DISSERTATION

Is the Charité Alarm Fatigue Questionnaire Construct Valid? An Examination Using Confirmatory Factor Analysis.

Ist der Charité Alarm Fatigue Fragebogen konstruktvalide? Eine Überprüfung mittels konfirmatorischer Faktorenanalyse.

zur Erlangung des akademischen Grades
Doctor rerum medicinalium (Dr. rer. medic.)

vorgelegt der Medizinischen Fakultät
Charité – Universitätsmedizin Berlin

von

Maximilian Markus Wunderlich

Erstbetreuung: Prof. Dr. med. Dr. rer. nat. Felix Balzer

Datum der Promotion: 29.11.2024

Inhaltsverzeichnis

Tabellenverzeichnis

Abbildungsverzeichnis

Abkürzungsverzeichnis

AF: Alarm Fatigue

AFC: Alarm-Fatigue-Score

CAFQa: Charité Alarm Fatigue Questionnaire

CI: Confidence Interval

CFA: Confirmatory Factor Analysis

CFI: Comparative Fit Index

DF: Degrees of freedom

DOI: Digital Object Identifier

e.g.: for example

et al.: and others

F: Factor

HTF: Healthcare Technology Foundation

ICU: Intensive Care Unit

i.e.: that is

KMO: Kaiser-Meyer-Olkin

Kurt.: Kurtosis

RNI: Relative Non-Centrality Index

RQ: Research Question

RMSEA: Root Mean Square Error of Approximation

SD: Standard Deviation

SRMR: Standardized Root Mean Squared Residual

TLI: Tucker-Lewis Index

ULS: Unweighted least-squares

Zusammenfassung

Hintergrund

Pflegefachkräfte und Ärzt:innen können „Alarmmüdigkeit“ (“Alarm Fatigue”; AF) entwickeln und so gegenüber Alarmen desensibilisiert sein. AF ist ein lange bekanntes Problem, jedoch konnte man sich bislang nicht auf eine einheitliche Methode zur Quantifizierung einigen. In einer vorangegangenen Arbeit entwickelten meine Kolleg:innen und ich einen Fragebogen mit neun Items: den Charité Alarm Fatigue Questionnaire (CAFQa). Wir postulierten, dass der CAFQa AF auf zwei assoziierten Skalen misst: der “Alarm-Stress-Skala” und der “Alarm-Bewältungs-Skala”.

Ziel

Diese Arbeit untersucht, ob sich die in der vorangegangenen Arbeit postulierte Zwei-Skalen-Struktur mittels einer konfirmatorischen Faktorenanalyse (CFA) in einer neuen Stichprobe bestätigen lässt. Dies würde auf die Konstruktvalidität des CAFQa hinweisen.

Methoden

Der CAFQa wurde als Online-Fragebogen an fünf großen deutschen Kliniken zwischen Oktober 2021 und Juli 2022 erhoben. Die CFA basierte auf dem “unweighted least squares”-Algorithmus und polychorischen Korrelationen. Teilnehmer:innen gaben auch eine Selbsteinschätzung ihrer AF sowie des Anteils falsch-positiver Alarme auf ihrer Station ab. Diese Daten dienten als Indikatoren konvergenter Validität, indem sie mit den durchschnittlichen Alarm-Fatigue-Scores (AFCs) der Teilnehmenden korreliert wurden. Cronbach’s Alpha und McDonald’s Omega dienten als Maß der internen Konsistenz.

Ergebnisse

Die Stichprobe umfasste N = 265 Personen (davon 56,6% Pflegefachkräfte und 35,8% Ärzt:innen). Der Chi-Quadrat-Test deutete auf eine schlechte Modellpassung hin ($\chi^2(26) = 44.932$, $p = 0.012$). Jedoch zeigten die alternativen Fit-Indizes eine gute Passung (SRMR = 0.052, RMSEA = 0.03, TLI = 0.985, CFI = 0.989, and RNI = 0.989). Die Ladungen der einzelnen Fragen waren statistisch signifikant (zwischen 0,35 und 0,73) und die Faktoren korrelierten untereinander (r = 0,4). Die Selbsteinschätzungen der Teilnehmenden korrelierten moderat mit den durchschnittlichen AFCs (r = 0,45); die

Schätzungen des Anteils falsch-positiver Alarme schwach (r = 0,3). Die interne Konsistenz ist mit Cronbach's alpha = 0,67 und McDonald's omega = 0,8 angezeigt.

Schlussfolgerungen

Das Zwei-Faktor-Modell konnte bestätigt werden. Eine explorative Modellmodifikation wäre theoretisch plausibel, würde aber die Generalisierbarkeit des Modells beeinträchtigen. Diese Ergebnisse deuten auf eine gute Konstruktvalidität des CAFQa hin. Eventuell misst die Alarm-Bewältungs-Skala jedoch ein Konstrukt, das zwar verwandt, aber nicht deckungsgleich mit AF ist. Mit dem CAFQa haben Alarmforscher:innen und Kliniker:innen eine Möglichkeit zur Quantifizierung der AF von Pflegefachkräften und Ärzt:innen. Bei der Entwicklung datengestützter Maßnahmen für das Alarmmanagement einer Station sollte der CAFQa routinemäßig erhoben werden.

Abstract

Background

When exposed to medical device alarms, nurses and physicians may develop alarm fatigue (AF) and become desensitized to alarms. AF is a long-recognized problem, but researchers have not yet agreed on a standardized method of measuring it. In previous work, my colleagues and I developed a nine-item questionnaire called Charité Alarm Fatigue Questionnaire (CAFQa). Based on exploratory analyses, we postulated that the CAFQa measures AF on two associated scales: the ‘alarm stress’ and the ‘alarm coping’ scale.

Aim

This work investigates whether the previously postulated two-scale structure can be confirmed in a new sample using confirmatory factor analysis (CFA). If so, this would indicate the CAFQa’s construct validity.

Methods

CAFQa data were collected as an online questionnaire in five large German hospitals between October 2021 and July 2022. The CFA was based on the unweighted least squares algorithm and polychoric correlations. Participants self-assessed their AF, as well as the proportion of false-positive alarms on their ward. These data were used as indicators of convergent validity by correlating them with participants' average alarm fatigue scores (AFCs). Cronbach's alpha and McDonald's omega served as measures of internal consistency.

Results

The sample included N = 265 subjects (of whom 56.6% were nurses and 35.8% were physicians). The chi-square test indicated poor model fit ($\chi^2(26) = 44.932$, $p = 0.012$). However, the alternative fit indices showed good fit (SRMR = 0.052, RMSEA = 0.03, TLI = 0.985, CFI = 0.989, and RNI = 0.989). The loadings of each question were statistically significant (ranging from 0.35 to 0.73) and the factors were correlated with each other (r = 0.4). Participants' self-ratings correlated moderately with mean AFCs (r = 0.45) and their estimates of the proportion of false-positive alarms correlated weakly with mean

AFCs (r = 0.3). Internal consistency is indicated with Cronbach's alpha = 0.67 and McDonald's omega = 0.8.

Conclusions

The two-factor model was confirmed. An exploratory model modification would be theoretically plausible but would negatively affect the generalisability of the model. These results suggest good construct validity of the CAFQa. However, it is possible that the alarm coping scale measures a construct that is related to but not congruent with AF. The CAFQa allows alarm researchers and clinicians to quantify AF in nurses and physicians. The CAFQa should be routinely collected when developing data-based alarm management measures for an ICU.

1 Introduction

1.1 Alarm fatigue in intensive care units

The alarm systems of patient monitoring devices in intensive care units (ICUs) are designed to alert ICU staff to events that could threaten patients' lives [1]. However, ICU staff are often overwhelmed by the sheer number of alarms. In a recent study, my colleagues and I counted 152.5 alarms per bed per day as the average of one ICU [2]. Other studies report even higher averages (e.g., Jones reports 771 alarms per bed per day as the average of one ICU [3]). Strikingly, most of these alarms are likely either false or do not require a medical response [4]. Exposure to a large number of alarms, of which many are false, can cause ICU staff to develop 'alarm fatigue' [5]. My colleagues and I provided the following definition of alarm fatigue in a previous publication [6]:

> "[Alarm fatigue is a] *sensory overload due to exposure to an excessive number of clinical alarms, which can lead to desensitisation and loss of competence in handling alarm-related procedures (such as dismissing alarms or adjusting monitoring thresholds). Alarm-fatigued ICU staff struggles to identify and prioritise clinical alarms efficiently*" (p. 2)

ICU nurses and physicians are professionals, which is why in the majority of cases, nothing happens where patients' lives are at risk. However, there remains a chance that a critical situation is being missed. In fact, Jones claims that in hundreds of cases, patients' deaths in the United States of America were due to alarms being missed or recognized with delay [3]. Even if no critical situation were ever to be missed, alarms are an omnipresent background noise that disturbs patients' sleep and stresses ICU staff [7–9].

1.2 Measuring the alarm fatigue of nurses and physicians

In an ideal world, there would be no alarm fatigue at all. In an almost ideal world, there would be a clear association between the observable alarm situation and staff's subjectively perceived alarm fatigue (e.g., the greater the number of alarms the stronger staff's alarm fatigue). Unfortunately, the situation is far from ideal, because there seems to be no such association [10,11]. This is also hinted at in a study by Sowan et al., who implemented a new alarm management routine on a 20-bed transplant/cardiac ICU [12].

Before and after the intervention, they collected alarm data and used a questionnaire to measure nurses' sentiment regarding alarms. While the number of alarms per patient per day decreased by roughly 32%, nurses' sentiment on alarms did not change. Wilken, Hüske-Kraus and Röhrig argue that measuring the number of alarms per patient bed per day is not an informative metric, because it does not inform how alarms are distributed across time and beds [10]. For example, if an ICU had 150 alarms per bed per day, these could be evenly distributed across 24 hours, or they could occur in random bursts of ten or twenty alarms at once. The latter distribution might be far more stressful for ICU staff.

