Received: 17 November 2021 | Revised: 27 February 2022 | Accepted: 1 March 2022

DOI: 10.1111/jch.14474

ORIGINAL ARTICLE

WILEY

A model for early failure prediction of blood pressure measurement devices in a stepped validation approach

Annina S. Vischer MD[1] | **Gilles Dutilh PhD**[2] | **Thenral Socrates MD**[1] | **Thilo Burkard MD**[1,3]

[1]Medical Outpatient Department and Hypertension Clinic, ESH Hypertension Centre of Excellence, University Hospital Basel, Basel, Switzerland

[2]Department of Clinical Research, University Hospital Basel, Basel, Switzerland

[3]Department of Cardiology, University Hospital Basel, Basel, Switzerland

Correspondence
Dr. Annina S. Vischer, MD, Medical Outpatient Department and Hypertension Clinic, University Hospital Basel, Petersgraben 4, CH-4031 Basel, Switzerland.
Email: annina.vischer@usb.ch

Abstract

Blood pressure monitoring (BPM) devices have to be validated according to strict international validation protocols. Each protocol requests a specific number of participants to be included. All protocols use vast amounts of resources, as three people have to be present for every measurement, making trials costly, especially when the manufacturer has no intention to execute a validation study, reflected in the low share of validated in the commercially available BPM devices. The aim of our study was to develop criteria, which could detect low accuracy devices that could not pass a validation protocol early in the course of the validation process. The 2010 European Society of Hypertension International Protocol (ESH-IP) and the Universal Standard for Validation of BPM devices (AAMI/ESH/ISO) were scrutinized for criteria which can be used for preclusion of passing. Based on this, we developed a fail model. We found that a BPM device cannot pass the ESH-IP protocol, if there are ≥27, 13, or 4 single measurements differing more than 5, 10, or 15 mmHg, respectively, from the reference. For the AAMI/ESH/ISO protocol, we developed a model, which calculates best-case standard deviations (SDs) to detect SDs which would prevent the passing of the protocol before its completion, making a stepwise validation process possible. In conclusion, we found that our model is able to predict failure of low-accuracy BPM devices early during a validation protocol if used in a stepwise-approach. This can be useful to keep costs of validation studies low and to enable investigator-initiated trials.

KEYWORDS
blood pressure measurement, fail criteria, home blood pressure measurement, validation protocol

1 | INTRODUCTION

Arterial hypertension (AHT) is the leading preventable cause of premature death worldwide, with almost a third of the world population affected.[1] To diagnose AHT it is crucial to measure the blood pressure (BP) accurately.[2] According to the guidelines, these measurements should be taken using a device which has been validated according to standardized conditions and protocols.[3] Additionally, we are

increasingly confronted with the occurrence of smart phone technology and wearable devices which claim to be able to measure BP, and therefore need to be properly assessed to prove reliability.[4]

There are several validation protocols for BP monitoring (BPM) devices, which have clear requirements consisting of a certain number of subjects and measurements to be included.[5–7] These protocols require copious amounts of time and staff, thus more efficient validation procedures would reduce the costs.[8] However, the authors of these validation protocols state, that a smaller sample size would decrease the study power and accuracy.[6,8]

To validate their BPM devices, however, is not always on the focus of manufacturers. This is demonstrated in the fact that the vast majority of commercially available BPM devices, for example, 82% of 278 upper-arm cuff BPM devices, 92% of 162 wrist-cuff BPM devices and, most strikingly, 100% of 532 wristband wearable BPM devices available online in Australia, did not undergo proper validation.[9] This situation may be supported by a system where regulatory authorities focus mainly on the physical safety features of BPM devices rather than accuracy and performance.[10] If there are so many manufacturers marketing their unvalidated devices, there is a large unmet need of stepped validation processes to enable investigator-initiated trials sorting out low-accuracy devices as long as there is no official claim by regulatory authorities to include accuracy into the regulatory clearance process.

Therefore, our aim was to define fail criteria which enable researchers to detect low accuracy devices unable to pass the validation protocols early in a stepped approach.

2 | METHODS

2.1 | 2010 European Society of Hypertension International Protocol for the validation of blood pressure measuring devices

2.1.1 | Pass criteria

The 2010 European Society of Hypertension International Protocol (ESH-IP) requires the inclusion of 33 participants with three measurement pairs each, resulting in 99 measurement pairs.[5] These measurement pairs are calculated from four measurements taken with the observer device (obsBP) and three measurements taken with the test device (TestBP).[5] The measurement pairs consist of the TestBP and the numerically nearer of the previous or next obsBP.[5] In the 2010 ESH-IP, there are two population criteria, called part 1.[5] First, two of the following criteria have to be fulfilled: the test device must deliver 73/99 measurement pairs with a difference ≤5 mmHg, 87/99 measurement pairs with a difference ≤10 mmHg, and 96/99 measurement pairs with a difference ≤15 mmHg.[5] Second, all of the following criteria have to be fulfilled: the test device must deliver at least 65/99 measurement pairs with a difference ≤5 mmHg, 81/99 measurement pairs with a difference ≤10 mmHg, and 93/99 measurement pairs with a difference ≤15 mmHg.[5] Furthermore, there are two criteria on an individual level, called part 2: ≥24/33 participants must have 2/3 measurement pairs with a difference of ≤5 mmHg (part 2.1), and maximally 3/33 participants with 0/3 measurement pairs with a difference of ≤5 mmHg (part 2.2).[5] A fail in any of these criteria results in an overall fail.

