



Missing data imputation techniques for wireless continuous vital signs monitoring

Mathilde C. van Rossum^{1,2,3} · Pedro M. Alves da Silva^{1,4} · Ying Wang^{1,5} · Ewout A. Kouwenhoven³ · Hermie J. Hermens¹

Received: 8 March 2022 / Accepted: 16 January 2023
© The Author(s) 2023

Abstract

Wireless vital signs sensors are increasingly used for remote patient monitoring, but data analysis is often challenged by missing data periods. This study explored the performance of various imputation techniques for continuous vital signs measurements. Wireless vital signs measurements (heart rate, respiratory rate, blood oxygen saturation, axillary temperature) from surgical ward patients were used for repeated random simulation of missing data periods (gaps) of 5–60 min in two-hour windows. Gaps were imputed using linear interpolation, spline interpolation, last observation- and mean carried forwards technique, and cluster-based prognosis. Imputation performance was evaluated using the mean absolute error (MAE) between original and imputed gap samples. Besides, effects on signal features (window's slope, mean) and early warning scores (EWS) were explored. Gaps were simulated in 1743 data windows, obtained from 52 patients. Although MAE ranges overlapped, median MAE was structurally lowest for linear interpolation (heart rate: 0.9–2.6 beats/min, respiratory rate: 0.8–1.8 breaths/min, temperature: 0.04–0.17 °C, oxygen saturation: 0.3–0.7% for 5–60 min gaps) but up to twice as high for other techniques. Three techniques resulted in larger ranges of signal feature bias compared to no imputation. Imputation led to EWS misclassification in 1–8% of all simulations. Imputation error ranges vary between imputation techniques and increase with gap length. Imputation may result in larger signal feature bias compared to performing no imputation, and can affect patient risk assessment as illustrated by the EWS. Accordingly, careful implementation and selection of imputation techniques is warranted.

Keywords Vital signs · Physiological monitoring · Imputation · Missing data · Telemonitoring

Abbreviations

°C degree Celsius
(A)E (Absolute) error

Bpm beats per minute
Brpm breaths per minute
CBP cluster-based prognosis
EWS early warning score
HR heart rate
IQR interquartile range
LI linear interpolation
LOCF last observation carried forward
MAE mean absolute error
MCF mean carried forward
MPE mean percentage error
RR respiratory rate
SD standard deviation
SI spline interpolation
SpO₂ blood oxygen saturation
Temp axillary temperature

Mathilde C. van Rossum and Pedro M. Alves da Silva contributed equally to this work.

✉ Mathilde C. van Rossum
m.c.vanrossum@utwente.nl

- ¹ Biomedical Signals and Systems, University of Twente, Enschede, The Netherlands
- ² Cardiovascular and Respiratory Physiology, University of Twente, Postbox 217, 7500 AE Enschede, The Netherlands
- ³ Department of Surgery, Hospital Group Twente, Almelo, The Netherlands
- ⁴ NOVA School of Science and Technology, NOVA University of Lisbon, Lisbon, Portugal
- ⁵ ZGT Academy, Hospital group Twente, Almelo, The Netherlands

1 Introduction

With the evolution of mobile health technology, the use of wireless sensors for remote vital signs monitoring is rapidly increasing. In a hospital ward setting, wireless monitoring provides the opportunity to measure vital signs continuously, which allows active notification of vital signs abnormalities and evaluation of trends [1, 2]. Accordingly, remote technologies have been deployed to assist early identification of patient deterioration in high-risk surgical or general ward patients [3, 4], and were proposed for monitoring of isolated patients during the COVID-19 pandemic [5]. Furthermore, the continuous data can be used for automated analysis and risk modelling, aiming to support patient monitoring and clinical decision-making. Although standards for the analysis of continuous data in ward patients have not been established as of yet, the sensor data can, for example, be used for the objectification of trends over time based on signal characteristics or for automated calculation of early warning scores (EWS) that are currently used as part of rapid response systems in ward patients [2, 6]. Likewise, the vital signs measurements or extracted signal characteristics can be used as features for advanced event detection algorithms and (machine learning-based) risk prediction models that are increasingly being developed [7].

Despite the potential clinical benefits of remote continuous monitoring and corresponding risk modelling, the processing and interpretation of the data is still a major challenge and hampered by missing and poor quality data [8, 9], resulting in data loss of up to 50% [10, 11]. Measurement disturbances or disruptions are often caused by motion artefacts, which occur frequently during continuous wireless measurements in mobilizing patients [12, 13]. In addition, sensor malfunction or displacement and wireless connection issues can lead to artefacts or data loss [8, 14]. In case the missing or erroneous data periods are not corrected adequately, these segments will hinder the evaluation of vital signs abnormalities and trends. Furthermore, missing data segments will hamper feature extraction and thereby reduce the performance of event detection algorithms, acuity scores, or risk prediction models that are used for clinical decision-making [7, 13–16].

In current practice, retrospective imputation is often applied to substitute periods of missing data or removed erroneous segments in physiological time series data for further analysis or risk modelling. Traditionally, imputation is performed using basic methods such as carry forward techniques or replacement by the patient mean [17, 18]. These basic methods are easy to interpret in clinical practice, and therefore widely used. Yet, various alternative imputation methods that model the dynamic or

personal characteristics of the data have been described more recently, which may be better suited for the evaluation of patterns or for personalized prediction models [16–19]. Although each imputation method has advantages and limitations, it is yet unclear how different imputation techniques perform when used for continuous vital signs monitoring in ward patients, and to what extent imputation could influence further analysis and clinical decision-making. Therefore, the current study aimed to evaluate and compare the performance of various techniques for retrospective imputation of missing data periods, and to explore the impact of imputation on patient monitoring by illustrating the effects on the extraction of basic signal features and calculation of early warning scores.

2 Methods

2.1 Data collection

The current study has a retrospective observational study design. Continuous vital signs recordings were obtained from an existing study database, including data from 60 adult patients that were admitted to the hospital ward for postoperative care after elective oesophageal or gastric surgery or hip fracture surgery in the Hospital Group Twente (ZGT, Almelo, the Netherlands) between 2018 and 2019. Vital signs were obtained every minute using wireless sensors connected to the Patient Status Engine (Isansys Lifecare Ltd., Oxfordshire, UK). The chest-worn LifeTouch sensor was used for measurements of heart rate (HR) and respiratory rate (RR), and the LifeTemp (Isansys Lifecare Ltd., Oxfordshire, UK) sensor was placed under the armpit to record axillary temperature (Temp). Blood oxygen saturation (SpO₂) was measured with a finger probe attached to the wrist-worn Nonin WristOx2 3150 (Nonin Medical Inc., Plymouth, MN, USA). Measurements were performed in parallel to standard care. Both caregivers and patients were blinded for the continuously measured vital signs data. Correct functioning of the sensors was checked regularly during office hours, and measurements were re-established after sensor repositioning, if needed. All data was uploaded to MATLAB (MathWorks, Inc.) for further analysis and simulation. Vital signs recordings were preprocessed by removing values that exceeded the expected physiological range [20] (HR > 200 or < 30 bpm, RR > 50 or < 5 brpm, SpO₂ < 70%, Temp > 50 or < 30 °C). Likewise, samples reporting error codes provided by the system in case of measurement interruptions caused by sensor displacement or disconnection were removed. Furthermore, a 4 min window-based median filter was applied [21].