Because alarm data alone cannot inform about ICU staff's alarm fatigue, researchers have tried using questionnaires. My colleagues and I argued in our previous work [6] that no gold standard exists for measuring alarm fatigue (citing Hüske-Kraus et al. and Lewandowska et al. [11,13]) despite two decades of research showing that it is a risk for patient safety and a burden on ICU staff (citing Wears and Perry [14]). We also outlined how previous studies often created their own, study-specific surveys to measure alarm fatigue by borrowing/modifying items from a survey of the Healthcare Technology Foundation (HTF). So far, the only systematic attempts at designing questionnaires were by Ashrafi et al. and by Torabizadeh et al. [15,16]. Both made an effort to gather questionnaire items from the scientific literature, to include nurses and physicians in expert panels to review a selection of items, and to collect data with their final questionnaire. Some studies have already started using the questionnaire by Torabizadeh et al. (see our previous work for an overview [6]). And recently, Rypicz et al. published a Polish translation of the questionnaire by Ashrafi et al. [17]. Clearly, there is a need among clinical alarm researchers to adapt and use one questionnaire as the gold standard for measuring alarm fatigue. This would free researchers from developing a new questionnaire for every intervention study and, importantly, it would allow comparing the effects of such intervention studies. My colleagues and I outlined three conditions that should be met by an alarm fatigue questionnaire before being accepted as a new standard [6]: First, the authors should be transparent about the language in which the questionnaire was developed, second, the questionnaire should be developed based on the best-practices of scale construction, and third, nurses and physicians should both be the target group. Unfortunately, neither Ashrafi et al. nor Torabizadeh et al. meet these three conditions.

1.3 The Charité Alarm Fatigue Questionnaire

My colleagues and I recently developed the Charité Alarm Fatigue Questionnaire (abbreviated CAFQa, pronounced like Franz Kafka's surname) [6]. The CAFQa aims to quantify the alarm fatigue of nurses and physicians. It consists of nine items and should not take more than five minutes to administer. Our aim was to adhere to the best practices of scale construction outlined in Boateng et al. [18]. First, we gathered a large pool of items from the scientific literature and asked experts in alarm fatigue to review each item regarding its relevance for measuring alarm fatigue. We then interviewed nurses and physicians to understand how they read and understood each item. At each step, we rephrased, added, or deleted items. Finally, we sent 27 items as an online survey to nurses and physicians of all ICUs in a large German University Hospital. Using exploratory factor analysis and other statistical measures we reduced the number of items to nine and identified two distinct, yet correlated scales: ‘alarm stress’ and ‘alarm coping’. The alarm stress scale captures items on the psychophysiological effects that alarms can have on ICU staff (e.g., headaches, confusion, and lack of motivation). The alarm coping scale captures items on systemic influences and the extent alarm management is practiced in a given ICU (e.g., the ward’s floor layout, whether procedural instructions for alarms exist, and whether alarm limits are customized to individual patients).

1.4 Gauging a questionnaire’s construct validity

Alarm fatigue is a hypothetical “construct” because there is no direct way of observing it [19]. On the other hand, the number of alarms that a patient’s cardiovascular condition causes in an ICU is not a construct, because one can hear, count, and even see them on the screens of the monitoring devices. Psychologists have developed methods to understand how well an instrument (such as a questionnaire) measures a given unobservable construct (for example, Cronbach and Meehl [20]). Construct validity refers to an instrument’s ability to truly measure what it was designed to measure.

Smith expands on Cronbach and Meehl’s theory and proposes that research on construct validity should start with a precise definition of the construct [19]. From that theory, it should then derive and test specific hypotheses. Finally, observations made during these tests can be used to argue in favor of the instrument’s construct validity or to revise the theory of the construct. In our previous work [6] we provided a precise definition of alarm

fatigue and conducted exploratory analyses that ultimately led us to hypothesize that the CAFQa measures alarm fatigue along two scales. Within Smith's Framework, the next step toward understanding CAFQ's construct validity is to test this hypothesis. A common way to do this is to gather questionnaire responses from a new sample and submit the data to confirmatory factor analysis (CFA) [18,21,22].

Confirmatory factor analysis differs from exploratory factor analysis in the sense that the factor model to be tested is pre-specified and fixed instead of exploratively discovered and malleable. Researchers simply test whether the model fits the data. To gauge model fit, they rely on the chi-square test, which if significant, indicates that the model does not fit the data. However, the chi-square test is sensitive to sample size and quickly becomes significant, so alternative fit indices have been developed [23]. If chi-square and alternative fit indices indicate that a model does not fit, researchers start exploring how the model can be modified to fit the data of the sample. Ideally, these modifications are theoretically plausible. However, a modified model would then need to be confirmed on another independent sample - just like the initial model of the exploratory factor analysis [22].

1.5 Aim of this thesis

With this thesis, I aimed to underpin the construct validity of the Charité Alarm Fatigue Questionnaire by testing whether our previously proposed factor structure (i.e., the two scales 'alarm stress' and 'alarm coping') could be confirmed on a different, independent sample using confirmatory factor analysis. The results and findings will be part of a forthcoming publication [24].

2 Methods

The methods and procedures described in this section are part of a forthcoming publication [24].

2.1 Ethics approval

The study was conducted in accordance with appropriate guidelines and regulations. All participants voluntarily consented to participate after being informed about the study and the Ethics Commission of the Charité – Universitätsmedizin Berlin granted their ethical approval (ethics application number: EA4/218/20).

2.2 Participants

Nurses and Physicians from nine ICUs across five major German hospitals were invited to fill out the questionnaire online using REDCap between October 2021 and July 2022. As an incentive for completing the questionnaire, participants were given the opportunity to enter a drawing for a €50 online shopping voucher. Each participant consented to the anonymous collection, analysis, and storage of their data.

2.3 Questionnaire

The questionnaire used in this study consisted of all nine items from the CAFQa ([6]; 1-9 in Table 1), five general questions on how alarms are perceived (10-14 in Table 1), two questions serving as criteria for evaluating the questionnaire (15 and 16 in Table 1), and demographic questions (17-19 in Table 1).

My colleagues and I arranged all the CAFQa's items and all five general questions pseudo-randomly. Responses had to be given on a Likert scale ranging from -2 (indicating "I do not agree at all") to 2 (indicating "I very much agree"). As in our previous study, I reversed the score of items with negative valences by multiplying responses by -1. The five general questions were not part of my analyses for this project.

The demographic items asked participants about how many days they work on average in an intensive care or monitoring area, how many years/months of ICU experience they have, at which campus and unit they are working most of the time, and their profession.

To make the questionnaire more comprehensible, we made small adjustments to the wording of two items (items 8 and 9). In item 8 we changed the word “urgency” (“Dringlichkeit”, in the German version of the questionnaire) that was used in the original study to “situation” (German version: “Situation”). In item 9 we changed the wording “clinical symptoms” (German version: “klinische Symptome”) to “clinical picture” (German version: “Krankheitsbild”).

Table 1: An overview of all items along with their response options of the questionnaire that was sent to participants (own illustration). Items 1-9 are items from the CAFQa [6]. Note that the original items were in German. I am using the translations from my colleague’s and mine previous and forthcoming publication [6,24].

	Item	Response options
1	With too many alarms on my ward, my work performance, and motivation decrease.	5-Point Likert Options[b]
2	Too many alarms trigger physical symptoms for me, e.g., nervousness, headaches, and sleep disturbances.	5-Point Likert Options[b]
3	Alarms reduce my concentration and attention.	5-Point Likert Options[b]
4	My or neighboring patients' alarms or crisis alarms frequently interrupt my workflow.	5-Point Likert Options[b]
5	There are situations when alarms confuse me.	5-Point Likert Options[b]
6	In my ward, procedural instruction on how to deal with alarms is regularly updated and shared with all staff.[a]	5-Point Likert Options[b]
7	Responsible personnel respond quickly and appropriately to alarms.[a]	5-Point Likert Options[b]
8	The acoustic and visual monitor alarms used on my ward floor and in my nurse station allow me to assign the patient, the device, and the situation clearly.[a]	5-Point Likert Options[b]
9	Alarm limits are regularly adjusted based on patients' clinical pictures (e.g., blood pressure limits for conditions after bypass surgery).[a]	5-Point Likert Options[b]
10	I check the alarm limits at the beginning of the shift.[a]	5-Point Likert Options[b]
11	Activities close to the patient (e.g., blood sampling, mobilization, aspiration of tracheal secretions) result in an unnecessary number of alarms.	5-Point Likert Options[b]
12	Alarms are too frequent in my ward.	5-Point Likert Options[b]
13	When alarms go off repeatedly, I become indifferent to them.	5-Point Likert Options[b]
14	Alarms are often triggered even when there is no risk to patients.	5-Point Likert Options[b]
15	Please estimate your alarm fatigue in percent.[c]	Placement of a slider on a scale from 0-100%.
16	In your opinion, what percentage of all alarms are false alarms (e.g., due to measurement errors, artifacts, incorrect settings)?	Placement of a slider on a scale from 0-100%.
17	In which intensive care unit do you currently work?	Site specific selection of ICUs

	Item	Response options
18	On average, how many times a month do you work in the intensive care unit?	Up to two months; Up to one year; More than one year.
19	What function do you perform in the intensive care unit?	Physician; Nurse; supporting nurses, nurses in training, medical students, or interns.
20	Any other comments or suggestions regarding the questionnaire:	Free text

[a] Item with a negative valence that was reversely scored.