2.2 | Extrapolation of fail criteria

By deducting the number of minimally required measurements or participants within certain limits from the number of measurement pairs or participants, respectively, required in the protocol, we receive the maximum allowed number of measurements which can exceed these limits. For part 1.1, two out of three requirements need to be fulfilled. Therefore, if two or more of these requirements have more than the maximal number of measurements outside of the limits, the device cannot pass. For part 1.2, all requirements need to be fulfilled. Therefore, if in any one of these categories, the number of outlier measurements exceeds the maximal number allowed, the device cannot pass. The same rule applies for part 2.

For each category in part 1, we subtracted from the total number of measurement pairs required, that is, 99 pairs, the minimal number of measurement pairs required pass the specific category (for part 1.1: 73 pairs for ≤5 mmHg, 87 pairs ≤10 mmHg, 96 pairs for ≤15 mmHg, and for part 1.2: 65 pairs for ≤5 mmHg, 81 pairs ≤10 mmHg, 93 pairs for ≤15 mmHg) to receive the maximally allowed number of measurement pairs outside of these limits. To this, we added one to receive the minimal number of measurements pairs for the "fail criteria" for each category. For part 2.1, we took the number of participants required, that is, 33 participants, subtracted the minimal number of participants needed within the limits (ie, 24 participants) to pass this part to receive the maximal number of participants allowed outside the limits, and added one to receive the minimal number of participants outside the limits of this part to fail the device. For part 2.2, we simply added 1 to the maximum number of participants allowed in this category (ie, three participants), to receive the minimal number of participants in this category needed to fail the device. See also Table 1.

TABLE 1 Extrapolation of fail criteria for the ESH-IP protocol

Part 1	Category	>5 mmHg	>10 mmHg	>15 mmHg
	Two of three measurements	99-73+1	99-87+1	99-96+1
	All three measurements	99-65+1	99-81+1	99-93+1
Part 2	**Category**	**>1/3 > 5 mmHg**		**0/3 ≤ 5 mmHg**
	Number per participant	33-24+1		3 + 1

2.3 | Universal Standard for the validation of blood pressure measuring devices (AAMI/ESH/ISO)

2.3.1 | Pass criteria

The Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Universal Standard for validation of BPM devices is regarded as the mandatory international standard and requests inclusion of at least 85 participants providing 255 paired BP measurements.[6,7] The observer and test device measurements are taken alternatingly, four measurements by the observers, three by the test device (TestBP).[7] The reference BP (RefBP) is defined as the average of the previous and the succeeding observer BP.[7]

The AAMI/ESH/ISO Universal Standard states, that a device is considered acceptable, if its estimated probability of an error ≤10 mmHg is at least 85%; however, this is not labeled as a criterion.[6] The authors state two criteria that need to be fulfilled: (1) calculating from all individual 255 pairs of measurements (TestBP – RefBP), the mean must be ≤5 mmHg and its SD ≤ 8 mmHg for systolic and diastolic BP. (2) Furthermore, for each of the 85 participants, the mean of the BP differences must be calculated. The maximal allowed SD of these averaged BP differences depends on their mean, and ranges from 4.79 mmHg for a mean of 5 to 6.95 mmHg for a mean of 0 mmHg.[7]

2.4 | Extrapolation of fail criteria and model development

The probability of an error ≤10 mmHg being at least 85% would conclude, that we should not see more than 39 measurements (15% of 255 measurements) with an error > 10 mmHg. However, the AAMI/ESH/ISO Universal Standard does not clearly state, if this would automatically lead to a failure for this device.[6] Therefore, we need to scrutinize the other criteria for potential fail criteria.

Both criteria state maximally allowed means.[7] The mean, that is, the arithmetic mean, is defined as the sum of the values in question divided by the number of values.[11] To predict the mean which will result from including the required number of measurements from a lower number of measurements is not possible, as positive and negative differences between TestBP and RefBP may cancel themselves out. Therefore, the mean is not predictable enough to use it as a fail criterion when including fewer than the required measurements.

Both criteria also state maximally allowed SD.[7] The uncorrected SD (σ) is defined as[12]

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(xi - \bar{x})^2} \quad (1)$$

Translated for not-mathematicians, this means that the variance (σ^2) is the mean of the squared deviation of each data point from the mean, divided by the number of samples. The square root of the variance is the SD. A TestBP – RefBP difference which is equal to the mean over all TestBP – RefBP differences will result in 0 for ((TestBP – RefBP) – mean)^2, whereas all other TestBP – RefBP differences will result in positive numbers for ((TestBP – RefBP) – mean)^2.

Therefore, a best-case SD from fewer than the required number of measurement pairs can be estimated by assuming means which would pass the requirements, the final number of required measurements and the so far truly observed differences between TestBP and RefBP. For this, we have to assume, that all so far missing TestBP-RefBP differences to complete the protocol would be equal to the assumed mean and, consequently, perfect matches.

