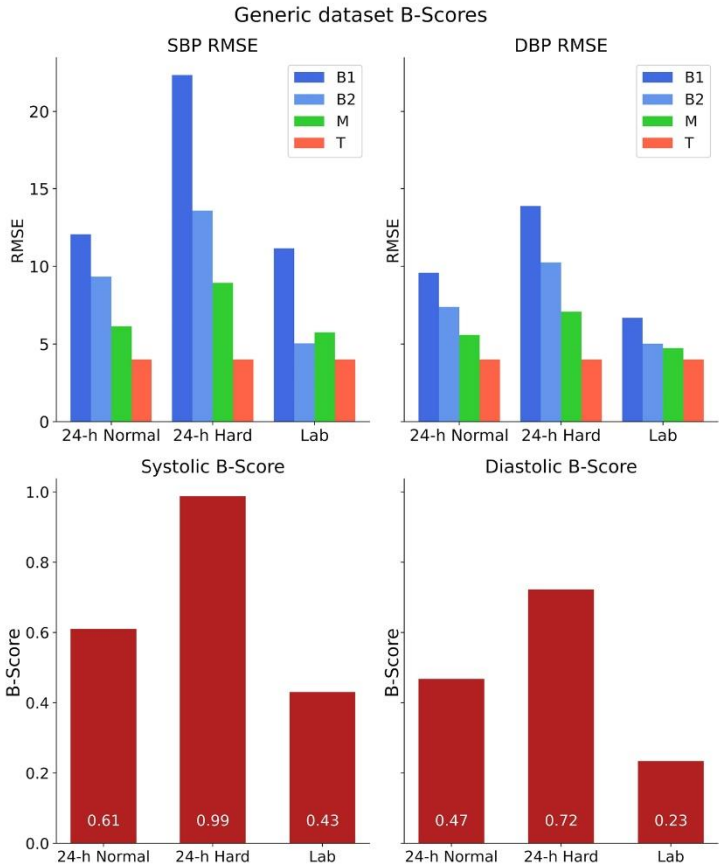


Figure 3. RMSE values (upper panel) and calculated B-Scores (lower panel) for all systolic (left) and diastolic (right) generated datasets. As expected, the "24-h Hard" dataset generated the highest B-Score for a fictional T-RMSE. The RMSE values indicate differences in dataset complexity ("Lab" easy to "24-h Hard" hard) which are reflected in the associated B-Scores.

Time complexity analysis. We split the MIMIC IV dataset into smaller subsets to analyse the time needed for B-Score calculations depending on the dataset size. The calculation was performed on a single core of an Intel i9 12900K CPU.

Programming packages and code. *Programming packages.* We wrote our programs in Python 3 (3.7.10) and primarily used the NumPy (1.19.5) and pandas (1.1.5) libraries for dataset cleaning and processing[29,30]. For model creation and evaluation, we relied on the Tensorflow2 (2.4.1) and sklearn (0.22.2) libraries[31,32]. We used the Matplotlib (3.2.2) library and NN-SVG for visualization[33,34].

Code. The code for B-Score calculation is provided as an Supplementary Appendix to this article (Supplementary Appendix 2).

Results

B-Score test with generic datasets. The created development datasets showed expected normal distributions, with highest variability in the "24-h Hard" and lowest in the "Lab" dataset (Supplementary Appendix 5).

We calculated the base performances (B1-, B2-, M-RMSE) values for each of the six datasets. We then calculated the B-Scores for all datasets with an assumed T-RMSE of 4.0 mmHg. Large differences between the T-RMSE and the base performances resulted in increased B-Scores. The B-Score scored highest for the "24-h Hard" dataset (as expected with constant T-RMSE; Fig. 3).