[b] The response options and the corresponding coding for data analyses (in brackets) were: “I do not agree at all” (-2), “I do not agree” (-1), “I agree in part” (0), “I agree” (1), "I very much agree" (2).

[c] Below the item, we added the following note: Alarm fatigue occurs when alarms desensitize personnel and reduce confidence in alarm systems, similar to the adage, "He who lies once is not believed, and even if he speaks the truth."

2.4 Statistical analyses

I conducted all analyses in R (Linux version 4.2.1) [25]. I handled the data using the packages *Tidyverse* [26] and *reshape2* [27] and I used the packages *psych* [28] and *lavaan* [29] for the factor analysis. The package *semPlot* [30] helped me diagram the factor models. The data can be accessed at www.zenodo.org/record/7801479.

2.4.1 Missing data

Deleting questionnaire submissions with missing data (i.e., “listwise deletion”) lowers statistical power and is not recommended [31]. Instead, missing data should be imputed for individual items [32]. Hence, I used the predictive mean matching of the mice package [33] to impute missing data that was assumed to be missing at random (MAR), as recommended by Heymans and Eekhout [34]. I did not impute and therefore deleted those questionnaires that were submitted completely blank. Likewise, I deleted questionnaires with signs of survey fatigue, as the assumption of MAR was not met in these cases, and imputation is not justified. I defined survey fatigue as participants not having answered at least the questionnaire's final 20% (i.e., three or more items of the CAFQa and the general questions). Given these definitions, 0.3% of the data were missing at random and subsequently imputed.

2.4.2 Assessment of data suitability for factor analysis

Traditional CFA using the maximum likelihood estimation method assumes that the data has a multivariate normal distribution [35]. According to Mardia's test, the multivariate skew was not normally distributed ($p < .001$), but kurtosis was ($p = 0.42$). When using ordered categorical variables it is recommended to calculate polychoric instead of product-moment correlations [36] and to use the unweighted least-squares (ULS) estimation method instead of maximum likelihood, especially, when the data are not multivariate normal distributed [36–38].

I aimed to identify outliers in the data using Mahalanobis distances [39] but detected none at $\chi^2(9)$ cutoff = 27.88 ($p < .001$). The Kaiser-Meyer-Olkin (KMO) statistic was 0.76, which is higher than the minimum value of 0.5 and close to the recommended value of 0.8 [40]. The correlation matrix was not an identity matrix as indicated by Bartlett's test of sphericity ($\chi^2(36) = 438.27$, $p < .001$). The determinant of the R matrix was greater than 0.00001 [40] and no correlations were greater than |0.7|. Hence there was no evidence of multicollinearity. In summary, these results indicate that the data is appropriate for conducting factor analysis.

2.4.3 Model specification and evaluation

I defined the model to be tested in the CFA in accordance with the findings of our previous publication [6], where my colleagues and I proposed a correlated two-factor solution where items 1-5 load on factor 1 (i.e., the alarm stress scale) and items 5-9 load on factor 2 (i.e., the alarm coping scale). I evaluated the fit of the model using the chi-square test. The model is rejected if the test is significant at alpha = 0.05. In addition to that, I used the following alternative fit indices with the cut-off values that Hu and Bentler [41] defined (cut-offs indicating good fit are provided in brackets): root mean square error of approximation (RMSEA; < 0.06), relative non-centrality index (RNI; > 0.95), Tucker-Lewis index (TLI; > 0.95), standardized root mean squared residual (SRMR; < 0.08), and comparative fit index (CFI; > 0.95).

2.4.4 Model modifications

I explored modification indices to improve the fit of the model. These modification indices had to be theoretically plausible and in line with the original theory of the CAFQa, as outlined in our previous work [6], to be added to the model. Because the items within factors correlated in our previous study, I prioritized modification indices that allow items

of the same factor to covary. Using the scaled chi-squared difference test, I accepted that a model fits the data better than another model if the difference of their chi-square value is statistically significant at alpha = 0.05.

2.4.5 Convergent validity

If two measurements of the same construct obtain similar results, convergent validity is indicated, which in turn supports construct validity [42]. So far, no alternative instrument for measuring alarm fatigue exists and hence there is no instrument that can serve as a comparison for the CAFQa. As a workaround, we included one item at the end of the questionnaire that asked participants to rate how much alarm fatigue they feel and another item asking them what percentage of alarms they perceive to be false. Both were answered by placing a slider on a scale of 0-100%. On the self-report item, a score of 0% indicated that participants feel no alarm fatigue at all and a score of 100% indicated that participants feel extreme alarm fatigue. On the false-alarm-estimation item, a score of 0% indicated that every alarm was true and a score of 100% indicated that no alarm was true. Participants were briefed about our definition of alarm fatigue in a short text above the self-report item. To analyze the convergent validity, I correlated each participant's mean score on the questionnaire and of each factor with the percentages of self-reported alarm fatigue and false alarm estimations.

2.4.6 Internal consistency

Internal consistency refers to the degree to which all items in a questionnaire measure the same construct. While Cronbach's coefficient alpha is the most popular estimator of internal consistency, it tends to underestimate if certain assumptions are not met [43]. It has been recommended to use McDonald's coefficient omega instead [43,44]. Here, I used both, Cronbach's coefficient alpha and McDonald's coefficient omega as estimators of the internal consistency of the questionnaire. I also calculated the mean inter-item correlation of both factors. High values in these measures suggests that the items are closely related. Conversely, low values suggest that the items are not closely related, suggesting they may not be reliably measuring the same construct.

3. Results

The results described in this section are part of a forthcoming publication [24].

3.1 Participant demographics

In total, 363 participants submitted a questionnaire, of which eight submissions matched our definition of survey fatigue and were therefore excluded. Twenty-three participants did not allow their data to be processed and 67 participants submitted blank questionnaires. Hence, I excluded the data in both cases. The remaining sample size was N = 265. The majority of participants were nurses (n = 150; 56.6%) and physicians (n = 95; 35.8%). Only nine participants were either interns, medical students, nurses in training or supporting nurses (3.4%). Eleven participants (4.2%) did not provide information about their professional background. Most participants are experienced in ICU settings: 219 (82.6%) worked more than eight days per month in an ICU and 203 (76.6%) have more than one year of ICU experience. Figure 1 gives an overview of participant's demographic data.

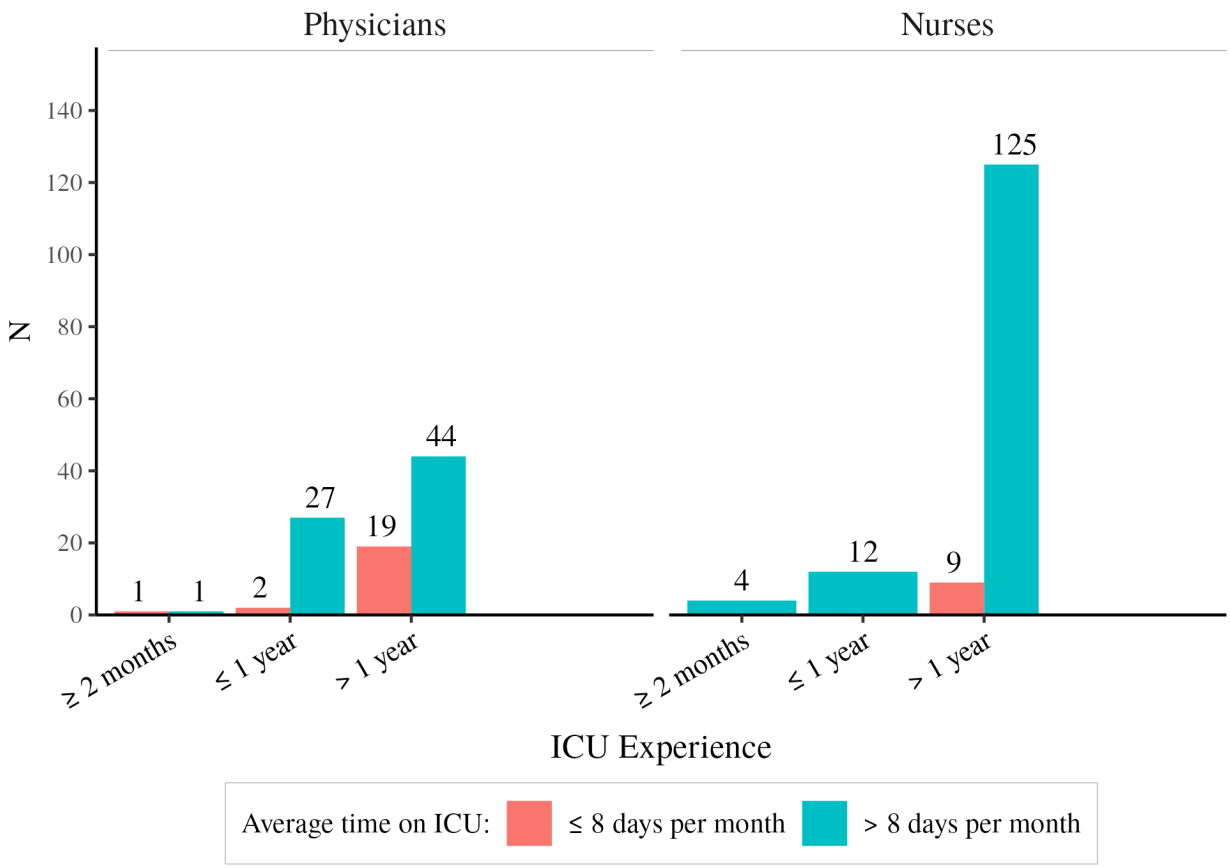

Figure 1. Participants' demographic data (own illustration). Most participants were experienced nurses and physicians. For the sake of clarity, I omitted some data from the figure: Not shown are the 11 missing data points, the data of one physician who did not provide their average monthly time in ICU, and the data of nine participants who fell into the category "intern, medical student, nurse in training or supporting nurse".