Accordingly, we repeatedly calculate a mock best-case SD using different means between -5 and 5 mmHg. 5 mmHg is stated as the maximum mean difference for passing; however, it is not stated if this an absolute value or not.[7] This means, that from every observed TestBP-RefBP difference, we subtract an assumed mean (always the same for each estimation), square the result and add all the squared numbers. This result is divided by the number of required measurements for the full protocol, that is, 255 for criterion 1 and 85 for criterion 2. From the result, the square root is taken. The full formula is reported in the results section and as Appendix S1.

2.5 | Simulation using Fail Criterion 1

In order to test our fail model and stopping rule, we created datasets comprising 255 random numbers defined by a mean between -10 and 10 and a SD between 0 and 16, simulating the results from full validation protocols with realistic deviations of a TestBP from the RefBP. We then simulated 1000 studies, which randomly picked a limited selection (n = 80, 160, or 255) from these full datasets and calculated the best-case SDs using our fail model. With the results of this, we built graphs which code in color how many simulations showed any or all best-case SDs for all assumed means between -5 and 5 mmHg below the SD threshold of 8 mmHg.

3 | STATISTICS

All calculations were completed using R version 4.0.4.[13] For the graphs in the simulation, we used the packages Tidyverse and Viridis.[14,15]

4 | RESULTS

4.1 | Fail criteria for the 2010 European Society of Hypertension International Protocol for the validation of blood pressure measuring devices

For part 1, the population level of the 2010 ESH-IP, the "fail criteria" are depicted in Table 2. If in two or more categories, the number of observed measurement pairs with a difference of > 5, > 10, or > 15 mmHg from the obsBP is equal to or larger than the number stated for part 1.1 in Table 2, the device cannot pass. If there are at least

TABLE 2 Fail criteria for the 2010 ESH International Protocol if fewer than 33 participants with 99 measurement pairs are included

	>5 mmHg	>10 mmHg	>15 mmHg
Part 1.1: Two of	27	13	4
Part 1.2: Either	35	19	7

For example, If a test device delivers four measurements with a difference of > 15 mmHg and 13 measurements with a difference of > 10 mmHg to the obsBP value, it cannot pass the 2010 ESH-IP validation criteria. Also, if, for example the test device delivers seven measurements with a difference of >15 mmHg to the obsBP, it cannot pass the ESH-IP validation criteria.

as many measurements as stated for part 1.2 in Table 2 with a difference of > 5, > 10, or > 15 mmHg from the obsBP, the device cannot pass.

For part 2, the individual level, the test device cannot pass the 2010 ESH-IP validation criteria, once ≥4 participants have 0 measurements with a difference of ≤5 mmHg to the obsBP, or ≥10 participants have only 2/3 measurements or less with a difference of ≤5 mmHg to the RefBP.

4.1.1 | Stopping rule

The validation protocol could be stopped once the count of outliers reaches the numbers stated in Table 2. For example, if a device delivers 27 measurements with a difference >5 mmHg and four measurements with a difference of >15 mmHg to the obsBP value, it cannot pass the validation criteria and the protocol could be stopped. Also, if, for example, the test device delivers 19 measurements with a difference of >10 mmHg to the obsBP, it cannot pass the ESH-IP validation criteria and the protocol could be stopped.

4.2 | Fail criteria and model for the Universal Standard for the validation of blood pressure measuring devices (AAMI/ESH/ISO)

4.2.1 | Criterion 1

The maximum allowed SD of the mean difference between TestBP and RefBP is 8 mmHg when all required 255 measurement pairs are included.[7] To get a sequentially updatable best-case SD if less than 255 measurement pairs are included, we keep on using 255 in the denominator, but include only those measurement pairs observed so far. This can be expressed in the following formula:

$$SD\_best\_case\_crit1 = \sqrt{\frac{1}{255}\sum_{i=1}^{N}(xi - \bar{x})^2} \quad (2)$$

In this formula, xi are the observed differences between TestBP – RefBP, and $\bar{x}$ corresponds to the assumed mean. It is relatively time-consuming to calculate this by hand, therefore, we developed an R code to easily analyze the observed values at regular intervals. Assumed_means stores a grid of means ranging from -5 to 5 (which would allow passing of the device), n_req is the number of measurements required for the full protocol (ie, 255 for criterion 1), and obs_crit1 stores the list of observed TestBP – RefBP differences. The full code, including an example and further guidance, is available as Appendix 1. Furthermore, if R is not available, a web-based application using our model can be found on dkfbasel.shinyapps.io/bpmvalidationfailprediction.

```
SDBestCase <- function(x, assumed_means, crit = 1){
  if (crit == 1){
      n_req = 255
  }
  if (crit == 2){
      n_req = 85
  }
  squared_deviances <- outer(x, assumed_means, '-')^2
  min_SDs <- apply(squared_deviances, 2,
 function(x){sqrt(sum(x)/n_req)})
  names(min_SDs) <- assumed_means
 return(min_SDs)
  }
SD_best_case_crit1 <- SDBestCase(x = obs_crit1,
                  assumed_means = assumed_means,
                  crit = 1)
```

The resulting minimal SD can either be printed in a table, or, for visual analysis, be plotted in a graph. Our code results in a graph with a red line for the maximum allowed SD for the full protocol, and black dots for the minimal SD reachable with the observed TestBP – RefBP differences. Three examples from 120 random numbers (as the observed TestBP – RefBP differences) can be found in Figure 1. The example in Figure 1, Panel A mimics a device which at this stage shows no predictors of failure. In this case, the protocol should be continued. Figure 1, Panel B shows an example of a BP device with an intermediate accuracy. The best-case SDs exceed the threshold when assuming some, but not all means. Therefore, the protocol should be continued and the fail model repeated after the next few participants. Figure 1, Panel C shows an example of a low accuracy device. All best-case SDs exceed the threshold of 8 mmHg and therefore, the validation protocol should be stopped, as this device is a certain failure.