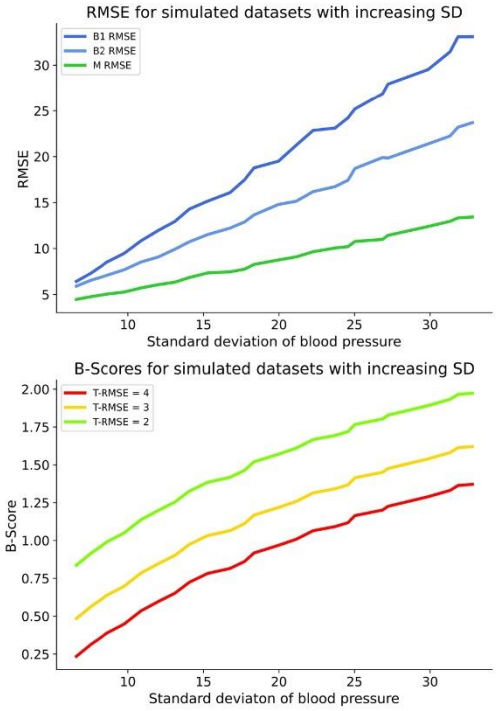


Figure 4. Base performances and B-Scores plotted against increasing standard deviations of simulated datasets. The upper panel shows rising base performance values (B1-, B2- and M-RMSE) for increasing standard deviation. Accordingly, the lower panel shows increasing B-Scores for constant T-RMSE values and increasing standard deviation. The B-Score discriminates between three tested T-RMSE values.

With the base-performance values calculated, we were able to calculate the B-Scores for each development dataset. Visual intuition about model performance based on base-performance values correlates with the calculated B-Scores (Fig. 3).

B-Score under increasing standard deviation. Simulating generic datasets with increasing BP standard deviation revealed a clear connection between increasing base performance values and increasing standard deviation. Subsequently, B-Scores rose with increasing BP standard deviation under constant T-RMSE values. The B-Score discriminated well between different tested T-RMSE values (Fig. 4).

Published small dataset (dobutamine). We calculated the B-Score for the "dobutamine"-dataset. For diastolic values, the proposed device's T-RMSE was larger than the M-RMSE. According to the B-Score's definition this results in a B-Score of < 0.00. The systolic B-Score was 0.94 (Fig. 5).

MIMIC IV dataset. We calculated the base performances (B1-, B2-, M-RMSE) for the systolic and diastolic MIMIC IV datasets (Fig. 6).

We used these base performance values to calculate the systolic and diastolic T-RMSE value. A new BP estimation would need to reach a systolic T-RMSE of 6.98 on the MIMIC IV dataset to perform coequally to the system proposed in the publication of the "dobutamine" dataset[20]. Smaller T-RMSE values would indicate a novel system outperforming the proposed device (Fig. 7).

We did not calculate a diastolic T-RMSE, as any T-RMSE smaller than the MIMIC IV B1-, B2- and M-RMSE values would be sufficient. Therefore, a T-RMSE smaller than the diastolic MIMIC IV M-RMSE (10.18, Fig. 5) outperforms the device tested on the "dobutamine" dataset.

Time complexity analysis. The time complexity analysis revealed a U-shaped dependency between dataset size and time needed for B-Score calculation. Calculation times were between 3 min for medium sized data-

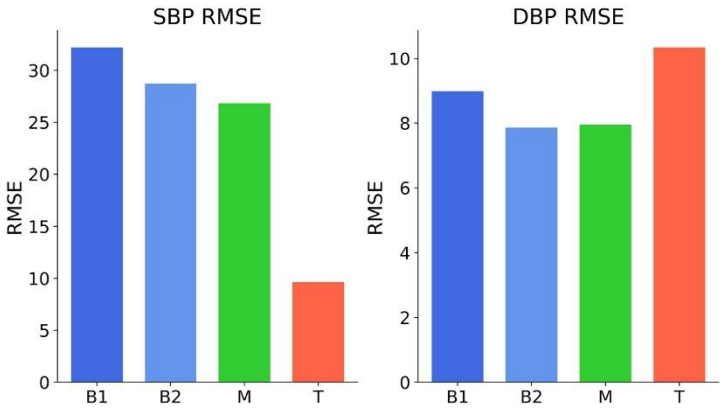


Figure 5. RMSE values calculated for the "dobutamine" dataset. The resulting systolic B-Score was 0.943. The diastolic B-Score is < 0.00 as the T-RMSE is not smaller than the M-RMSE. Diverging systolic and diastolic relative performances are apparent. The systolic RMSE values were B1 = 32.25, B2 = 28.71, M = 26.83, T = 9.64. The diastolic RMSE values were B1 = 8.98, B2 = 7.86, M = 7.95, T = 10.35.