3.2 Confirmatory factor analysis

Fitting the correlated two-factor model to the data yielded mixed results: While the chi-square test rejected the model at $\chi^2(26) = 44.932$, p = 0.012, all alternative fit indices accepted it (SRMR = 0.052, RMSEA = 0.03, TLI = 0.985, CFI = 0.989, and RNI = 0.989). The factor loadings ranged from 0.35 to 0.73 and were all statistically significant (see Table 2). The correlation between the two factors was moderate but statistically significant (r = 0.4, p < .001, 95% CI = 0.21-0.59). Figure 2 shows a diagram of the model. An overview of the model fit statistics can be found in Table 3.

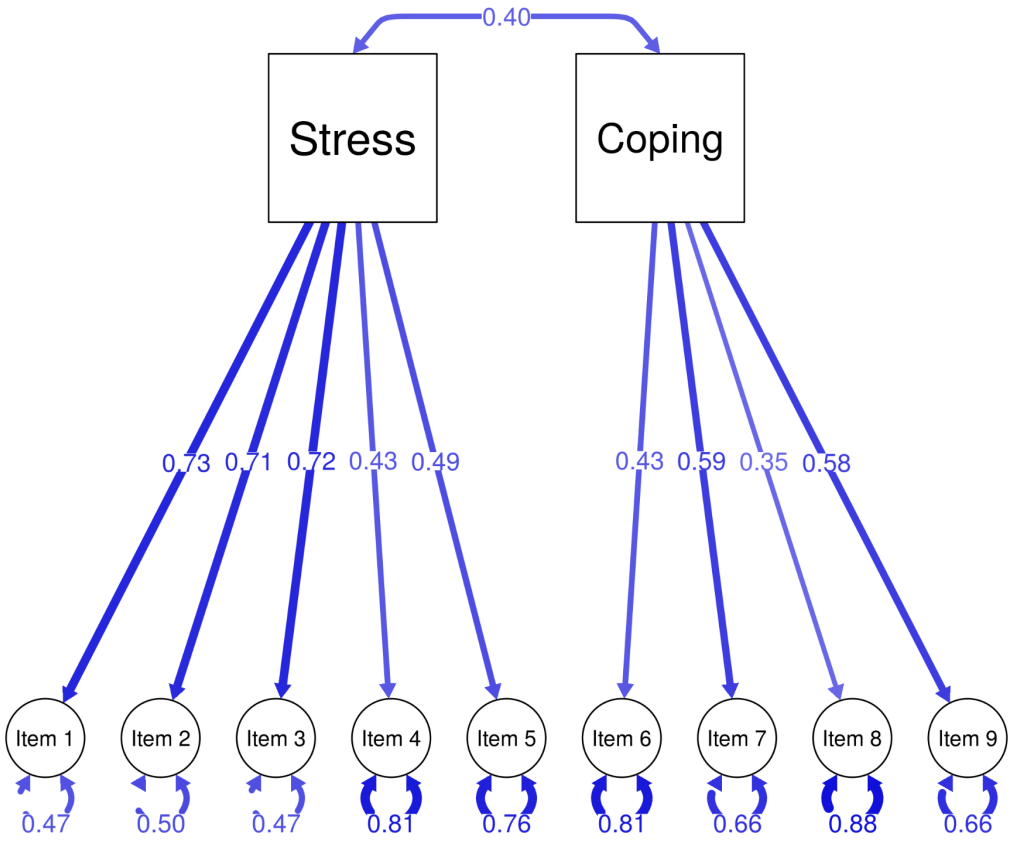

Figure 2. Diagram of the model (modified after Figure 1 of the forthcoming publication [24]). The squares represent the two factors, and each circle represents one of the nine

items. The correlation between Factors is represented by the arrow connecting them. Residuals are shown as circular arrows below items.

3.2.1 Model modifications

I accepted the original two-factor model because the fit indices indicated a good fit and because model modifications can make a model less generalizable. However, I explored modification indices to see what it would take to gain a non-significant result on the chi-square test.

The five modification indices with the largest impact on the fit of the model were between 3.76 and 10.2. However, freeing the error covariance of items 4 and 5, which are similar in content and load onto the same factor, was the only option that fulfilled the criteria of theoretical plausibility outlined above. After adding this term to the model specification, the chi-square test became significant: $\chi^2(25) = 37.158$, $p = 0.056$. All alternative fit indices improved as well: RNI = 0.998, TLI = 0.996, CFI = 0.998, SRMR = 0.047, and RMSEA = 0.015. The difference in chi-square between the original model and the modified model was statistically significant at χ^2diffence(1) = 9.1886, $p = 0.0024$. The correlation between the factors barely changed ($r = 0.41$, $p < .001$, 95% CI = 0.22-0.59), while most items had slightly larger factor loadings (now between 0.35 and 0.74; all statistically significant at $p < .001$). Figure 3 visualizes the modified model and Table 3 summarizes all model fit statistics.

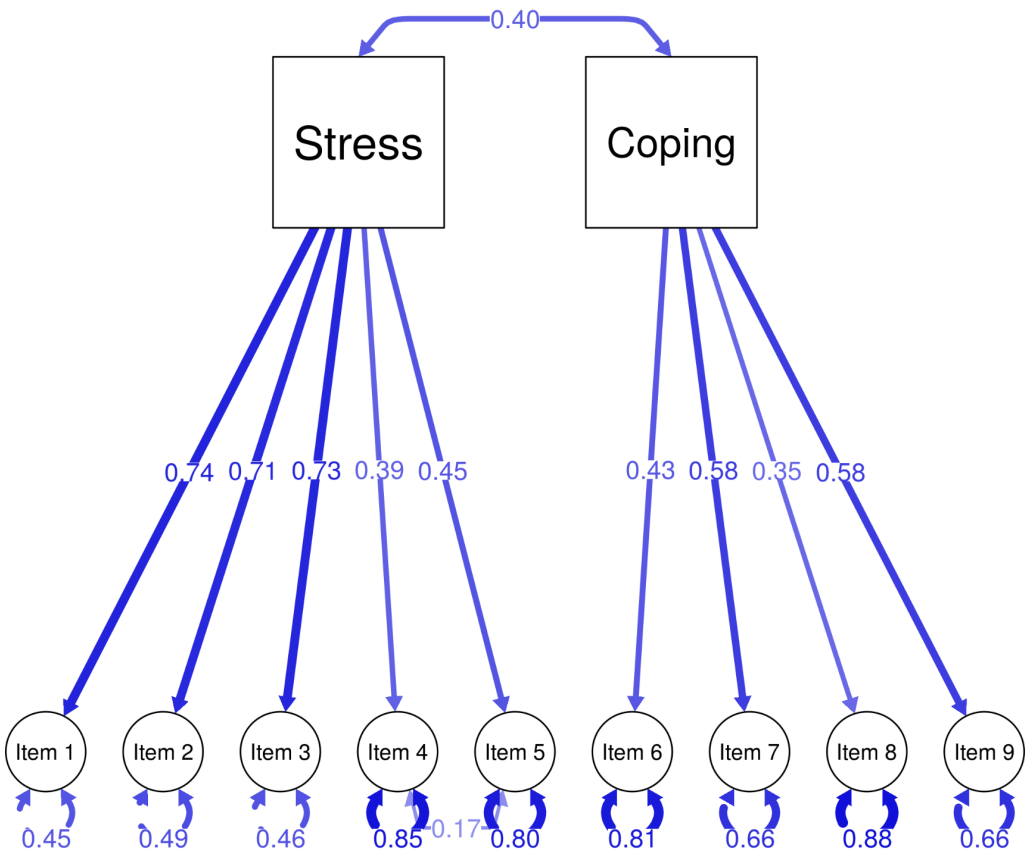

Figure 3. Diagram of the modified model (own illustration). The model modification (i.e., error covariance) is indicated by the arrow connecting items 4 and 5.

Table 2: Descriptive item statistics and the pattern coefficients of the original two-factor model and of the exploratively modified model (modified after Table 1 of the forthcoming publication [24]). All numbers were rounded to two digits. F = factor, CI = confidence interval, SD = standard deviation, Kurt. = kurtosis.

		Original Model			**Modified Model**			**Item Statistics**			
	Item	F1[b]	F2[b]	95% CI	F1[b]	F2[b]	95% CI	Mean	SD	Kurt.	Skew
1	With too many alarms on my ward, my work performance, and motivation decrease.	0.73	-	0.64-0.82	0.74	-	0.65-0.83	0.47	1.01	-0.62	-0.3
2	Too many alarms trigger physical symptoms for me, e.g., nervousness, headaches, and sleep disturbances.	0.71	-	0.61-0.80	0.71	-	0.62-0.81	0.23	1.26	-1.13	-0.16
3	Alarms reduce my concentration and attention.	0.72	-	0.63-0.81	0.73	-	0.64-0.82	0.43	1.07	-0.91	-0.21
4	My or neighboring patients' alarms or crisis alarms frequently interrupt my workflow.	0.43	-	0.32-0.55	0.39	-	0.27-0.51	0.87	0.83	-0.41	-0.39
5	There are situations when alarms confuse me.	0.49	-	0.38-0.59	0.45	-	0.34- 0.56	0.08	1.09	-0.73	-0.09
6	In my ward, procedural instruction on how to deal with alarms is regularly updated and shared with all staff.[a]	-	0.43	0.27-0.60	-	0.43	0.27-0.60	0.77	1.24	-0.68	-0.7
7	Responsible personnel respond quickly and appropriately to alarms.[a]	-	0.59	0.42-0.75	-	0.58	0.42-0.75	-0.32	0.82	-0.12	-012
8	The acoustic and visual monitor alarms used on my ward floor and in my nurse station allow me to assign the patient, the device, and the situation clearly.[a]	-	0.35	0.18-0.52	-	0.35	0.18-0.52	-0.46	1.06	-0.52	0.36
9	Alarm limits are regularly adjusted based on patients' clinical pictures (e.g., blood pressure limits for conditions after bypass surgery).[a]	-	0.58	0.43-0.73	-	0.58	0.43-0.73	-0.38	0.93	-0.24	0.26

[a] Item with a negative valence that was reversely scored.