Stopping Rule Criterion 1

If all best-case SDs exceed the threshold of 8 mmHg, the validation protocol cannot be passed and should therefore be stopped.

4.2.2 | Criterion 2

The concept of validation criterion 2 is similar. The mean is equally difficult to estimate in advance; however, we can use the same model as

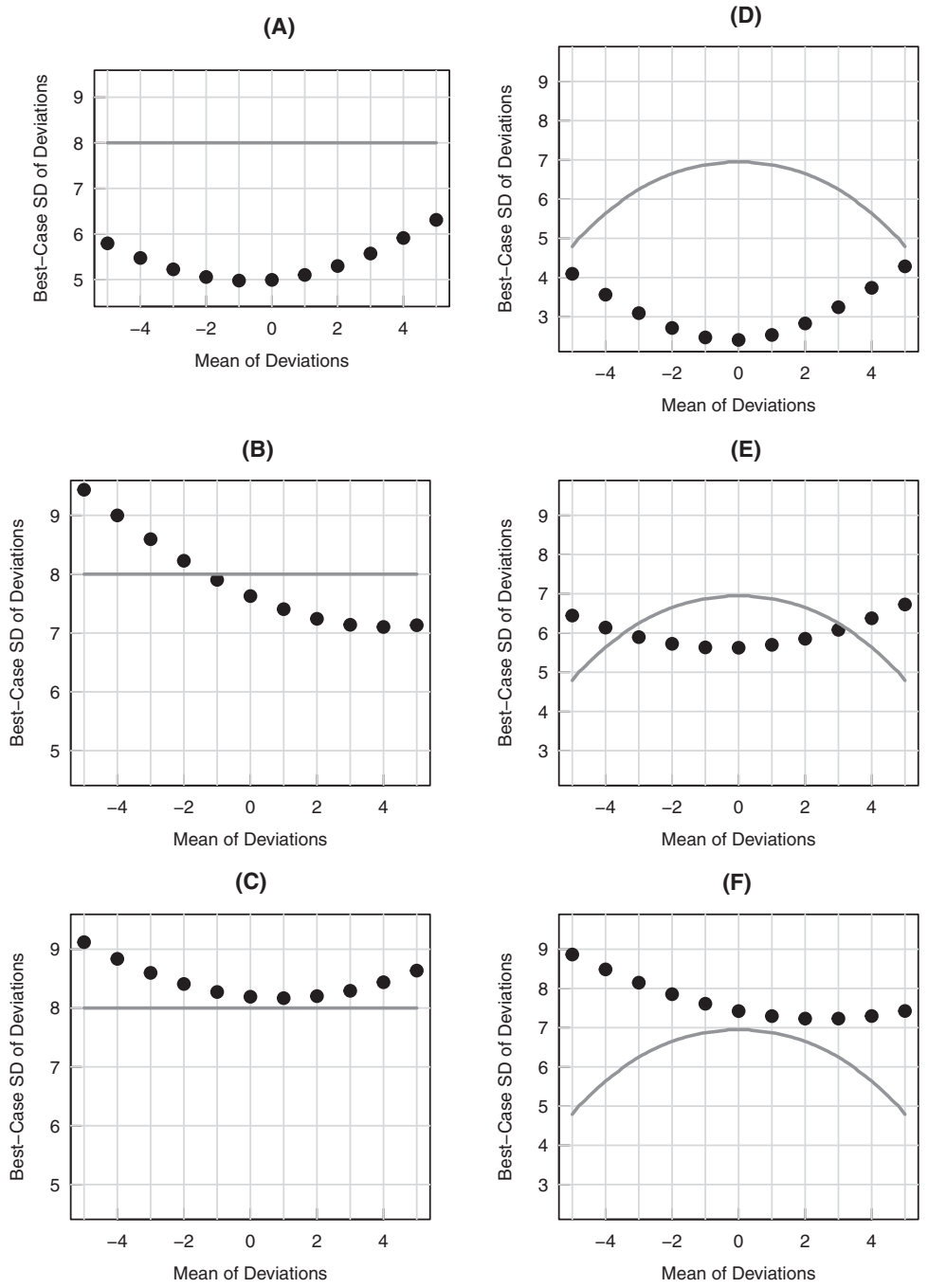

FIGURE 1 Examples of the results from hypothetical TestBP-devices by using the fail model for criterion 1 (panels A, B, C) and criterion 2 (panels D, E, F) using 120 random numbers for criterion 1 and 40 random numbers for criterion 2. Panels A, D: potentially passing device. Panels B, E: intermediate accuracy device. Panels C, F: Failing device. Black dots: best-case SDs calculated with the fail model. Red line: maximum permissible SD

for criterion 1, but use 85 instead of 255 as the denominator:

$$SD\_best\_case\_crit2 = \sqrt{\frac{1}{85}\sum_{i=1}^{N}(xi - \bar{x})^2} \quad (3)$$