2.2 Data loss evaluation

To explore the degree of missing data in the current database and thereby evaluate the clinical relevance of data imputation, the percentage of the total recording time where one-minute vital signs samples were missing before and after preprocessing was calculated. In addition, the amount and duration of missing data periods were assessed for each vital parameter. Interruptions longer than 4 h were not included in this count, as these comprise a major part of an eight-hour nurse shift and were therefore not regarded as part of continuous measurements.

2.3 Missing data simulation

Missing data periods ('gaps') were simulated in real uninterrupted continuous vital signs recordings to evaluate the performance of different imputation methods. Figure 1 provides an overview of the main steps of the simulation and evaluation process. In each patient, a maximum of ten windows of three hours each was selected for analysis for each of the vital signs ('analysis window'). Analysis windows were selected subsequently using a sliding window approach, allowing no overlap. Furthermore, windows were only selected in case the concerning vital sign measurement did not contain any missing values. The last two hours of each window was allocated as 'simulation window' and used for simulation of gap segments (Fig. 2). This simulation window size was selected based on the assumption that—although there is no consensus regarding the optimal monitoring frequency [22]—the (average) vital signs values would ideally be updated at least every two hours to enable evaluation of the risk level of ward patients which typically deteriorate in a period of hours [23]. Gap segment simulation was performed by randomly generating one artificial period of missing data within the simulation window. Simulation was repeated 30 times per simulation window, and for gap segment lengths of 5, 10, 15, 20, 30, and 60 min, respectively. For each simulated gap segment, the one-hour window preceding the gap was assigned as the 'pre-gap window', which was used for extraction of prior data characteristics by some of the imputation techniques.

2.4 Imputation techniques

Five different imputation techniques were tested, including the last observation carried forward (LOCF), mean carried forward (MCF), linear interpolation (LI), and spline interpolation (SI) techniques [24], and a cluster-based prognosis technique (CBP). The first four methods were selected because these represent traditional and basic imputation methods that are widely used for physiological signal processing and imputation of vital signs [17, 18, 25–30],

whereas the last method was selected to explore a more advanced technique performing personalized estimation of vital sign patterns [31]. The differences in imputation techniques are illustrated in Fig. 3.

The LOCF technique substitutes all samples in the gap segment by the last sample value prior to the data gap. The MCF technique is a variant of the LOCF method, aiming to estimate the missing data based on a longer measurement period. Accordingly, the MCF technique uses the mean value of the one-hour pre-gap window to fill the gap segment. In the LI technique, the gap segment is substituted by a linear function, which is estimated using the latest sample value prior to the data gap and the first sample value after the gap. Similarly, the SI technique imputes the gap segment with a cubic spline function. The CBP technique is adapted from imputation methods described by Sun et al. [31], where a regression model is used to impute missing data using similar data segments obtained in similar patients. Details of the CBP technique and modifications that were made as compared to Sun's method are described in Supplementary file 1.

2.5 Performance evaluation

The performance of each imputation technique was assessed using the mean absolute error (MAE) and mean percentage error (MPE). The MAE and MPE were calculated for each simulated gap by respectively averaging the absolute or relative difference between the imputed data value (\hat{x}_i) and corresponding original data value (x_i) for all data samples (i) in the gap segment with length l , following Eqs. 1 and 2:

$$MAE_{gap} = \frac{\sum_{i=1}^l |x_i - \hat{x}_i|}{l} \quad (1)$$

$$MPE_{gap} = \frac{\sum_{i=1}^l \left| \frac{x_i - \hat{x}_i}{x_i} \right|}{l} * 100\% \quad (2)$$

As simulation was performed 30 times per analysis window for all combinations of simulated gap length and vital parameters, the MAE_{gap} and MPE_{gap} were averaged across these iterations to obtain the results per analysis window for each of these combinations. The MAE_{gap} values of all analysis windows were evaluated separately for the different gap segment lengths and different vital parameters, to evaluate the range of performance for each imputation technique. The MPE_{gap} was used to explore differences in overall performance between imputation techniques and between vital parameters.

Last, for each vital parameter, the median MAE_{gap} of all simulations performed in assessment windows with 10% lowest and 10% highest original mean value were compared with the median MAE_{gap} of the remaining windows,