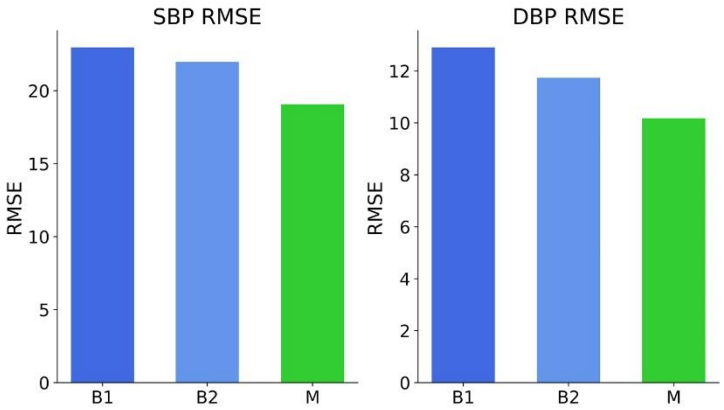


Figure 6. Base performances (B1-, B2-, M-RMSE) calculated for the systolic and diastolic MIMIC IV dataset. Systolic data variance is higher than diastolic. The M-RMSE shows a predictive benefit over the B1- and B2-RMSE for both systolic and diastolic values. The systolic RMSE values were B1 = 22.96, B2 = 21.97, M = 19.07. The diastolic RMSE values were B1 = 12.91, B2 = 11.74, M = 10.18.

sets (50,000 BP values) and 50 min for very small (250 BP values) and extremely large (2.3 million BP samples) (Fig. 8).

On multicore processors, systolic and diastolic B-Scores can be calculated simultaneously without noticeable time delay. Noticeable, GPU acceleration does negatively affect calculation times. The standard deviation of dataset reshuffles (which are averaged for M-RMSE calculation) was below 3 mmHg for all subsamples and below 1 mmHg for all samples with more than 1250 samples.

Discussion

The B-Score is a tool for comparing the relative performances of BP estimation systems. It sets measures of absolute model performance (regularly reported) in contrast to dataset specific parameters (base performances). It is based on the RMSE, combining insights about absolute error (cf. mean-absolute error) and measurement consistency (cf. correlation coefficient). The B-Score allows the comparison of performances between various systems tested on different datasets as higher B-Scores equal better relative performance.

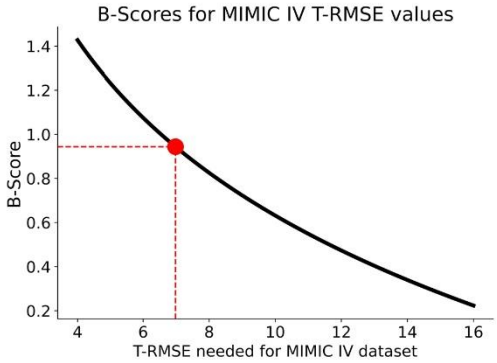


Figure 7. B-Scores calculated for various systolic MIMIC IV T-RMSE values (black line). The red dot indicates the T-RMSE which reaches a coequal performance to the device proposed by Patzak et al. The red dashed lines indicate the B-Score (0.96) and T-RMSE (6.98) of coequality.

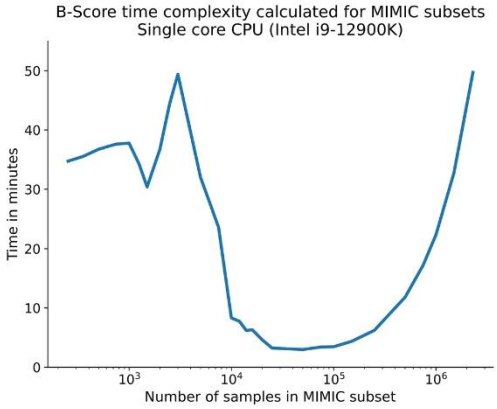


Figure 8. Time complexity analysis for B-Score calculation derived from MIMIC IV subsamples increasing in size. The x-axis is scaled logarithmically to allow visual interpretation.

We ensured reliable B-Score results and have shown its applicability by using generic datasets with differing inter- and intrapersonal BP variability. Further, the B-Score discriminated correctly between "easy" ("Lab") and "hard" ("24 h Hard") datasets, for a set fictional T-RMSE (= same absolute performance for all datasets). Further, the B-Score showed expected results for generic datasets with increasing BP standard deviation. It discriminated between different tested T-RMSE values for increasing base performance values.