In this formula, xi is the mean observed differences between TestBP – RefBP per participants, and $\bar{x}$ corresponds to the assumed mean. Again, we developed an R code to easily analyze the observed values at regular intervals. Assumed_means stores a grid of means ranging from -5 to 5 (which would allow passing of the device), n_req is the number of measurements required for the full protocol (ie, 85 for criterion 2), and obs_crit2 stores the list of the mean observed TestBP – RefBP differences per participant. The full code, including an example, is available as Appendix1. Furthermore, if R is not available, a web-based application using our model can be found on dkfbasel.shinyapps.io/bpmvalidationfailprediction.

```
SDBestCase < - function(x, assumed_means, crit = 1){
    if (crit = = 1){
        n_req = 255
    }
    if (crit = = 2){
        n_req = 85
    }
    squared_deviances < - outer(x, assumed_means, '-')^2
    min_SDs < - apply(squared_deviances, 2, function(x){sqrt(sum(x)/
            n_req)})
    names(min_SDs) < - assumed_means
    return(min_SDs)
}
SD_best_case_crit2 < - SDBestCase(x = obs_crit2,
                    assumed_means = assumed_means,
                    crit = 2)
```

Three examples from 40 random numbers (serving as the observed mean TestBP – RefBP differences per participant) can be found in Figure 1. The example in Figure 1, Panel D mimicks a device which at this stage shows no signs of failure. In this case, the protocol should be continued. Figure 1, Panel E shows an example of a BP device with an intermediate accuracy. The best-case SDs exceed the threshold when assuming some, but not all means. Therefore, the protocol should be continued and the fail model repeated after the next few participants. Figure 1, Panel F shows an example of a low accuracy device. All best-case SDs exceed the threshold and therefore, the validation protocol should be stopped, as this device is a certain failure.

4.2.3 | Stopping Rule Criterion 2

If all best-case SDs exceed the predefined threshold for each mean, in our plot marked by the red line, the validation protocol cannot be passed and should therefore be stopped.

4.3 | Simulation to Test the Fail Rule for Criterion 1

The 1000 simulations on a random selection of measurements from defined datasets each consisting of 255 measurements showed that no datasets which would pass the full validation protocol were predicted to fail by using our fail model (Figure 2). On the other hand, they showed that the more measurements were included, the more likely it would be that at least one best-case SD would be above the threshold (Figure 3). This means that, especially in cases with a true mean at the extreme end of the allowed means, our model could result in some (but not all) base-case SDs above threshold. Therefore, a single best-case SD above threshold does not rule out that a device passes the validation protocol. The full code for this simulation is available as Appendix S2.

5 | DISCUSSION

We showed that it is possible to predict from fewer than the required number of measurements for BPM validation protocols whether a device in question will certainly fail the full protocol, or has a potential to pass to full protocol. For the ESH-IP protocol,[5] an absolute number of measurements outside certain limits are enough to predict definite failure. For the AAMI/ESH/ISO protocol,[6,7] which serves as the current mandatory international standard, a model to predict best-case SDs is needed. All protocol requirements such as sex, blood pressure category, and arm circumference distribution must be followed strictly. The model we present is useful to stop this BPM device validation protocol prematurely, if, after partial completion of the protocol, all best-case SDs are above the threshold required to pass the protocol. This process aids in the detection of low accuracy BPM devices even before the completion of a full validation protocol.

By using our fail model on randomly picked numbers from simulated datasets with a defined mean and SD, we showed that the more datapoints are included, the more likely it is that, especially for datasets with a true mean at the extremes of the permissible means, at least one best-case SD would be above the threshold. Therefore, solely because one best-case SD from one assumed mean is above threshold does not preclude the device from passing the validation protocol. On the other hand, our simulation showed that no datasets with true means and SD within the requested limits would show all best-case SDs above threshold using our model.

We offer an R code that allows for easy calculation of the best-case SDs as interim analyses to detect low-accuracy devices before completion of the full protocol. In practical means, this translates, that during the completion of a validation protocol, the results received can be checked with the fail model after every few patients if the device still has a potential to pass the protocol or not. We recommend starting to test after as few as 25 or 30 participants, as this number may be sufficient to preclude passing of the full validation protocol in very low-accuracy devices, and continuing with the protocol and repeating the procedure after next few, for example, 5 or 10, participants, as long as the best-case SDs are lower than the threshold. If these interim analyses show that all best-case SDs are higher than the threshold defined by the protocol, we can stop the study early and mark the device as failed. The less accurate the TestBP is, the lower the number of examinations needed to predict its failure. If an interim analysis shows some, but not all best-case SDs above the threshold, the true arithmetic mean of the dataset can be taken into account. If this is in the range where the SD is above threshold, it is more likely that the device would fail the protocol. However, to prevent unfair stopping of a validation protocol, we recommend stopping the protocol only once all best-case SDs exceed the threshold. This also prevents unfair stopping in case that one or two additional participants are needed to fulfill the requirements of the protocol, such as BP category or cuff size. To avoid a confirmation bias, we suggest having these interim analyses completed by an independent monitoring person and keeping the study personnel performing the measurements blinded for the results.