[b] All loadings were statistically significant at $p < 0.001$.

Table 3: Model fit statistics of the original two-factor model and of the exploratively modified model (own illustration). DF = degrees of freedom, RNI = Relative Non-Centrality Index, TLI = Tucker-Lewis Index, CFI = Comparative Fit Index, SRMR = Standardized Root Mean Squared Residual, RMSEA = Root Mean Square Error of Approximation.

	χ^2 (DF)	P	RNI	TLI	CFI	SRMR	RMSEA
Original Model	44.932 (26)	0.012	0.989	0.985	0.989	0.052	0.03
Modified Model	37.158 (25)	0.056	0.998	0.996	0.998	0.047	0.015
Difference	9.1886 (1)	0.0024	-	-	-	-	-

3.3 Convergent validity

There was a moderate, statistically significant correlation between participant's self-reported alarm fatigue and their mean scores on the questionnaire: r(242) = 0.45 (p < .001, 95% CI = 0.34-0.54). This association was similar when isolating the mean scores of factor 1 (r(242) = 0.42, p < .001, 95% CI = 0.31-0.52), but weaker when isolating the mean scores of factor 2 (r(242) = 0.29, p < .001, 95% CI = 0.17-0.4).

The correlation between participants' estimated percentage of false alarms and their mean scores on the questionnaire was weak, yet statistically significant: r(247) = 0.3, (p < .001, 95% CI = 0.18,-0.41). Here, the association was stronger for the mean scores of factor 2 (r(247) = 0.29, p < .001, 95% CI = 0.17-0.4), than for those of factor 1 (r(247) = 0.2, p = 0.0016, 95% CI = 0.08-0.32).

3.4 Internal consistency

Across factors, Cronbach's coefficient alpha = 0.67 and McDonald's coefficient omega = 0.8. Within factors, Cronbach's coefficient alpha of factor 1 was 0.72, and of factor 2 0.49. McDonald's coefficient omega of factor 1 was 0.77 and of factor 2 0.55. Items on factor 1 had a mean correlation of 0.38, and items on factor 2 had a mean correlation of 0.23.

4. Discussion

ICU staff and patients can be exposed to hundreds of false or non-actionable alarms from medical devices and both groups can suffer serious consequences from that. Patients can be disturbed in their sleep and recovery, while ICU staff can develop alarm fatigue, i.e. a desensitization to alarms [5,8]. Clinical alarm researchers and clinicians may try to remedy this problem by implementing alarm management, but they lack a reliable way of assessing nurses' and physicians' alarm fatigue before and after such changes [6]. Hence, they do not know if their efforts were fruitful. Therefore, my colleagues and I developed the Charité Alarm Fatigue Questionnaire in a previous work [6]. It consists of nine items along two scales: the alarm stress and the alarm coping scale. The alarm stress scale captures staff's psychophysiological responses to alarm overload, while the alarm coping scale captures how staff's ICU manages alarms in general. However, these scales were discovered by exploration. If these scales could be shown to re-emerge by collecting questionnaire data on a new and independent sample, it would be an important contribution towards understanding the CAFQa's construct validity. In other words, it would indicate that the questionnaire indeed measures the construct "alarm fatigue", as previously defined [6]. That is why, for this thesis, I issued the CAFQa to nurses and physicians in the ICUs of five major German hospitals and submitted the data to a confirmatory factor analysis [24]. My aim was to find out whether I could replicate the exploratively discovered factor structure and thus underpin the CAFQa's construct validity. The findings of this study will form part of a forthcoming publication [24].

4.1 Principal findings

I analyzed the data of 265 participants, of which most were experienced nurses [24]. In the confirmatory factor analysis, all alternative fit indices indicated that the two-factor model indeed fits the data, while the chi-squared test indicated the opposite. All in all, however, I considered the model as confirmed: the CAFQa seems to measure alarm fatigue in line with theory.

The convergent validity of the instrument was supported by a moderate correlation between participants' mean scores on the CAFQa and the alarm fatigue that they

estimated for themselves. There was a weak correlation between participants' mean scores and their perceived rate of false alarms.

The traditional measure of internal consistency, the coefficient Cronbach's alpha, yielded moderate results. Coefficient McDonald's Omega, the more appropriate measure of internal consistency for this study, yielded good results. Items of the alarm stress scale correlated moderately, and items of the alarm coping scale had a low inter-item correlation. The alarm coping scale had consistently lower coefficients of internal consistency than the alarm stress scale.

4.2 Interpretation of the findings

4.2.1 The CAFQa's construct validity

In our previous work [6], my colleagues and I proposed that the construct "alarm fatigue" should not only be measured by asking respondents about the psychophysiological effects alarms have on them. Instead, we argued that structural and systemic effects should be considered too. These two aspects align with our proposed factor structure: factor 1, the alarm stress scale, measures psychophysiological aspects of alarm fatigue, and factor 2, the alarm coping scale, measures systemic and structural aspects. The findings of this thesis support the two-factor model of the CAFQa. The exploratively derived factor structure was thus replicated on a new sample with participants from different ICUs in Germany. This is an important insight into the construct validity of the CAFQa.

A small caveat to this conclusion is, however, that the model would need to be modified to reach a non-significant chi-squared test result along with the alternative fit indices (by allowing the error terms of item 4 and 5 to covary). This adjustment would be theoretically plausible because both items loaded onto factor 1 and both have similar content. Item 4 reads "[…] alarms or crisis alarms frequently interrupt my workflow." and item 5 reads "There are situations when alarms confuse me." When ICU staff are interrupted by alarms in their current task, they may feel briefly confused. This claim is supported by a psychological theory outlined in D'Mello [45] (citing Mandler [46,47]): being interrupted can create confusion. It is therefore reasonable to assume that the participants of the

present study answered items 4 and 5 similarly, which lead to a covariation of the two and in turn to the covariance of their error terms.

Nonetheless, I suggest not modifying the model, because of three reasons: First, even the slightest model modification can make the model less generalizable. Second, it is likely that the chi-squared test rejected the original model simply because it is sensitive to large sample sizes [48]. And third, all alternative fit indices unequivocally indicated a good model fit. While some argue that chi-square should be the single source of truth when evaluating model fit [49], the general consensus is that alternative fit indices are useful [48]. I am looking forward to finding out whether future studies can fit the original model with the results of chi-squared pointing in the same direction as the alternative fit indices of the present study.

4.2.2 Convergent validity

When being asked to estimate their own alarm fatigue in percent, those participants who had a high mean score on the CAFQa were more likely to estimate a higher percentage for themselves – and vice versa. This was indicated by a moderate positive correlation between the mean scores on the questionnaire and the self-report item. The alarm stress scale seems to correlate more with self-reported alarm fatigue than the alarm coping scale. My colleagues and I observed the same pattern in our previous work [6]: There was a correlation of $r = 0.56$ between the whole questionnaire and self-reported alarm fatigue ($r = 0.45$ in the present study), and a correlation of $r = 0.54$ and $r = 0.3$ for the alarm stress and alarm coping scale, respectively ($r = 0.42$ and $r = 0.29$, in the present study, respectively).

Similarly, participants were more likely to report a higher rate of false alarms in their day-to-day work when they have a higher mean score on the CAFQa and vice versa. This was indicated by a weak correlation. Interestingly, the correlation between the alarm coping scale alone and the estimated false alarm rate seems to be higher than the correlation between the alarm stress scale and the estimated false alarm rate. To me, this pattern is plausible, because the alarm coping scale encompasses items about the alarm management practices of a given ICU (e.g., using procedural instructions or individualizing patient's alarm limits) and because ICUs that actively manage their alarms

can reduce the number of alarms – and with that the number of false alarms [24]. The weak association between participants' mean scores and the estimated false alarm rate in general and the association between the alarm stress scale and the estimated false alarm rate in particular might be explained by the observation that alarm fatigue is not caused by the number of (false) alarms alone. For example, Sowan et al. [12] found that the alarm fatigue of ICU staff did not decrease, regardless of the decreased number of alarms after implementing new alarm management. However, all these explanations remain speculations and no causation can be inferred from the observed correlations. Nonetheless, I am looking forward to seeing ingenious research designs of future studies helping to untangle these mechanisms. I wonder, for example, whether ICU staff that perceives many alarms to be false develops stronger alarm fatigue, or whether alarm fatigued staff simply *perceives* more alarms as being false [24]. To answer this question, it would be valuable to know the true positive predictive value of the alarms in a given ICU. Since such data is unavailable, automatically annotated datasets are a promising alternative [50].

4.2.3 Internal consistency

The CAFQa had an acceptable internal consistency, as indicated by the coefficients Cronbach's Alpha and McDonald's Omega. However, the internal consistency of the alarm coping scale was underwhelming on both coefficients. Interestingly, a similar pattern emerged in my colleague's and my previous work [6]: the questionnaire as a whole and the alarm stress scale had good internal consistency, but the alarm coping scale did not. Building a questionnaire with strong internal consistency is a double-edged sword: the more similar the items' content, the higher the questionnaire's internal consistency – but at the cost of potentially missing capturing some facets of a construct. The CAFQa is a very short questionnaire that has the ambition to cover the many facets of alarm fatigue. Particularly the alarm coping scale covers a wide range of topics. This might explain its repeatedly low internal consistency. If this pattern is continued in future studies, we recommend finding out how the internal consistency of the alarm coping scale might be improved. One option is to enrich it with more items, which comes at the cost of making the questionnaire longer and hence more burdensome to fill out.