To test the B-Score in real-world data, we used a small, published dataset with a proposed device for BP estimation ("dobutamine" dataset) and the to our knowledge largest available BP dataset ("MIMIC IV"). We calculated the B-Score for the "dobutamine" dataset and retrieved greatly differing results for systolic and diastolic performance. The B-Score revealed a markedly better systolic performance even though the absolute performance measures were largely the same between systolic and diastolic values. This illustrated the important additional information provided by a measure of relative performance (B-Score).

Further, we calculated the base performances (B1-, B2-, M-RMSE) for the MIMIC IV dataset. These parameters allowed us to calculate the T-RMSE value a new system would need to reach on the MIMIC IV dataset to provide coequal performance to the device tested by Patzak et al. ("dobutamine" dataset). The inter- and intrapersonal BP variability in the MIMIC IV dataset was smaller than in the "dobutamine" dataset. Consequently, the needed T-RMSE to reach coequal relative performance was lower than the one derived from the "dobutamine" dataset.

This analysis revealed important insights into the described datasets but more importantly proved that the B-Score is easily calculatable even in extreme (comparing very small vs. very large datasets) real-world use cases. This is further underlined by the quick calculation times which allows to derive the B-Score within one hour on

using a modern CPU. This is true for virtually all dataset sizes, with minimum calculation times for medium-sized datasets. The U-shaped time complexity curve is product of increasing calculation time per training episode for larger datasets and simultaneous reduction of repeated and reshuffled calculations due to increased trust in result reliability. We consider the resulting calculation times reasonable, especially noting that the calculation time needed only applies once per dataset, as the base performances can be used to recalculate the B-Score for any new model using the same dataset within seconds.

Additionally, time complexity analysis revealed narrow standard deviations between reshuffles, even for small datasets. This supports the assumption the B-Score calculation generates reliable and repeatable results.

We designed the B-Score to be intuitively understandable (higher B-Scores equal better relative performance) and easily calculable. Any researcher using a modern machine can calculate the B-Score for their data within 1 workday using the provided code (Supplementary Appendix 2).

As the B-Score is partly determined by inter- and intrapersonal BP variability within a given dataset, researchers aiming for high B-Scores are incentivized to develop their models for high variability datasets. This is important, because systems which perform well on heterogenous data are more likely to generalize to real-world applicability.

We envision investigators in the field of BP estimation devices to calculate base performance (B1-, B2-, M-RMSE) and B-Score values for their respective datasets and proposed systems. This will allow intuitive comparability between the plethora of systems available and under development.

We did not create the B-Score to replace validation studies and standardized validation protocols. These serve an important role in guaranteeing methodological comparability, which cannot be displaced by the B-Score.

We anticipate the B-Score to be an important tool for systems and devices which have not yet reached the stage of full-on clinical validation. It empowers researchers and engineers to quickly assess their system's relative performance on whichever dataset they have available. Researchers will be able to detect promising trends in the scientific literature more quickly and securely when B-Score are reported in the scientific literature during early stages of development. Further, the B-Score can become a tool for advanced, post-validation system testing. It allows to compare performances for distinct groups (e.g., pregnant women, children, etc.) or under special circumstances (e.g., sport) which are not covered by validation protocols.

Conclusion

The B-Score is a novel, functional measure of relative BP estimation performance. We proved its reliable results, ease of calculation even for large datasets and desired properties with generated datasets. Followingly, the B-Score revealed important insights in an extreme, real-world use case (comparing a very small vs. very large dataset).

The B-Score is easily interpreted and quickly calculated for any given dataset. We envision the B-Score to be used in pre-validation studies for system development and in advanced, post-validation analyses for special subgroups or measurement circumstances.

We hope that the B-Score will in the future become a useful and broadly applied tool for model performance comparison. It enables the quick and secure detection of promising trends in the scientific literature and allows scientists and engineers to quickly assess the performance of their systems.

Data availability

The "dobutamine" dataset is a re-analysis of already published data (Ref.[20]) and is available from the corresponding author of this publication upon reasonable request. The MIMIC IV dataset is available following the instructions from reference 25. The Code mentioned in the article is openly available in a public repository (Supplementary Appendix 2). For further information about data availability, please contact the corresponding author.