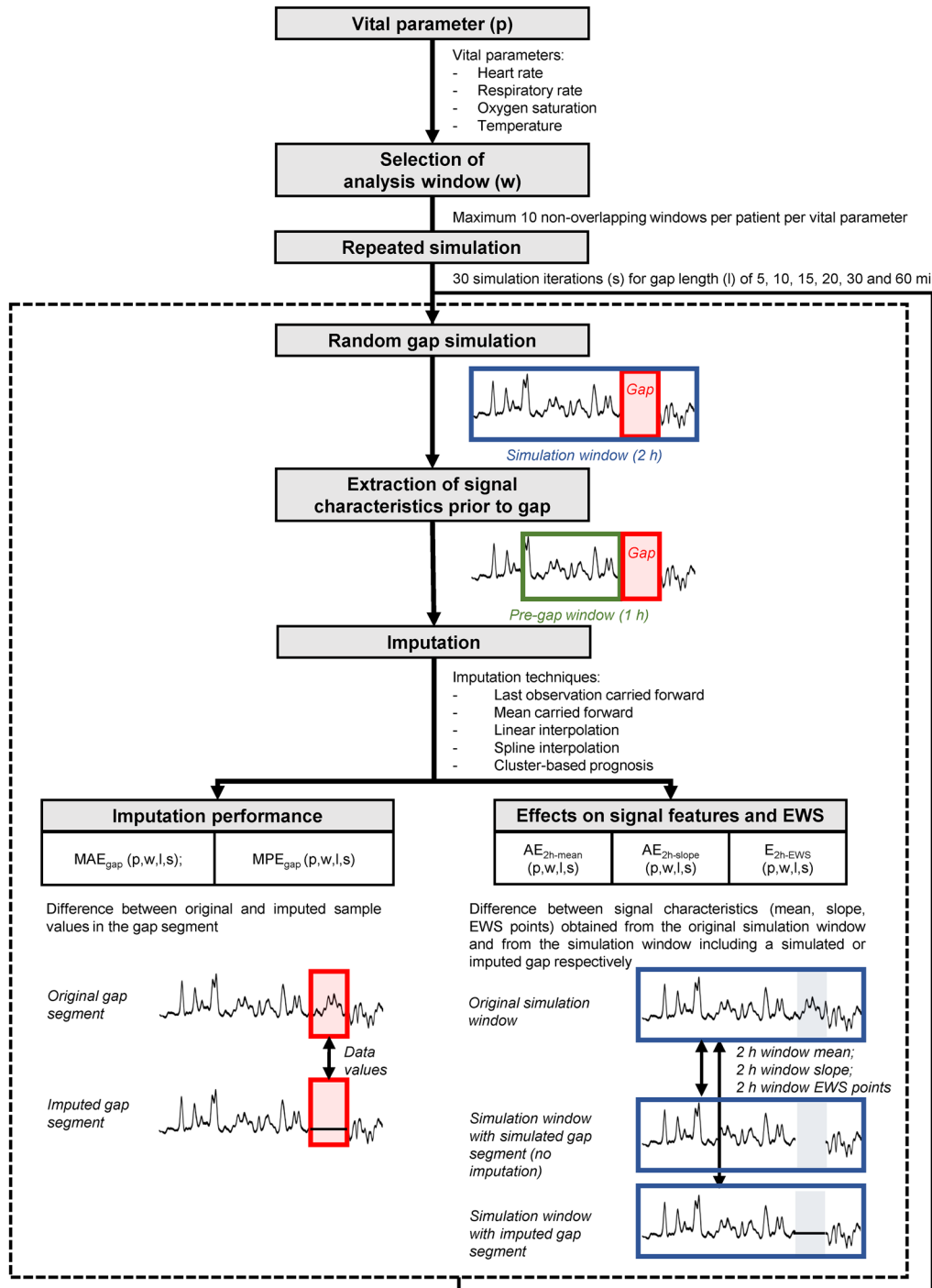


Fig. 1 Overview of the missing data simulation process and evaluation of imputation techniques. MAE_{gap} mean absolute error of the imputed gap segment, MPE_{gap} mean percentage error of the imputed data gap, $AE_{2h-mean}$ absolute error of the mean value of the two-hour

simulation window, $AE_{2h-slope}$ absolute error of the slope of the two-hour simulation window, E_{2h-EWS} error of the EWS points assigned to the two-hour simulation window, EWS : early warning score, p parameter, w window, s simulation iteration, l gap length

aiming to explore the influence of vital sign levels on imputation performance. Likewise, the MAE_{gap} was compared

for assessment windows with highest and lowest standard deviation to investigate the effect of data variability.

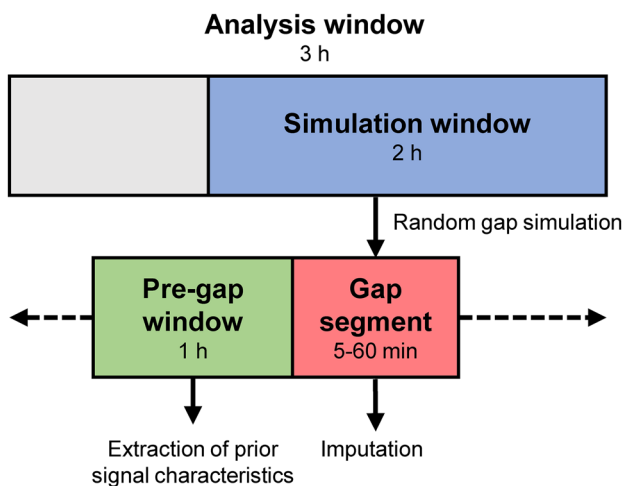


Fig. 2 Illustration of the windows used in the gap simulation process. In each iteration of the simulation process, a missing data period (gap segment) of a predefined length (5, 10, 15, 20, 30, or 60 min) is generated at random within the simulation window and used to test imputation techniques. The pre-gap window is used to extract signal characteristics prior to the gap segment

2.6 Clinical impact exploration

2.6.1 Effects on signal features

In clinical practice, the evaluation of vital signs measurements by caregivers does not only rely on individual vital signs values but also involves evaluation of vital signs trends, i.e., whether vital signs are stable or increase or decrease over time [32]. Although there is still little evidence regarding the clinical value of automated trend assessment methods for vital signs monitoring, studies have indicated

that basic trend metrics such as the average value or slope can contribute to clinical risk prediction models [33, 34]. To explore to which extent imputation may influence the extraction of signal features that could be relevant for trend identification or risk modelling, we compared the mean value and linear slope of the two-hour simulation window before and after imputation. Accordingly, the absolute error (AE) between the mean value of the original two-hour simulation window and the mean value of the simulation window with an imputed gap segment was calculated, resulting in the $AE_{2h-mean}$. In addition, the $AE_{2h-mean}$ was also calculated for the simulation window after deletion of the gap samples, i.e., following an available-case analysis approach, which served as a reference for trend estimation without imputation. Like the $AE_{2h-mean}$, the absolute error was also computed for the slope ($AE_{2h-slope}$), for all imputation techniques, and for the situation without imputation. For the $AE_{2h-slope}$, windows with an original absolute slope value < 0.0025 per hour were excluded as the slope feature was considered clinically irrelevant for stable measurements.

2.6.2 Effects on early warning scores

Early warning scores (EWS) are used widely in clinical wards to assess the risk of patient deterioration. Although many variants exist, the EWS is obtained by assigning points for every vital sign, where the number of points increases for larger deviations from their normal range. The EWS is calculated as the sum of all assigned points and used to trigger further patient assessment or care escalation in case the total EWS exceeds a pre-set threshold [6]. Although vital sign measurements currently rely on nurse observations, there is growing interest to use sensor technologies for (partial)