4.2.4 Is ‘alarm coping’ a different construct?

The results described above make me wonder whether the alarm coping scale could be measuring a slightly different construct that is closely related to but not directly measuring the construct of alarm fatigue. To me, this is hinted at by the consistently lower correlation between participants’ self-reported alarm fatigue and the mean scores on the alarm coping scale, compared to the alarm stress scale. Additionally, in my colleague’s and my previous work, the factor loadings and communalities of items on the alarm coping scale were lower than those of the alarm stress scale [6]. In this article, we added that it would be interesting to see if the two scales could be dissociated and that future studies should find out if they could be used independently. I suspect the alarm stress scale directly measures alarm fatigue, while the alarm coping scale measures alarm management (thus only indirectly alarm fatigue). Structural equation modeling (specifically path analysis) can be used to find hierarchical relationships between the two scales. If it was the case that the two scales measure different constructs, future research should find ways to enhance the consistenly lower coefficients of internal consistency for the alarm coping scale (as presented in this thesis and the previous work [6,24]).

4.3 Using the CAFQa in combination with alarm-logs

When assessing the alarm situation in an ICU, staff’s alarm fatigue is one of many variables to consider. Together with Poncette et al., I developed an iterative framework, where data plays a key role for designing new alarm management interventions [2]. We proposed to show ICU staff alarm log data, to jointly design interventions, and to use alarm log data again to evaluate whether the interventions had an effect on the alarm burden. I suggest adding the CAFQa as a routine measurement to this loop and to use it alongside alarm log data to evaluate the status quo and the effect of alarm management interventions on staff’s subjective alarm burden. This can help identify situations as the one described in Sowan [12], where no change in staff’s alarm fatigue was detected, despite a significant reduction in the number of alarms.

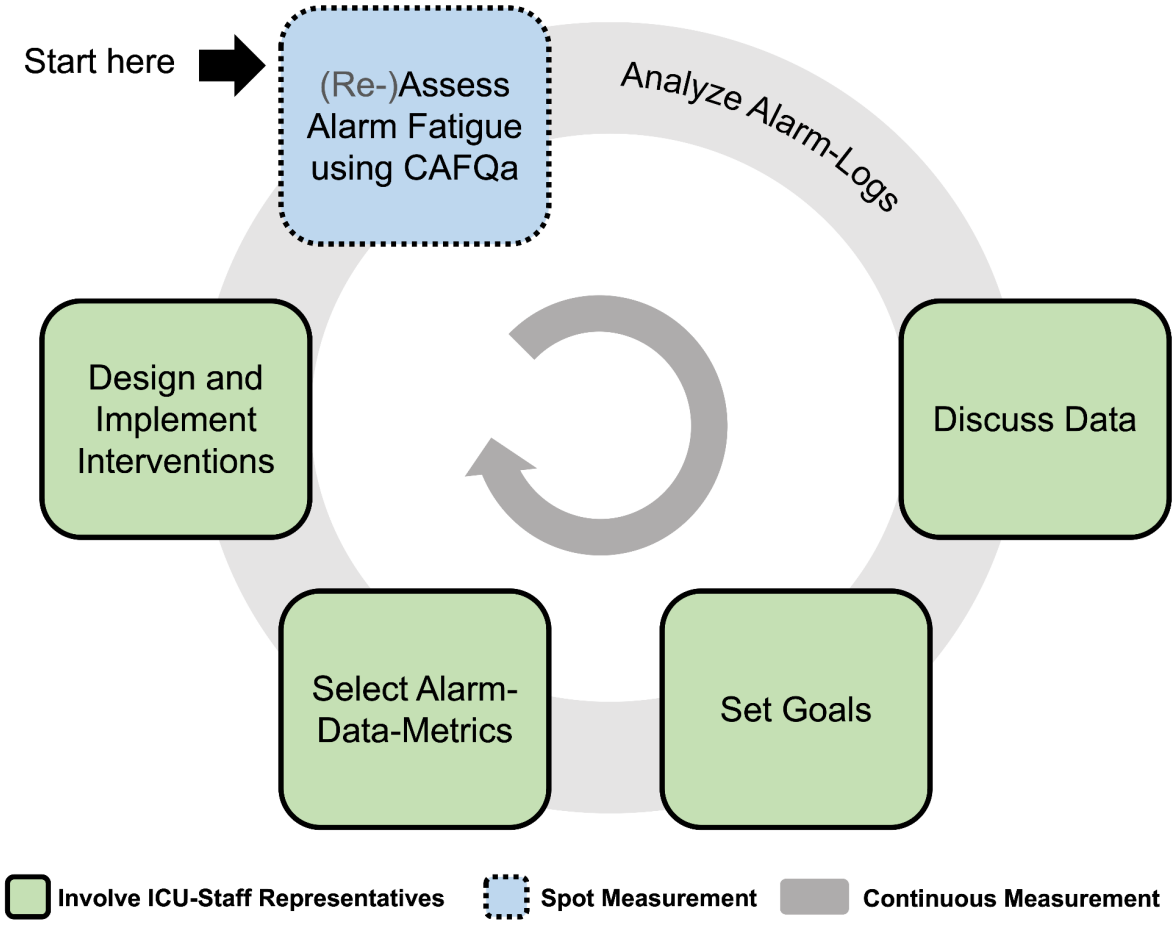

Figure 4. My recommended procedure for designing data-driven alarm management interventions in ICUs (adapted from Poncette et al. [2]). First, the ICU's alarm situation should be assessed by quantifying staff's alarm fatigue using the CAFQa and by analyzing alarm-logs. In focus groups with ICU-staff representatives, discuss the data and set goals for improving the alarm situation. Choose alarm-data-metrics that allow evaluating progress towards goals. Finally, design and implement concrete alarm management interventions. Pay special attention to the results of the CAFQa's alarm coping scale, as it might hint at specific needs for improvement. Alarm-logs should be continuously collected and analyzed. After a few weeks, consider measuring staff's alarm fatigue again. I recommend iterating through this loop, because, in my opinion, data driven alarm management is most effective when it becomes an integral part of an ICU's culture. My colleagues and I are piloting this approach in an ongoing clinical study [51] (DRKS00029655).

4.4 Strengths and limitations of this work

Xia and Yang as well as Savalei have shown that RMSEA, CFI, and TLI can overestimate model fit under the Unweighted Least Squares estimator [24,52,53]. The possibility remains that I accepted a model because these fit indices indicated a good model fit, while in fact the model fitted badly. However, I am optimistic that this was not the case, because I took other fit indices into account to evaluate model fit – including chi-squared. The former unequivocally indicated a good model fit. The latter indicated a bad fit initially, but indicated a good fit too, after a minor (theoretically plausible) model modification. Also, chi-squared is known to be sensitive to the number of model parameters and to large sample sizes and therefore a rather conservative estimate of model fit.

My colleagues and I repeatedly highlight that it could be misleading to assume that nurses and physicians can express their own alarm fatigue in percent [6,24]. As long as no other validated questionnaire for alarm fatigue exists however, I am confident that asking participants directly to estimate their own alarm fatigue is a worthwhile tool for approximating the questionnaire's convergent validity. For the present study, we ensured that participants understand what they are asked to do by providing them with a brief description of alarm fatigue along with the self–report-item. Likewise, asking participants to estimate their perceived rate of false alarms can only crudely approximate the real positive predictive value of alarms. However, I believe it is valuable nonetheless until automatically annotated alarm data becomes reliable reality. With sufficiently large sample sizes, one might even approximate true positive predictive value due to the law of large numbers.

I outlined above how we slightly modified the wording of items 8 and 9 to improve their readability. In hindsight, this was a risky operation because it might have changed the items' performance. Luckily, this was not the case. Hence, I recommend continuing to use the new wording "clinical pictures" in item 9. Regarding the change of words in item 8, however, I recommend reverting to the word choice of "urgency" instead of "situation", because, in hindsight, it seems to capture more of the meaning that we originally intended it to have in our previous work [6]. "Urgency" seems to be more straightforward than "situation" when asking nurses and physicians how they assess an alarm's criticality.

Regardless of these limitations, I believe the study of this is invaluable for advancing the design of a reliable instrument for measuring alarm fatigue in nurses and physicians. My

colleagues and I were the first to systematically design a questionnaire that adheres to the best practices of scale construction [6] and the present study is the first in clinical alarm research to test and confirm a previously defined construct model of alarm fatigue. Thus, I am confident that the hypothesized two-factor model is generalizable. Similarly, I consider the results in terms of internal consistency and convergent validity to be generalizable, because the present and the previous study had similar results in this domain [6,24]. Only in the present study have I correlated participants' mean scores on the CAFQa with their subjectively estimated false alarm rate. Therefore, unfortunately, there are no values with which to compare my results. However, all correlations found were statistically significant, which hints at them being generalizable.

4.5 Practical implications and future research

An important next step is to validate the English translation of the questionnaire using a large sample in English-speaking countries. I recommend conducting cognitive interviews with nurses and physicians to find out how the English translation can be improved. I also recommend paying special attention to cultural differences that might influence the way a question is understood. If the two-factor model of the CAFQa can be confirmed in a different language and in a different culture, it would indicate that the translation was successful. Such a finding would also add to the construct validity of the CAFQa in general.

It would be interesting to see whether future studies can assess convergent validity by using non-validated surveys on alarm fatigue that other authors used before the CAFQa was developed (we provided a list of studies in our previous work [6]), or by using questionnaires that measure constructs related to alarm fatigue (e.g., stress or burn-out). Likewise, future studies should investigate discriminant validity by administering questionnaires that measure something unrelated to alarm fatigue (e.g., team cohesion).