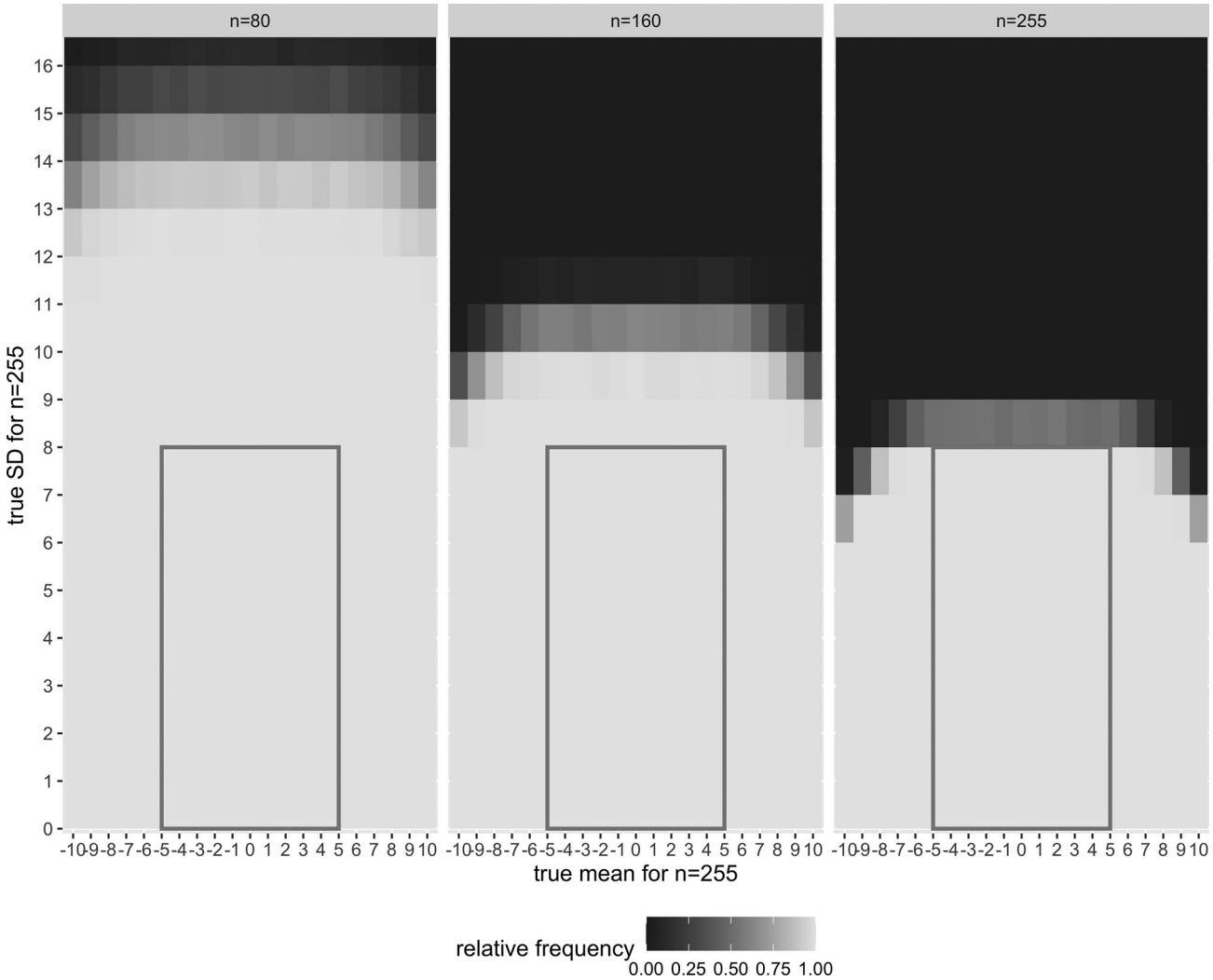

FIGURE 2 Heat map showing the frequency of **any** minimal SDs below the threshold for assumed means between -5 and 5 mmHg calculated with the fail model for criterion 1 in 1000 simulations using 80, 160, and 255 randomly selected numbers from random datasets with a defined mean of -5 to 5 mmHg and SD between 0 and 16 mmHg over 255 "measurements". On the x axis are the true means for the full dataset with $n = 255$. On the y axis are the true SDs for the full dataset with $n = 255$. Yellow represents at least one SD below threshold for all simulations, purple represents no SD below threshold in all simulations. The red box encompasses the area, in which would pass the protocol using all 255 "measurements"