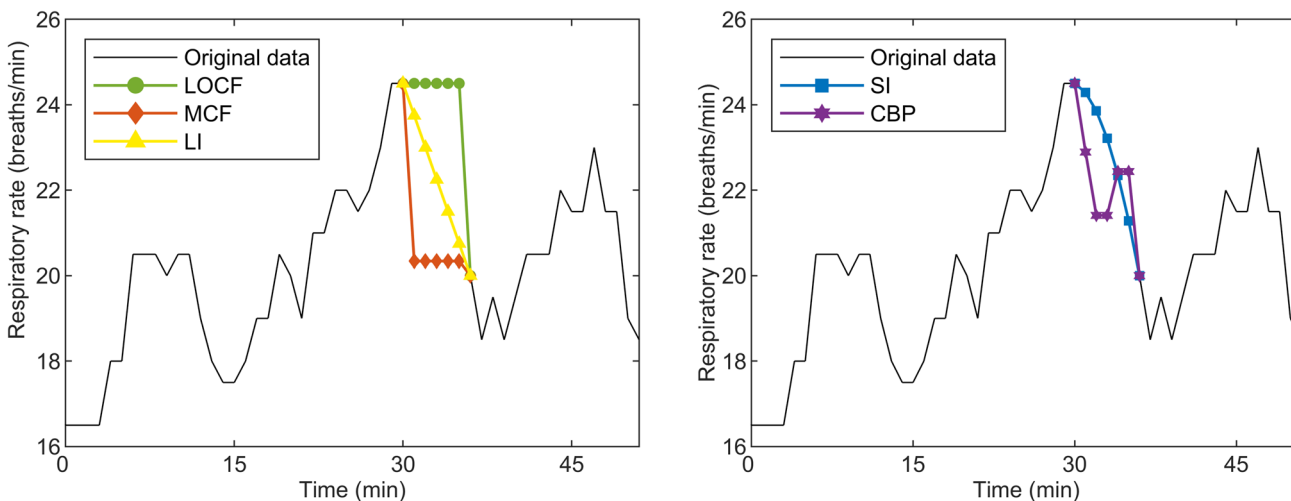


Fig. 3 Illustration of the imputation of missing data by different imputation techniques. *LOCF* last observation carried forward, *MCF* mean carried forward, *LI* linear interpolation, *SI* spline interpolation, *CBP* cluster-based prognosis

automation of EWS measurements [2]. To investigate the possible consequences of imputation on the EWS, we investigated for each vital parameter to what extent the points assigned to the vital parameters obtained from the sensor recordings were affected by imputation. Accordingly, for each simulation, the mean value of the two-hour assessment window was categorized according to the criteria described in Table 1 before and after imputation. The criteria of HR, RR, and Temp were based on the Modified Early Warning Score (MEWS), which is widely used [13]. As SpO₂ is not included in the MEWS, the SpO₂ criteria were obtained from the National Early Warning Score (NEWS) criteria [18]. For each parameter, the error (E_{2h-gap}) between the points assigned to the original window and the window after gap simulation or imputation was assessed. Correspondingly, the number of simulations which resulted in misclassification of the EWS (i.e., $E_{2h-gap} \neq 0$) was calculated.

3 Results

3.1 Data collection

The database included vital signs recordings obtained from 60 hospitalized post-surgical patients, of which 8 patients were excluded due to incomplete demographic

data. A total of 52 patients were included, of which 15 patients experienced one or more complications (Clavien Dindo Class I–III) during the monitoring period. The demographics of the included patients are reported in Table 3 (Supplementary file 2).

3.2 Data loss

The original dataset of included patients contained vital signs recordings with a median duration of 119 h (IQR: 93–147) per vital sign, resulting in a total of 6792 h of monitoring data. The median data availability in these recordings was 86% (IQR: 72–94%) for HR, 86% (IQR: 72–94%) for RR, 46% (IQR: 38–61%) for SpO₂, and 96% (IQR: 81–99%) for Temp. In total, 0.2% of the missing data was related to outlier removal whereas 60% was related to sensor displacement or disconnection as reported by the system. For the remaining missing samples, data was missing without further information. Figure 4 reports the number and total duration of missing data periods up to 4 h that was observed in the original dataset. Most of the gaps that were observed had a duration of 1–5 min, whereas larger gaps were observed less frequently. Nevertheless, the total duration of larger gaps was higher compared to short data gaps.

Table 1 Criteria for early warning score (EWS)

Parameter	EWS points						
	3	2	1	0	1	2	3
Heart rate (bpm)		≤40	41–50	51–100	101–110	111–129	≥130
Respiratory rate (brpm)		≤8		9–14	15–20	21–29	≥30
Oxygen saturation (%)	≤91	92–93	94–95	≥96			
Temperature (°C)		≤34.9		35–38.4		≥38.5	

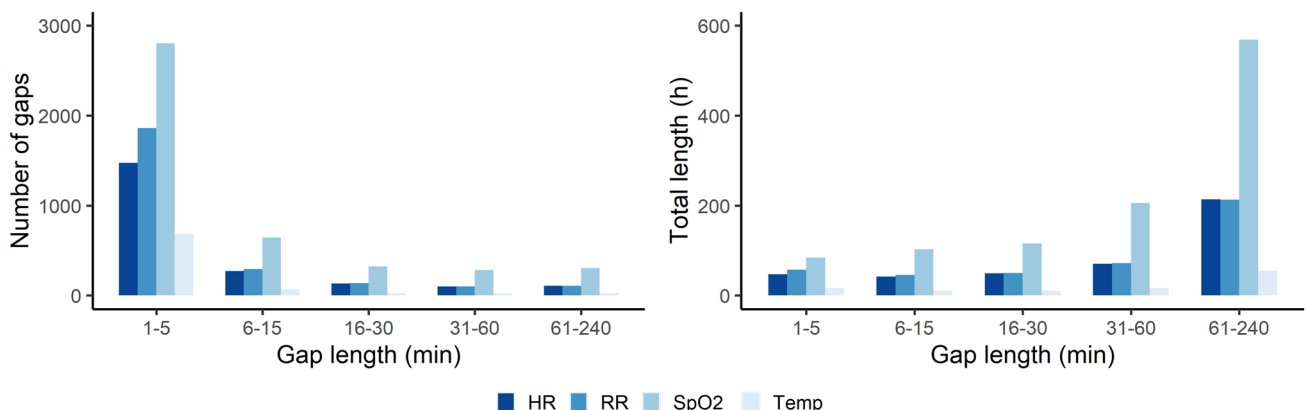


Fig. 4 Number and total length of missing data periods (gaps) with a gap length of up to 4 h observed in the vital signs recordings of the included patient population. *HR* heart rate, *RR* respiratory rate, *SpO₂* blood oxygen saturation, *Temp* temperature

3.3 Missing data simulation

From the original data recordings, a total of 1743 three-hour analysis windows (497 for HR, 492 for RR, 264 for SpO₂, 490 for Temp) were eligible for simulation, with a median of 34 (IQR: 31–39) windows per patient. As gap simulation was repeated 30 times for each gap size in every analysis window, a total of 313,740 gaps were simulated.

3.4 Performance evaluation

Figure 5 reports the MPE_{gap} observed across all gap lengths for each parameter. For the HR, RR, and SpO₂, the median MPE_{gap} and corresponding upper quartile ranges were lowest for the LI technique followed by the CBP and LOCF techniques, but interquartile ranges were relatively large and overlapping. The median and upper quartiles of the MPE_{gap} were highest for the MCF and SI methods. The same performance ranking was found for Temp, except for the fact that SI showed the second lowest median MPE_{gap} . Comparing results between vital parameters, MPE_{gap} ranges were largest for the RR with median MPE_{gap} ranging between 5.5% for LI to 9.7% for SI, followed by the HR (2.0% for LI to 4.1% for MCF), SpO₂ (0.5% for LI to 1.0% for MCF) and Temp respectively (0.2% for LI to 0.7% for MCF).