5. Conclusion

So far, the Charité Alarm Fatigue Questionnaire (CAFQa) remains the only available questionnaire for measuring alarm fatigue in nurses and physicians that was developed according to the best practices of scale construction. It is freely accessible, easy to use, quick and intuitive. The results of this thesis support the questionnaire's construct validity and indicate that it indeed measures alarm fatigue. The alarm stress scale has five items and measures the psychophysiological effects of alarms; the alarm coping scale has four items and measures the influence of systemic variables and alarm management practices in an ICU's alarm situation. Future studies should investigate the internal consistency of the alarm coping scale, find out more about the questionnaire's discriminant and convergent validity and validate its English translation. Ideally, such studies also aim to replicate the two-factor model, as in this thesis. When designing data driven alarm management interventions jointly with ICU staff, the CAFQa should be used to assess the baseline levels of alarm fatigue and to measure changes in response to any new interventions.

Literaturverzeichnis

1. Schmid F, Goepfert MS, Reuter DA. Patient monitoring alarms in the ICU and in the operating room. Crit Care Lond Engl. 2013 Mar 19;17(2):216.
2. Poncette AS, Wunderlich MM, Spies C, Heeren P, Vorderwülbecke G, Salgado E, Kastrup M, Feufel MA, Balzer F. Patient Monitoring Alarms in an Intensive Care Unit: Observational Study With Do-It-Yourself Instructions. J Med Internet Res. 2021 May 28;23(5):e26494.
3. Jones K. Alarm fatigue a top patient safety hazard. CMAJ Can Med Assoc J J Assoc Medicale Can. 2014 Feb 18;186(3):178.
4. Joint Commission. Medical device alarm safety in hospitals. Sentin Event Alert. 2013 Apr 8;(50):1–3.
5. Sendelbach S, Funk M. Alarm Fatigue: A Patient Safety Concern. AACN Adv Crit Care. 2013;24(4):378–86.
6. Wunderlich MM, Amende-Wolf S, Krampe H, Kruppa J, Spies C, Weiß B, Memmert B, Balzer F, Poncette AS. A brief questionnaire for measuring alarm fatigue in nurses and physicians in intensive care units. Sci Rep. 2023 Aug 24;13(1):13860.
7. AAMI Foundation. Clinical Alarm Management Compendium [Internet]. 2015 [cited 2020 Jan 16]. Available from: https://www.aami.org/docs/default-source/foundation/alarms/alarm-compendium-2015.pdf?sfvrsn=2d2b53bd_2
8. Simons KS. Impact of light and noise exposure on critically ill patients [Internet] [Dissertation]. [Radboud]: Radboud University; 2018. Available from: https://repository.ubn.ru.nl/handle/2066/194289
9. Ferreira VR, Pereira AR, Vieira J, Pereira F, Marques R, Campos G, Sampaio A, Crego A. Capturing the attentional response to clinical auditory alarms: An ERP study on priority pulses. PLOS ONE. 2023 Feb 16;18(2):e0281680.
10. Wilken M, Hüske-Kraus D, Röhrig R. Alarm Fatigue: Using Alarm Data from a Patient Data Monitoring System on an Intensive Care Unit to Improve the Alarm Management. Stud Health Technol Inform. 2019;273–81.
11. Hüske-Kraus D, Wilken M, Röhrig R. Measuring Alarm System Quality in Intensive Care Units. Zuk Pflege Tagungsband 1 Clust 2018. 2018;89.
12. Sowan AK, Gomez TM, Tarriela AF, Reed CC, Paper BM. Changes in Default Alarm Settings and Standard In-Service are Insufficient to Improve Alarm Fatigue in an Intensive Care Unit: A Pilot Project. JMIR Hum Factors [Internet]. 2016 Jan 11 [cited 2020 Oct 3];3(1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4797663/
13. Lewandowska K, Weisbrot M, Cieloszyk A, Mędrzycka-Dąbrowska W, Krupa S, Ozga D. Impact of Alarm Fatigue on the Work of Nurses in an Intensive Care Environment—A Systematic Review. Int J Environ Res Public Health. 2020 Nov 13;17(22):8409.
14. Wears RL, Perry SJ. Human factors and ergonomics in the emergency department. Ann Emerg Med. 2002 Aug 1;40(2):206–12.
15. Ashrafi S, Najafi Mehri S, Nehrir B. Designing an Alarm Fatigue Assessment Questionnaire: Evaluation of the Validity and Reliability of an Instrument. J Crit Care Nurs [Internet]. 2017 Nov 30 [cited 2023 Jan 23];10(4). Available from: http://jccnursing.com/en/articles/11647.html
16. Torabizadeh C, Yousefinya A, Zand F, Rakhshan M, Fararooei M. A nurses' alarm fatigue questionnaire: development and psychometric properties. J Clin Monit Comput. 2017 Dec;31(6):1305–12.

17. Rypicz Ł, Rozensztrauch A, Fedorowicz O, Włodarczyk A, Zatońska K, Juárez-Vela R, Witczak I. Polish Adaptation of the Alarm Fatigue Assessment Questionnaire as an Element of Improving Patient Safety. Int J Environ Res Public Health. 2023 Jan;20(3):1734.
18. Boateng GO, Neilands TB, Frongillo EA, Melgar-Quiñonez HR, Young SL. Best Practices for Developing and Validating Scales for Health, Social, and Behavioral Research: A Primer. Front Public Health. 2018 Jun 11;6:149.
19. Smith GT. On Construct Validity: Issues of Method and Measurement. Psychol Assess. 2005 Dec;17(4):396–408.
20. Cronbach LJ, Meehl PE. Construct validity in psychological tests. Psychol Bull. 1955;52(4):281.
21. DiStefano C, Hess B. Using confirmatory factor analysis for construct validation: An empirical review. J Psychoeduc Assess. 2005;23(3):225–41.
22. Bryant FB, Yarnold PR, Michelson EA. Statistical Methodology.: VIII. Using Confirmatory Factor Analysis (CFA) in Emergency Medicine Research. Acad Emerg Med. 1999 Jan;6(1):54–66.
23. Beaujean AA. Latent variable modeling using R: A step-by-step guide. Routledge; 2014.
24. Wunderlich M, Krampe H, Amende-Wolf S, Spies C, Weiß B, Balzer F, Poncette AS, Study Group. Evaluating the Construct Validity of the Charité Alarm Fatigue Questionnaire using Confirmatory Factor Analysis. Forthcoming. 2023;
25. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing [Internet]. Vienna, Austria; 2022. Available from: https://www.R-project.org/
26. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J. Welcome to the Tidyverse. J Open Source Softw. 2019;4(43):1686.
27. Wickham H. Reshaping data with the reshape package. J Stat Softw. 2007;21(12):1–20.
28. Revelle W. psych: Procedures for Psychological, Psychometric, and Personality Research [Internet]. 2020 [cited 2020 Mar 26]. Available from: https://CRAN.R-project.org/package=psych
29. Rosseel Y. lavaan: An R package for structural equation modeling. J Stat Softw. 2012;48:1–36.
30. Epskamp S. semPlot: Unified visualizations of structural equation models. Struct Equ Model Multidiscip J. 2015;22(3):474–83.
31. Nakagawa S, Freckleton RP. Missing inaction: the dangers of ignoring missing data. Trends Ecol Evol. 2008 Nov;23(11):592–6.
32. Eekhout I, de Vet HCW, Twisk JWR, Brand JPL, de Boer MR, Heymans MW. Missing data in a multi-item instrument were best handled by multiple imputation at the item score level. J Clin Epidemiol. 2014 Mar 1;67(3):335–42.
33. Van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. J Stat Softw. 2011;45(1):1–67.
34. Heymans M, Eekhout I. Applied missing data analysis with SPSS and (R) studio [Internet]. Amsterdam, The Netherlands; 2019. Available from: https://bookdown.org/mwheymans/bookmi/missing-data-in-questionnaires.html
35. MacCallum RC. Factor analysis. In: The SAGE Handbook of Quantitative Methods in Psychology. Thousand Oaks, CA: SAGE Publications; 2009. p. 123–47.
36. Flora DB, LaBrish C, Chalmers RP. Old and new ideas for data screening and assumption testing for exploratory and confirmatory factor analysis. Front Psychol.