A stepwise approach has been recommended by the British Hypertension Society validation protocol in 1990, to keep the costs for validation as low as possible.[16] The costs of validation protocols have not properly been analyzed. However, all commonly applied protocols request two observers using a double stethoscope and one supervisor in the room for all measurements, therefore, although material costs are relatively low, the expenses for the staff (and their time) are relatively high especially considering the number of participants necessary and the strict in- and exclusion criteria, some of which can only be assessed during the validation process.[5–7] The relatively high number of participants necessary for each protocol has been chosen to improve the statistical power of the validation protocol, considering that most BP monitors currently available are a moderate accuracy level.[6,8] However, new technologies are likely to bring new possibilities for BP measurements, such as cuffless measurements. Though the current protocols are formally not appropriate for continuous or cuffless measurement devices, it is absolutely necessary that such protocols are developed in the near future. For the validation of such devices, which have a high risk of being of low accuracy, our fail model and code may be helpful to keep costs for validation at a minimum. Given the high number of unvalidated devices on the market, our stepped approach may enable investigator-initiated trials to detect devices with low accuracy early and help in the development of validation protocols for cuffless devices. With our fail model, we hope to motivate and enable universities, research groups and clinicians to initiate validation studies for BPM devices to improve the prevalence of validated BPM devices. BP values obtained with a device in which the validation protocol was stopped due to our fail model should not be used for clinical decision-making.

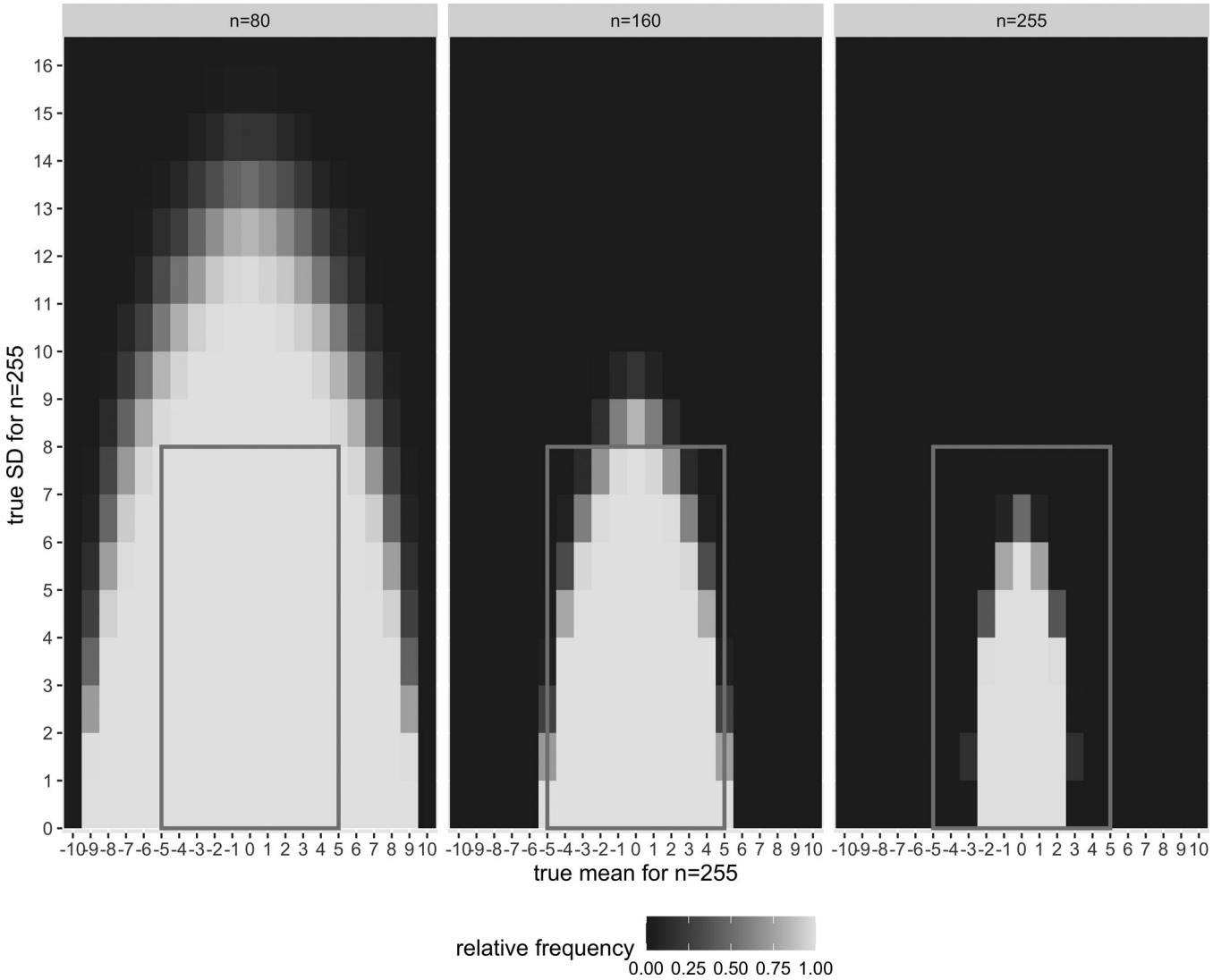

FIGURE 3 Heat map showing the frequency of **all** minimal SDs below the threshold for assumed means between -5 and 5 mmHg calculated with the fail model for criterion 1 in 1000 simulations using 80, 160, and 255 randomly selected numbers from random datasets with a defined mean of -5 to 5 mmHg and SD between 0 and 16 over 255 "measurements". On the x axis are the true means for the full dataset with n = 255. On the y axis are the true SDs for the full dataset with $n = 255$. Yellow represents all SD below threshold for all simulations, purple represents at least one SD above threshold in all simulations. The red box encompasses the area, in which would pass the protocol using all 255 "measurements"

6 | LIMITATIONS

The fail model for the AAMI/ESH/ISO results always in an estimation, since we assume that the true mean lies within -5 and 5 mmHg. There is no guarantee, that a device which does not meet up with these fail criteria, will pass the AAMI/ESH/ISO standard. However, if a device meets the fail criteria, it is not possible for this device to pass the AAMI/ESH/ISO standard by adding more measurements.