Looking at the absolute errors across different gap sizes (Fig. 6), MAE_{gap} ranges increased with gap size for all vital parameters, in particular for the SI method. The order of performance was similar as found for the MPE_{gap} results, where LI showed the lowest median MAE_{gap} . The MAE_{gap}

of the LI technique for gaps of 5 to 60 min ranged from HR: 0.9–2.6 bpm, RR: 0.8–1.8 brpm, SpO₂: 0.3–0.7%, and Temp: 0.04–0.17 °C. For small gap sizes, highest error rates were typically found for MCF whereas large gap sizes showed highest errors for SI. The median MAE_{gap} reached values up to 6.5 bpm (SI technique) for the HR, 5.9 brpm for RR (SI technique), 2.1% for SpO₂ (SI technique), and 0.31 °C for Temp (MCF technique) for gaps of 60 min.

Supplementary file 3 reports the MAE_{gap} ranges for all simulations performed in the assessment windows with 10% lowest and 10% highest mean value or standard deviation, respectively. For the HR and RR, the median MAE_{gap} and interquartile ranges were largest for windows with the highest mean value, and lowest for windows with lowest mean, whereas the opposite effect was observed for the SpO₂ and Temp. For all vital parameters, MAE_{gap} ranges were lowest for windows with the lowest standard deviation and highest for the windows with the highest standard deviation. MAE_{gap} varied most between assessment window clusters for the MCF method, followed by the LOCF method.

3.5 Clinical impact exploration

3.5.1 Effects on signal features

The $AE_{2h-mean}$ and $AE_{2h-slope}$ obtained by comparing the mean value and slope of the simulation window before and after simulation are shown in Fig. 7 for the HR, and in Supplementary file 4 for RR, SpO₂ and Temp. As for the MAE_{gap} , the $AE_{2h-mean}$ and $AE_{2h-slope}$ increased with gap

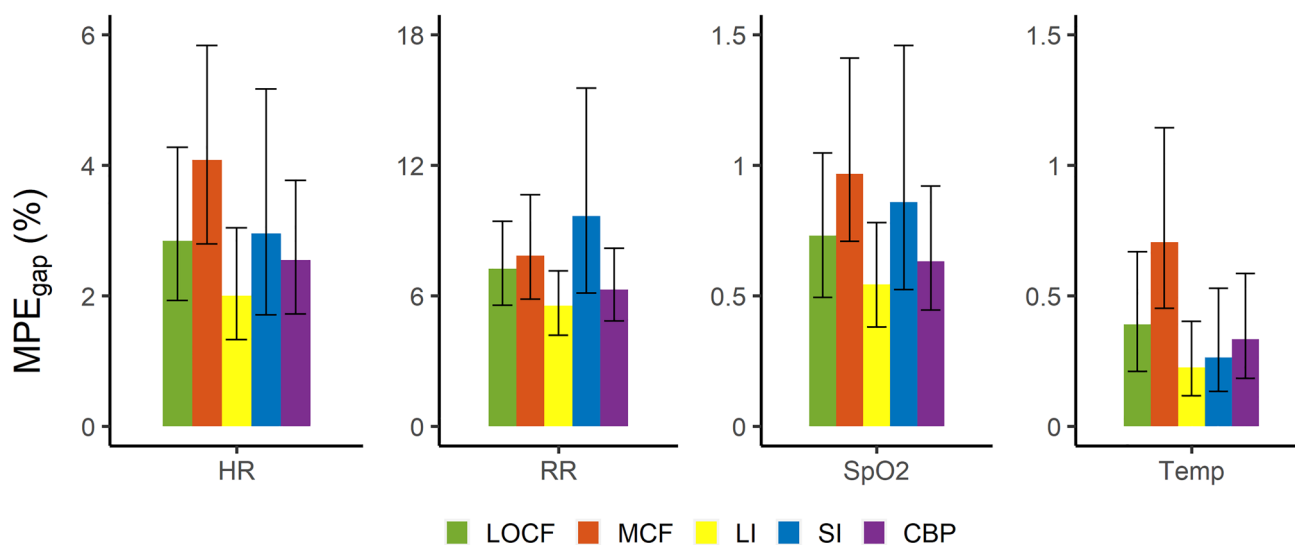


Fig. 5 Mean Percentage Errors (MPE_{gap}) observed for imputation of missing data periods simulated in individual vital signs. The MPE_{gap} is shown as median with interquartile range, and described results found for all simulated gap lengths. *LOCF* last observation carried

forward, *MCF* mean carried forward, *LI* linear interpolation, *SI* spline interpolation, *CBP* cluster-based prognosis, *HR* heart rate, *RR* respiratory rate, *SpO₂* blood oxygen saturation, *Temp* temperature