Received: 25 February 2022; Accepted: 12 July 2022
Published online: 16 July 2022

References

1. Whelton, P. K. *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ ASH/ASPC/NMA/PCNA guideline for the prevention, detection, evaluation, and management of high blood pressure in adults a report of the American College of Cardiology/American Heart Association Task Force on Clinical practice guidelines. *Hypertension* **71**, E13–E115 (2018).
2. Williams, B. *et al.* 2018 ESC/ESH guidelines for themanagement of arterial hypertension. *Eur. Heart J.* **39**, 3021–3104 (2018).
3. Hermida, R. C. *et al.* Ambulatory blood pressure monitoring recommendations for the diagnosis of adult hypertension, assessment of cardiovascular and other hypertension-associated risk, and attainment of therapeutic goals (summary). Joint recommendations from the International Society for Chronobiology (ISC), American Association of Medical Chronobiology and Chronotherapeutics (AAMCC), Spanish Society of Applied Ch. *Clinica e Investig. en Arterioscler.* **25**, 74–82 (2013).
4. Gijón-Conde, T. & Banegas, J. R. Use of ambulatory blood pressure monitoring. *Hipertension y Riesgo Vasc.* **34**, 15–18 (2017).
5. Gijón Conde, T. & Banegas, J. R. Ambulatory blood pressure monitoring for hypertension diagnosis? *Hipertension y Riesgo Vasc.* **34**, 4–9 (2017).
6. Bilo, G. *et al.* Validation of the Somnotouch-NIBP noninvasive continuous blood pressure monitor according to the European Society of Hypertension International Protocol revision 2010. *Blood Press. Monit.* **20**, 291–294 (2015).
7. Ding, X. R., Zhang, Y. T., Liu, J., Dai, W. X. & Tsang, H. K. Continuous cuffless blood pressure estimation using pulse transit time and photoplethysmogram intensity ratio. *IEEE Trans. Biomed. Eng.* **63**, 964–972 (2016).
8. Zheng, Y. L., Yan, B. P., Zhang, Y. T. & Poon, C. C. Y. An armband wearable device for overnight and cuff-less blood pressure measurement. *IEEE Trans. Biomed. Eng.* **61**, 2179–2186 (2014).
9. Agarwal, R. & Light, R. P. The effect of measuring ambulatory blood pressure on nighttime sleep and daytime activity—Implications for dipping. *Clin. J. Am. Soc. Nephrol.* **5**, 281–285 (2010).
10. Sherwood, A., Hill, L. K., Blumenthal, J. A. & Hinderliter, A. L. The effects of ambulatory blood pressure monitoring on sleep quality in men and women with hypertension: Dipper vs. nondipper and race differences. *Am. J. Hypertens.* **32**, 54–60 (2019).
11. Davies, R. J. O., Jenkins, N. E. & Stradling, J. R. Effect of measuring ambulatory blood pressure on sleep and on blood pressure during sleep. *BMJ* **308**, 820 (1994).