2012;3:55.
37. Forero CG, Maydeu-Olivares A, Gallardo-Pujol D. Factor analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. Struct Equ Model. 2009;16(4):625–41.
38. Koğar H, Yilmaz Koğar E. Comparison of Different Estimation Methods for Categorical and Ordinal Data in Confirmatory Factor Analysis. Eğitimde Ve Psikolojide Ölçme Ve Değerlendirme Derg. 2015 Dec 28;6(2).
39. McLachlan GJ. Mahalanobis distance. Resonance. 1999;4(6):20–6.
40. Field A, Miles J, Field Z. Discovering Statistics Using R. SAGE Publications Ltd; 2012.
41. Hu L, Bentler PM. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. Struct Equ Model Multidiscip J. 1999 Jan 1;6(1):1–55.
42. Convergent validity. In: Wikipedia [Internet]. 2021 [cited 2023 Jan 25]. Available from: https://en.wikipedia.org/w/index.php?title=Convergent_validity&oldid=1053029785
43. Kalkbrenner MT. Alpha, Omega, and *H* Internal Consistency Reliability Estimates: Reviewing These Options and When to Use Them. Couns Outcome Res Eval. 2021 Jul 30;1–12.
44. Peters GJY. The alpha and the omega of scale reliability and validity: why and how to abandon Cronbach's alpha and the route towards more comprehensive assessment of scale quality. Eur Health Psychol. 2014;16(2):56–69.
45. D'Mello S, Lehman B, Pekrun R, Graesser A. Confusion can be beneficial for learning. Learn Instr. 2014 Feb 1;29:153–70.
46. Mandler G. Mind and body: Psychology of emotion and stress. New York, NY, USA: W.W. Norton & Company Incorporated; 1984.
47. Mandler G. Interruption (discrepancy) theory: review and extensions. In S. Fisher, & C. L. Cooper (Eds.), On the move: The psychology of change and transition (pp. 13e32). Chichester: Wiley.
48. Alavi M, Visentin DC, Thapa DK, Hunt GE, Watson R, Cleary M. Chi-square for model fit in confirmatory factor analysis. J Adv Nurs. 2020 Sep;76(9):2209–11.
49. Barrett P. Structural equation modelling: Adjudging model fit. Personal Individ Differ. 2007 May;42(5):815–24.
50. Chromik J, Klopfenstein SAI, Pfitzner B, Sinno ZC, Arnrich B, Balzer F, Poncette AS. Computational approaches to alleviate alarm fatigue in intensive care medicine: A systematic literature review. Front Digit Health [Internet]. 2022 [cited 2023 Feb 1];4. Available from: https://www.frontiersin.org/articles/10.3389/fdgth.2022.843747
51. DRKS - Deutsches Register Klinischer Studien [Internet]. [cited 2023 Sep 27]. Available from: https://drks.de/search/de/trial/DRKS00029655
52. Xia Y, Yang Y. RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. Behav Res Methods. 2019 Feb 1;51(1):409–28.
53. Savalei V. Improving Fit Indices in Structural Equation Modeling with Categorical Data. Multivar Behav Res. 2021 May 4;56(3):390–407.

Eidesstattliche Versicherung

„Ich, Maximilian Markus Wunderlich, versichere an Eides statt durch meine eigenhändige Unterschrift, dass ich die vorgelegte Dissertation mit dem Titel “Ist der Charité Alarm Fatigue Fragebogen konstruktvalide? Eine Überprüfung mittels konfirmatorischer Faktorenanalyse/Is the Charité Alarm Fatigue Questionnaire Construct Valid? An Examination Using Confirmatory Factor Analysis” selbstständig und ohne nicht offengelegte Hilfe Dritter verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel genutzt habe.

Alle Stellen, die wörtlich oder dem Sinne nach auf Publikationen oder Vorträgen anderer Autoren/innen beruhen, sind als solche in korrekter Zitierung kenntlich gemacht. Die Abschnitte zu Methodik (insbesondere praktische Arbeiten, Laborbestimmungen, statistische Aufarbeitung) und Resultaten (insbesondere Abbildungen, Graphiken und Tabellen) werden von mir verantwortet.

Ich versichere ferner, dass ich die in Zusammenarbeit mit anderen Personen generierten Daten, Datenauswertungen und Schlussfolgerungen korrekt gekennzeichnet und meinen eigenen Beitrag sowie die Beiträge anderer Personen korrekt kenntlich gemacht habe (siehe Anteilserklärung). Texte oder Textteile, die gemeinsam mit anderen erstellt oder verwendet wurden, habe ich korrekt kenntlich gemacht.

Meine Anteile an etwaigen Publikationen zu dieser Dissertation entsprechen denen, die in der untenstehenden gemeinsamen Erklärung mit dem Erstbetreuer angegeben sind. Für sämtliche im Rahmen der Dissertation entstandenen Publikationen wurden die Richtlinien des ICMJE (International Committee of Medical Journal Editors; www.icmje.og) zur Autorenschaft eingehalten. Ich erkläre ferner, dass ich mich zur Einhaltung der Satzung der Charité – Universitätsmedizin Berlin zur Sicherung Guter Wissenschaftlicher Praxis verpflichte.

Weiterhin versichere ich, dass ich diese Dissertation weder in gleicher noch in ähnlicher Form bereits an einer anderen Fakultät eingereicht habe.

Die Bedeutung dieser eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unwahren eidesstattlichen Versicherung (§§156, 161 des Strafgesetzbuches) sind mir bekannt und bewusst.“

Datum Unterschrift

Lebenslauf

Mein Lebenslauf wird aus datenschutzrechtlichen Gründen in der elektronischen Version meiner Arbeit nicht veröffentlicht.

Komplette Publikationsliste

1. **Wunderlich, M. M.**, Amende-Wolf, S., Krampe, H., Kruppa, J., Spies, C., Weiß, B., Memmert, B., Balzer, F., & Poncette, A.-S. (2023). A brief questionnaire for measuring alarm fatigue in nurses and physicians in intensive care units. *Scientific Reports*, *13*(1), https://doi.org/10.1038/s41598-023-40290-7
2. **Wunderlich, M. M.**, Couque-Castelnovo, E., Giesa, N., Hueske-Kraus, D., Amende-Wolf, S., Poncette, A.-S., & Balzer, F. (2022). Entwicklung eines Scoring Systems zur Darstellung der Alarmbelastung auf der Intensivstation: Eine Mixed Methods Studie. *Deutscher Anästhesiecongress*, Hamburg, Germany.
3. **Wunderlich, M. M.**, Amende-Wolf, S., Poncette, A.-S., Krampe, H., Kruppa, J., Spies, C., & Balzer, F. (2021). Ein deutschsprachiger Fragebogen zur Messung der Alarmmüdigkeit bei Pflegekräften und Ärzt:innen auf Intensivstationen. *21. Kongress der Deutschen Interdisziplinären Vereinigung für Intensiv- und Notfallmedizin e.V.*, Hamburg, Germany.
4. **Wunderlich, M. M.**, Poncette, A.-S., Spies, C., Heeren, P., Vorderwülbecke, G., Salgado, E., Kastrup, M., Feufel, M. A., & Balzer, F. (2021). ICU Staff's Perception of Where (False) Alarms Originate: A Mixed Methods Study. *International Anesthesia Research Society Annual Meeting*, Hamburg, Germany.
5. Poncette, A.-S., **Wunderlich, M. M.**, Spies, C., Heeren, P., Vorderwülbecke, G., Salgado, E., Kastrup, M., Feufel, M. A., & Balzer, F. (2021). Patient Monitoring Alarms in an Intensive Care Unit: Observational Study With Do-It-Yourself Instructions. *Journal of Medical Internet Research*, 23(5), e26494. https://doi.org/10.2196/26494

Danksagung

Viele Personen haben dazu beigetragen, dass diese Studie erfolgreich verlaufen ist. Besonders möchte ich mich bei Akira-S. Poncette bedanken, dafür dass er mich inspirierte, ehrgeizig zu sein und die bestmögliche Studie zu designen, dabei aber stets pragmatische und kreative Lösungen für Probleme fand. Ein großer Dank geht auch an Nicolas Frey für seine Hilfe in der Logistik der Studie – besonders in der Anfangsphase. Ich danke Henning Krampe für sein kritisches Feedback, mit dem er half, die Arbeit und die Methodik maßgeblich zu verbessern. Ich danke auch Daniel Schulze für seine exzellente Beratung zum methodischen Vorgehen. Auch danke ich Halley Ruppel für die interessanten Diskussionen über das Konzept von Alarm Fatigue, die zum Diskussionsteil dieser Arbeit beigetragen haben. Schließlich bedanke ich mich bei allen Kolleginnen und Kollegen der Kliniken, die diese Studie realisiert haben.

Bescheinigung des akkreditierten Statistikers

CHARITÉ

ChariteCentrum für Human- und Gesundheitswissenschaften

Charité | Campus Charité Mitte | 10117 Berlin

Institut für Biometrie und klinische Epidemiologie (iBikE)

Direktor (komm.): Prof. Dr. Frank Konietschke

Postanschrift:
Charitéplatz 1 | 10117 Berlin
Besucheranschrift:
Sauerbruchweg 3 | CCM

Tel. +49 (0)30 450 562171
frank.konietschke@charite.de
https://biometrie.charite.de/

Name, Vorname: **Wunderlich, Maximilian Markus**

Emailadresse: **maximilian-markus.wunderlich@charite.de**
Personalnummer: **155343**
PromotionsbetreuerIn: **Prof. Dr. Dr. Felix Balzer**

Promotionsinstitution / Klinik: **Institut für medizinische Informatik**

Bescheinigung

Hiermit bescheinige ich, dass Herr *Maximilian Markus Wunderlich* innerhalb der Service Unit Biometrie des Instituts für Biometrie und klinische Epidemiologie (iBikE) bei mir eine statistische Beratung zu einem Promotionsvorhaben wahrgenommen hat. Folgende Beratungstermine wurden wahrgenommen:

- Termin 1: 31.05.2022
- Termin 2: 03.02.2023

Dabei wurde folgende Punkte zu konfirmatorischer Faktoranalyse diskutiert:

- Wahl des Schätzers (ggf. ULS, Bayesianischer Schätzer wurde dargestellt)
- Modellgüte: Fitindizes besprochen.
- Modellmodifikationen: Sollen theoriegeleitet vorgenommen werden. Sollen nicht erfolgen, wenn bereits ein akzeptabler Modellfit erreicht wurde.

Diese Bescheinigung garantiert nicht die richtige Umsetzung der in der Beratung gemachten Vorschläge, die korrekte Durchführung der empfohlenen statistischen Verfahren und die richtige Darstellung und Interpretation der Ergebnisse. Die Verantwortung hierfür obliegt allein dem Promovierenden. Das Institut für Biometrie und klinische Epidemiologie übernimmt hierfür keine Haftung.

Datum: 15.09.23 [redacted] Name des Beraters/der Beraterin: Daniel Schulze

Unterschrift BeraterIn, Institutsstempel