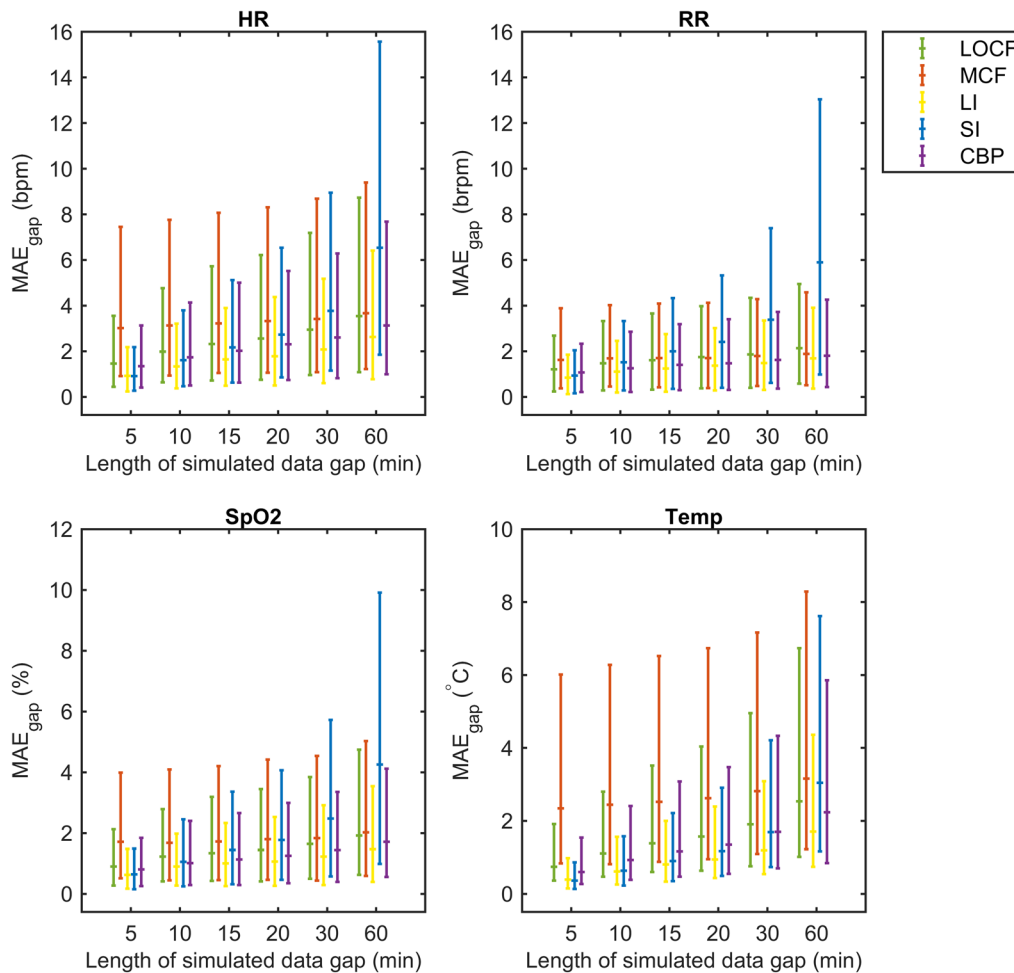


Fig. 6 Mean Absolute Error (MAE_{gap}) observed for imputation of simulated missing data periods (gaps) of 5–60 min length. The MAE_{gap} is shown as median with interquartile range. *LOCF* last observation carried forward, *MCF* mean carried forward, *LI* linear

interpolation, *SI* spline interpolation, *CBP* cluster-based prognosis, *HR* heart rate, *RR* respiratory rate, *SpO2* blood oxygen saturation, *Temp* temperature

segment length. Comparing estimations of the two-hour window mean, the median $AE_{2h-\text{mean}}$ and upper quartiles were lowest for the *LI* or *CBP* techniques for all gap sizes, although interquartile ranges highly overlapped with other techniques. For the slope, the *LI* technique was associated with the lowest median $AE_{2h-\text{slope}}$ for almost all gap sizes, ranging between 0.05 and 0.8 bpm/hour for *HR*, 0.04–0.5 brpm/hour for *RR*, 0.00–0.08%/hour for *SpO2* and 0.02–0.23 °C/hour for *Temp* for gaps of 5–60 min. Comparing trend estimations after imputation to estimations based on non-imputed data, the median $AE_{2h-\text{mean}}$ and $AE_{2h-\text{slope}}$ of the *LI* and *CBP* method and corresponding upper quartiles were lower as compared to performing no imputation for almost all gap sizes in all vital parameters. In contrast, in comparison to no imputation, median $AE_{2h-\text{mean}}$ and $AE_{2h-\text{slope}}$ and upper quartiles were larger

for the highest gap size(s) for the *LOCF* and *SI*, and for all gap sizes for the *MCF* technique.

3.5.2 Effects on early warning scores

Figure 8 presents the percentage of simulations performed in each parameter where the EWS was misclassified (i.e., $E_{2h-\text{gap}} \neq 0$) after gap simulation and imputation respectively. Overall, imputation led to different EWS points in 1–2% of all simulations for *HR* and *Temp*, and between 2 and 7% for *RR* and 2–8% for *SpO2*. Changes were observed in both directions, where the number of simulations with increased points was comparable with the number of simulations with decreased points. In most cases, the EWS increased or decreased one level, resulting in $E_{2h-\text{EWS}}$ of ± 1 points for *HR*, *RR*, and *SpO2* and ± 2 points for *Temp* (see

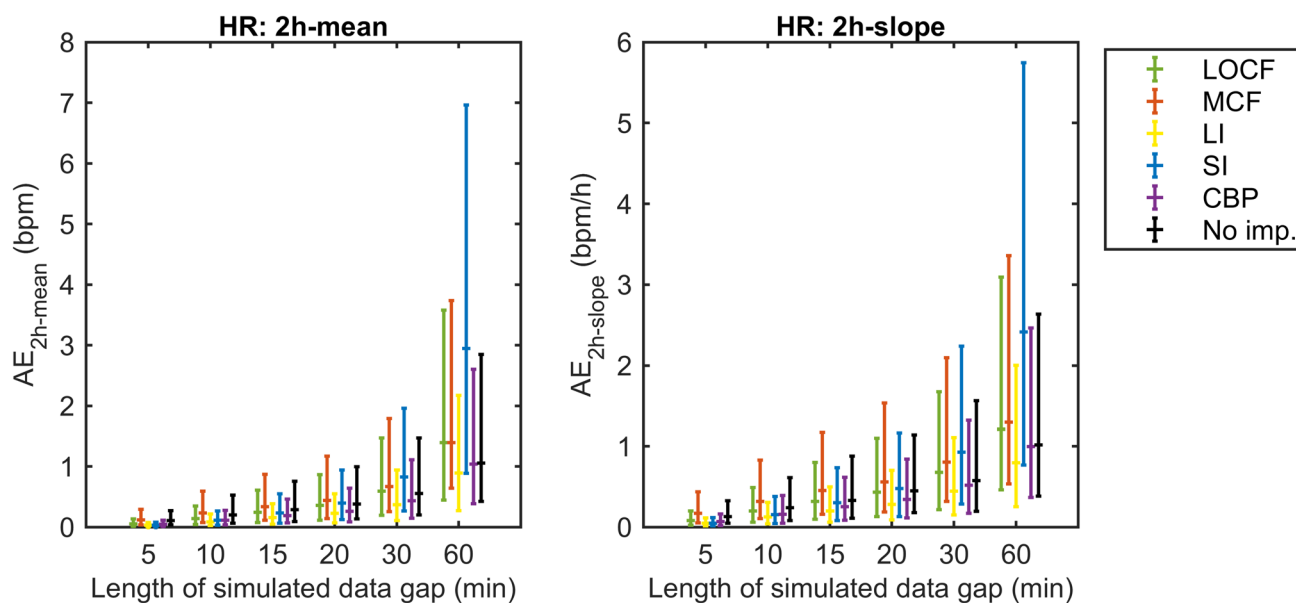


Fig. 7 Absolute error of the mean value ($AE_{2h\text{-mean}}$) and the slope ($AE_{2h\text{-slope}}$) of the two-hour simulation window found for the heart rate (HR). The absolute error is shown as median with interquartile range for different imputation techniques and for the situation with

out imputation. *LOCF* last observation carried forward, *MCF* mean carried forward, *LI* linear interpolation, *SI* spline interpolation, *CBP* cluster-based prognosis, *No imp.* no imputation,

Table 1). Similar to the results presented for the extraction of signal features, imputation using the LI and CBP techniques had a lower impact on EWS calculation compared to performing no imputation, whereas the LOCF, MCF, and SI methods showed more or higher changes in EWS points for several parameters.