12. Mancia, G. & Parati, G. The role of blood pressure variability in end-organ damage. *J. Hypertension* **21**, S17 (2003).
13. Stevens, S. L. *et al.* Blood pressure variability and cardiovascular disease: Systematic review and meta-analysis. *BMJ* **354**, 4098 (2016).
14. Parati, G., Stergiou, G. S., Dolan, E. & Bilo, G. Blood pressure variability: Clinical relevance and application. *J. Clin. Hypertension* **20**, 1133–1137 (2018).
15. Beutel, F., van Hoof, C., Rottenberg, X., Reesink, K. & Hermeling, E. Pulse arrival time segmentation into cardiac and vascular intervals—Implications for pulse wave velocity and blood pressure estimation. *IEEE Trans. Biomed. Eng.* https://doi.org/10.1109/TBME.2021.3055154 (2021).
16. Ibrahim, B. & Jafari, R. Cuffless blood pressure monitoring from an array of wrist bio-impedance sensors using subject-specific regression models: Proof of concept. *IEEE Trans. Biomed. Circuits Syst.* https://doi.org/10.1109/TBCAS (2019).
17. Yan, C. *et al.* Novel deep convolutional neural network for cuff-less blood pressure measurement using ECG and PPG signals. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 1917–1920 (Institute of Electrical and Electronics Engineers Inc., 2019). https://doi.org/10.1109/EMBC.2019.8857108.
18. Jorge, J. *et al.* Machine learning approaches for improved continuous, non-occlusive arterial pressure monitoring using photoplethysmography. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Vol. 2020 (Institute of Electrical and Electronics Engineers Inc., 2020).
19. Shimazaki, S., Kawanaka, H., Ishikawa, H., Inoue, K. & Oguri, K. Cuffless blood pressure estimation from only the waveform of photoplethysmography using CNN. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Vol. 2019, 5042–5045 (Institute of Electrical and Electronics Engineers Inc., 2019).
20. Patzak, A., Mendoza, Y., Gesche, H. & Konermann, M. Continuous blood pressure measurement using the pulse transit time: Comparison to intra-arterial measurement. *Blood Press.* **24**, 217–221 (2015).
21. Lin, W. H., Wang, H., Samuel, O. W. & Li, G. Using a new PPG indicator to increase the accuracy of PTT-based continuous cuff-less blood pressure estimation. In *Proc. Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, Vol. 2017, 738–741 (Institute of Electrical and Electronics Engineers Inc., 2017).
22. Ding, X. *et al.* Pulse transit time based continuous cuffless blood pressure estimation: A new extension and a comprehensive evaluation. *Sci. Rep.* **7**, 3 (2017).
23. Wang, R., Jia, W., Mao, Z. H., Sclabassi, R. J. & Sun, M. Cuff-free blood pressure estimation using pulse transit time and heart rate. In *International Conference on Signal Processing Proceedings, ICSP*, Vol. 2015, 115–118 (Institute of Electrical and Electronics Engineers Inc., 2014).
24. Goldberger, A. L. *et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation* **101**, e215 (2000).
25. Johnson, A. *et al.* MIMIC-IV (version 1.0). *PhysioNet*. https://doi.org/10.13026/s6n6-xd98 (2021).
26. Pinsky, M. R. Cardiopulmonary interactions: Physiologic basis and clinical applications. *Ann. Am. Thorac. Soc.* **15**, S45–S48 (2018).
27. Douma, L. G. & Gumz, M. L. Circadian clock-mediated regulation of blood pressure. *Free Radic. Biol. Med.* **119**, 108–114 (2018).
28. Julien, C. The enigma of Mayer waves: Facts and models. *Cardiovasc. Res.* **70**, 12–21 (2006).
29. Harris, C. R. *et al.* Array programming with NumPy. *Nature* **585**, 357–362 (2020).
30. Mckinney, W. Data structures for statistical computing in python. In *Proc. of the 9th Python in Science Conf.* (eds. van der Walt, S. & Millman, J.) 56–61. https://doi.org/10.25080/Majora-92bf1922-00a (2010).
31. Abadi, M. *et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems* (2015).
32. Pedregosa, F. *et al.* Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825 (2011).
33. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
34. LeNail, A. NN-SVG: Publication-ready neural network architecture schematics. *J. Open Source Softw.* **4**, 747 (2019).

Author contributions

T.L.B. wrote the main manuscript text. All authors reviewed the manuscript and contributed to designing and conducting the analyses.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

T.L.B. and A.P. are advisors for SOMNOmedics on blood pressure measurement. N.P. declares no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-16527-2.

Correspondence and requests for materials should be addressed to T.L.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Curriculum Vitae

// My curriculum vitae is not published in this electronic record for data security purposes //