7 | CONCLUSIONS

With our fail model, we show that it is possible to safely detect low accuracy BP monitors before completing a conventional full validation protocol. Our analysis can be repeated at regular intervals as a step-wise approach. This is important to keep costs for validation studies at a minimum.

ACKNOWLEDGEMENTS

Many thanks to Dr. Constantin Sluka, Data Scientist at the Department of Clinical Research at the University Hospital in Basel, Switzerland, for programming the web-based application.

CONFLICTS OF INTEREST

None.

AUTHOR CONTRIBUTIONS

Annina S. Vischer contributed to the conception and design of the study; development of the formula; coding of the model; drafted the manuscript; critically revised the manuscript; and gave final approval

for the manuscript. Gilles Dutilh contributed to the conception and design of the study; the final coding of the model and the validation simulation; critically revised the manuscript; and gave final approval for the manuscript. Thenral Socrates contributed to the conception and design of the study; drafted the manuscript, critically revised the manuscript; and gave final approval for the manuscript. Thilo Burkard contributed to the conception and design of the study; critically revised the manuscript; and gave final approval for the manuscript.

ORCID

Annina S. Vischer MD https://orcid.org/0000-0001-5188-1723

REFERENCES

1. Mills KT, Bundy JD, Kelly TN, et al. Global disparities of hypertension prevalence and control: a systematic analysis of population-based studies from 90 countries. *Circulation*. 2016;134(6):441-450.
2. Vischer AS, Burkard T. How should we measure and deal with office blood pressure in 2021?. *Diagnostics (Basel)*. 2021;11(2).
3. Williams B, Mancia G, Spiering W, et al. 2018 ESC/ESH Guidelines for the management of arterial hypertension. *Eur Heart J*. 2018;39(33):3021-3104.
4. Kario K. Management of hypertension in the digital era: small wearable monitoring devices for remote blood pressure monitoring. *Hypertension*. 2020;76(3):640-650.
5. O'Brien E, Atkins N, Stergiou G, et al. European Society of Hypertension International Protocol revision 2010 for the validation of blood pressure measuring devices in adults. *Blood Press Monit*. 2010;15(1):23-38.
6. Stergiou GS, Alpert B, Mieke S, et al. A universal standard for the validation of blood pressure measuring devices: Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO) Collaboration Statement. *Hypertension*. 2018;71(3):368-374.
7. Stergiou GS, Palatini P, Asmar R, et al. Recommendations and practical guidance for performing and reporting validation studies according to the Universal Standard for the validation of blood pressure measuring devices by the Association for the Advancement of Medical Instrumentation/European Society of Hypertension/International Organization for Standardization (AAMI/ESH/ISO). *J Hypertens*. 2019;37(3):459-466.
8. Friedman BA, Alpert BS, Osborn D, Prisant LM, Quinn DE, Seller J. Assessment of the validation of blood pressure monitors: a statistical reappraisal. *Blood Press Monit*. 2008;13(4):187-191.
9. Picone DS, Deshpande RA, Schultz MG, et al. Nonvalidated home blood pressure devices dominate the online marketplace in Australia: major implications for cardiovascular risk management. *Hypertension*. 2020;75(6):1593-1599.
10. Sharman JE, O'Brien E, Alpert B, et al. Lancet Commission on Hypertension group position statement on the global improvement of accuracy standards for devices that measure blood pressure. *J Hypertens*. 2020;38(1):21-29.
11. Weissstein EW, "Mean." From MathWorld—A Wolfram Web Resource. https://mathworld.wolfram.com/Mean.html Accessed February 27, 2021
12. Weissstein EW, "Standard Deviation." From MathWorld—A Wolfram Web Resource. https://mathworld.wolfram.com/StandardDeviation.html Accessed February 7, 2021
13. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2019. https://www.R-project.org/
14. Wickham H, Averick M, Bryan J, et al. Welcome to the Tidyverse. *Journal of Open Source Software*. 2019;4:(43):1686.
15. *Rvision – Colorblind-Friendly Color Maps for R*. Version R package version 0.6.1. 2021.
16. O'Brien E, Petrie J, Littler W, et al. The British Hypertension Society protocol for the evaluation of automated and semi-automated blood pressure measuring devices with special reference to ambulatory systems. *J Hypertens*. 1990;8(7):607-619.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Vischer AS, Dutilh G, Socrates T, Burkard T. A model for early failure prediction of blood pressure measurement devices in a stepped validation approach. *J Clin Hypertens*. 2022;24:582–590. https://doi.org/10.1111/jch.14474