4 Discussion

4.1 Main findings

This study explored the performance and related clinical impact of various techniques for imputing missing data periods in continuous vital signs recordings obtained using wearable wireless sensors in postoperative surgical patients. The results indicated that the performance of imputation techniques varied largely between simulation windows, and that imputation errors strongly increased with gap segment length. Of all vital parameters, imputation had the most impact on respiratory rate measurements as suggested by the percentage error rates. Although the error ranges found for the different imputation techniques overlapped, we observed structural differences between the median errors and corresponding interquartile ranges. The LI technique resulted in the lowest median errors and smallest error ranges compared to the other imputation techniques. The largest median errors and error ranges were observed for the SI and MCF techniques. Similar results were found for the signal features

extracted from the two-hour simulation window, where error ranges varied between and within vital parameters, techniques, and gap lengths. The LI and CBP techniques led to lower median bias and a smaller interquartile range of the windows' slope and mean as compared to the deletion of missing data periods. In contrast, however, the MCF, SI, and LOCF techniques were associated with a larger (range of) bias compared to performing no imputation for most gap sizes. Therefore, these techniques can have adverse effects on the accuracy of signal features, and create most uncertainty in further analysis. Imputation led to an increase or decrease in the number of EWS points assigned to vital parameters in up to 8% of all simulations, which illustrates that imputation can affect clinical decision-making.

4.2 Implications

Missing data is a relevant issue in remote vital signs monitoring in ward patients, as observed by the large missing data rates observed in the present study and other studies [10, 11]. Although most data gaps observed in the original recordings had a short duration, larger gaps contributed most to the total duration of missing data, which indicates that imputation is relevant for gaps of variable lengths. The current study highlights the importance of careful implementation and selection of imputation techniques, as error rates strongly varied between and within techniques, in particular for larger gap sizes.

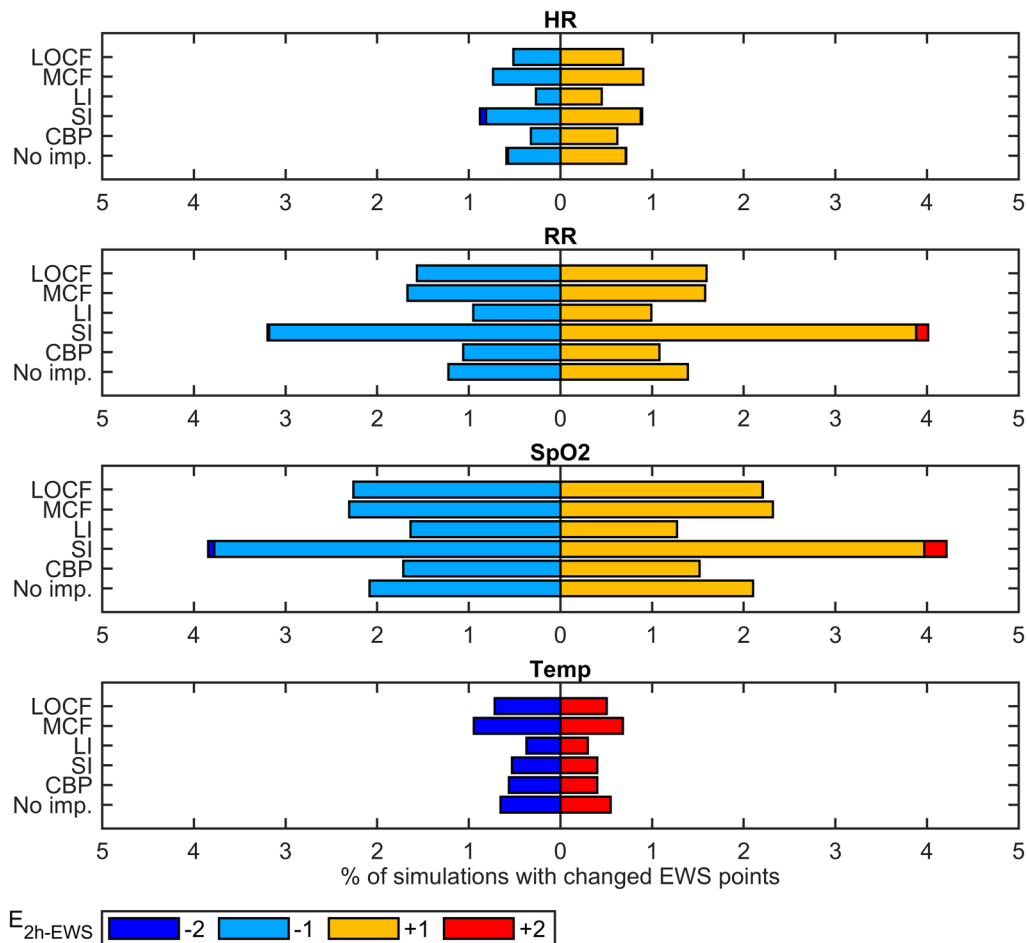


Fig. 8 Percentage of simulations where (no) imputation led to a different number of EWS (early warning score) points assigned to individual vital parameters, compared to the original data. The colors present the error of the EWS points (E_{2h-EWS}), indicating an increase (+1 or +2 points) or decrease (-1 or -2 points) in EWS points.

Although the performance ranges of imputation techniques overlap, LI is suggested as the preferred method for retrospective imputation since this method showed the lowest median error rates and corresponding interquartile ranges and therefore brings the lowest risks of high error rates. Furthermore, this method is simple and therefore relatively easy to implement and intuitively understood by clinicians. This finding is in line with other studies reporting that linear interpolation generally provides higher imputation accuracy in vital signs data compared to other methods [18], and improves the performance of classification models based on physiological data [16]. The CBP technique showed the second-best performance for most parameters. As the CBP technique relies on model training, it can be expected that the performance of this technique will improve with further model optimization using larger datasets tailored to the population of interest. Since the CBP method estimates the dynamical characteristics of the missing data, this or

similar personalized approaches may thereby be considered for intelligent models [15, 16].