Publication list

Original articles

1. Pilz, N.; Heinz, V.; Parati, G.; Haberl, R.; Hofmann, E.; Küchler, G.; Patzak, A.; **Bothe, T.L.** Assessment of Nocturnal Blood Pressure: Importance of Determining the Time in Bed—A Pilot Study. J. Clin. Med. 2024, 13, 2170. doi: 10.3390/jcm13082170 IF: 3.900 (2022, Q2)
2. **Bothe TL**, Kreutz R, Glos M, Patzak A, Pilz N. Simultaneous 24-h ambulatory blood pressure measurement on both arms: a consideration for improving hypertension management. J Hypertens. 2023 Dec 7. doi: 10.1097/HJH.0000000000003632. IF: 4.900 (2022, Q2)
3. Wang H, **Bothe TL**, Deng C, Lv S, Khedkar PH, Kovacs R, Patzak A, Wu L. Comparison of Prognostic Models for Functional Outcome in Aneurysmal Subarachnoid Hemorrhage Based on Machine Learning. World Neurosurg. 2023 Oct 9:S1878-8750(23)01402-X. doi: 10.1016/j.wneu.2023.10.008. IF: 2.000 (2022, Q3)
4. **Bothe TL**, Hulpke-Wette M, Barbarics B, Patzak A, Pilz N. Accuracy of cuff-less, continuous, and non-invasive blood pressure measurement in 24-h ABPM in children aged 5-17. Blood Press. 2023 Dec;32(1):2255704. IF: 1.800 (2022, Q4)
5. Kagelmann N, Janke D, Maggioni MA, Gunga H-C, Riveros Rivera A, Genov M, Noppe A, Habazettl H, **Bothe TL**, Nordine M, Castiglioni P and Opatz O (2023) Peripheral skin cooling during hyper-gravity: hemodynamic reactions. Front. Physiol. 14:1173171. doi: 10.3389/fphys.2023.1173171. IF: 4.755 (2021, Q1)
6. Roth B, **Bothe TL**, Patzak A, Pilz N. Validation of the ABPMpro ambulatory blood pressure monitor in the general population according to AAMI/ESH/ISO Universal Standard (ISO 81060-2:2018). Blood Press Monit. 2023 Apr 5. IF: 1.430 (2021, Q4)
7. Pilz N, Patzak A, **Bothe TL**. The pre-ejection period is a highly stress dependent parameter of paramount importance for pulse-wave-velocity based applications. Front Cardiovasc Med. 2023 Feb 15;10:1138356. IF: 5.848 (2021, Q2)
8. **Bothe TL**, Bilo G, Parati G, Haberl R, Pilz N, Patzak A. Impact of oscillometric measurement artefacts in ambulatory blood pressure monitoring on estimates of average blood pressure and of its variability: a pilot study. J Hypertens. 2022 Oct 21; IF: 4.844 (2020, Q2)
9. **Bothe TL**, Patzak A, Pilz N. The B-Score is a novel metric for measuring the true performance of blood pressure estimation models. Scientific Reports 2022 12:1. 2022 Jul 16;12(1):1–10. IF: 4.380 (2020, Q1)

Review articles

1. **Bothe TL**, Gunga HC, Pilz N, Heinz V, Opatz OS. Relativistic aspects of physiology: Expanding our understanding of conventional control loops. Acta Physiol (Oxf). 2023 Dec;239(4):e14064. doi: 10.1111/apha.14064. IF: 6.4 (2022, Q1)
2. **Bothe TL**, Pilz N, Patzak A, Opatz OS. Bridging the gap: The dichotomy between measurement and reality in physiological research. Acta Physiol (Oxf). 2023 Jun 24:e14015. doi: 10.1111/apha.14015. IF: 7.532 (2021, Q1)
3. Pilz N, Patzak A, **Bothe TL**. Continuous cuffless and non-invasive measurement of arterial blood pressure—concepts and future perspectives. https://doi.org/101080/0803705120222128716. 2022 Dec 31;31(1):254–69. IF: 2.835 (2020, Q3)
4. **Bothe TL**, Pilz N, Dippel LJ. The compass of biomedicine. Acta Physiol 2022 Sep 1;236(1):e13856. IF: 6.311 (2020, Q1)
5. **Bothe TL**, Dippel LJ, Pilz N. The Art of Planning-How many samples are enough? Acta Physiol (Oxf). 2022 Feb 1;234(2). IF: 6.311 (2020, Q1)

6. **Bothe TL**, Patzak A, Schubert R, Pilz N. Getting it right matters! Covid-19 pandemic analogies to everyday life in medical sciences. Acta Physiol. 2021 Sep 1;233(1). IF: 6.311 (2020, Q1)
7. **Bothe, T.L.**, Patzak, A. Significant significance?. Acta Physiol (2021) 232: e13665. IF: 5.542 (2019, Q1)

Acknowledgments

Firstly, I would like to express my most cordial gratitude to Prof. Dr. med. Andreas Patzak without whom none of this work would have been possible. He incited my deep interest in human physiology and turned from lecturer to supervisor to co-researcher, who I now consider a close friend and role model, both in terms of scientific approach and character.

Similarly, I would like to thank Mr. Niklas Pilz, who turned from fellow student and friend into co-researcher and who now is an integral part of our scientific endeavours. I am especially glad that we were able to combine a close friendship with continued and meaningful scientific output.

My special thanks go to my parents who sparked my scientific interest and made me believe that a life led by curiosity is attainable even with a non-academic background.

Lastly, I owe my deepest gratitude to Ms. Laura Dippel who gave me the energy for and supported my dedication towards building scientific projects without taking a break from medical school with her continued emotional and intellectual support. I admire her strength of character and ambition which serve as model for approaching my work.