LOCF last observation carried forward, MCF mean carried forward, LI linear interpolation, SI spline interpolation, CBP cluster-based prognosis, HR heart rate, RR respiratory rate, SpO2 blood oxygen saturation, Temp temperature

similar personalized approaches may thereby be considered for intelligent models [15, 16]. In the investigation of the window slope and mean, we observed lower median errors and corresponding upper quartiles, compared to performing no imputation for the LI and CBP methods. Therefore, these techniques can improve the accuracy of signal feature extraction in measurements containing missing data periods and reduce the uncertainty in further data analysis. Conversely, we observed that the MCF, LOCF, and SI techniques were associated with larger error ranges as compared to performing no imputation for some or all gap lengths and resulted most often in EWS misclassification. A possible explanation for these observations is that these methods do not (adequately) estimate the variability of data estimations and are affected most by outliers prior to or after the data gap. Correspondingly, we observed that signal variability had the most influence on error rates in these

methods. Therefore, we do not recommend using these techniques for retrospective imputation. These findings are of clinical relevance, as the LOCF and MCF or similar imputation methods are commonly applied for vital signs imputation in early warning scores or other risk prediction models [25–29].

Independent of the technique that is selected, one should be aware that imputation by definition results in data uncertainty, where the possible benefits—compared to performing no imputation at all—but also the risks for clinical decision-making will depend on the size and variability of errors. The median percentage of errors found across all simulations remained below 10% for each vital parameter, which indicates that the clinical risks of imputation are limited in most cases. Correspondingly, the risk that imputation affects the EWS points assigned to individual parameters was 1–8%, which could be reasonable in non-acute settings. On the other hand, the performance of the imputation techniques varied considerably between simulation windows, as reflected by the large interquartile ranges, creating uncertainty for further risk modelling. Besides, the relatively high upper quartiles indicate that there is a considerable risk of large imputation errors, in particular for larger gap sizes. Last, it is likely that missing data periods will be present simultaneously in multiple vital parameters, since measurements often rely on the same sensor or data connection. In this case, the uncertainty of risk models that rely on multiple parameters—such as the EWS—will increase even more. For some clinical applications, these (risks of) high errors are unacceptable, for example when it compromises safety by underestimating risk in unstable patients. As such, it is highly important to assess when the use of imputation is no longer justified.

In practice, the clinical team has to decide which level of uncertainty is acceptable for which patient, and for how long. Obviously, the clinical condition of the patient and corresponding suspicion for deterioration is paramount, as this defines the required level of monitoring. For example, for patients that have been stable for 2 days and are nearing hospital discharge, it will suffice if the care team evaluates general vital sign trends or the risks computed by computer models only once every nurse shift. In these patients, the imputation of gaps of up to one hour could be acceptable, as the overall risks for clinical decision-making and patient safety will be limited. However, patients that have just been discharged from the intensive care unit are often less stable and have a larger risk of serious deterioration. Accordingly, vital sign levels and patient risks need to be assessed more frequently and with higher accuracy levels, as small vital deviations could be critical. In these cases, it can be decided to allow imputation only for data containing short gaps to restrict the uncertainty of data and corresponding decisions, especially because imputation errors seem to be larger in

recordings with larger variability and more extreme measurement values.

In any case, applying imputation should be weighted against alternative methods to compensate for missing data, such as performing weighted or available-case analysis, or abstaining from analysis or decisions in case of incomplete data [35]. In this consideration, relevant factors include not only the possible error rates but also the understandability for clinical staff, the computational time [16], and whether complete data availability is needed for clinically used algorithms or for decision-making [36]. Last, the prevalence, duration, and nature of missing data should be taken into account. According to the classification of missing data as defined by Rubin [37], most of the tested techniques assumed data ‘missing completely at random’ (MCAR) and were also tested by randomly simulating missing data in the current study. However, MCAR assumptions may not always hold in clinical practice [38, 39]. Although technical disturbances such as connection issues are likely to occur completely at random, factors such as skin type or patient activities could systematically influence the likelihood of missing data related to sensor detachment or motion artefacts. In case the missingness is related to known factors and is not related to the signal characteristics of the vital parameter itself, data ‘missing at random’ (MAR) can be assumed. Furthermore, situations where the reasons for missing data are unknown or where missingness is associated with (pathological) vital sign abnormalities can occur, for example when measurements are disturbed by sweating in patients with fever or by motion artefacts related to delirium in deteriorating patients. In these cases, data is assumed to be ‘missing not at random’ (MNAR). As the performance of imputation techniques can be influenced in MAR and MNAR situations, as illustrated by the increased errors ranges found in data windows with larger variability or extreme vital sign levels, further investigation of the circumstances and possibilities to correct for these factors, for example by using accelerometry data, is of interest. Nevertheless, it should be realized that it will often be difficult to identify underlying reasons for missingness as context information is often lacking or cannot be objectified automatically. Therefore, it is recommended that the effects of imputation are validated in the intended care setting.

4.3 Limitations and recommendations

To our knowledge, this is the first study that evaluated imputation techniques for wireless vital signs monitoring in a ward setting. The data used for simulation included many hours of recording but was obtained in a relatively small population including only two patient groups from one hospital. As vital signs characteristics vary between and within patient groups, this could specifically have influenced the results of the CBP method which relies on population

data. To minimize the selection bias, we used random and repeated gap simulation and limited the number of simulation windows per patient. However, gap segments generated in the simulation iterations may have overlapped, in particular, for large gap lengths. Furthermore, gaps were only simulated in data segments with complete data to allow performance evaluation, and may therefore underrepresent situations where missing data is (most) likely to occur in real practice. Together, external validation of results in a larger dataset and for other patient groups is recommended, where MAR or MNAR scenarios are also explored in more detail. Besides, verification of the performance for other sensor systems is desired, taking into account the variable accuracy and different measurement techniques of wearable devices [40, 41].

By comparing estimations of the window slope and mean before and after imputation, we aimed to gain insights into the range of bias that can be expected when extracting signal features relevant for ward patient monitoring. Likewise, we explored possible consequences on clinical decision-making by evaluating changes in EWS points. However, as no standard guidelines for the analysis of continuous data in ward patients exist as of yet, these results are only illustrative. The effects were only investigated for single parameters, whereas a full EWS and other risk prediction models typically rely on multiple vital parameters and also include other clinical variables. Besides, the signal features and EWS points were only obtained in two-hour windows, while dynamic characteristics vary per vital parameter and per individual due to differences in underlying (patho)physiology. Last, the effect of imputation was only studied for a limited range of gap sizes and was not explored for windows with multiple gaps or other data sampling frequencies. Therefore, depending on the diagnostic aims and data characteristics, it might be relevant to verify the effects of imputation on other signal features or when using shorter or longer data windows. Likewise, it is recommended to evaluate the performance of imputation techniques for patterns of clinical interest, for example by exploring pathophysiological data or by comparing stable, linear, and non-linear trend patterns [19].

The current study only investigated a selection of imputation techniques for retrospective monitoring, while many other techniques for imputing missing data in physiological waveforms or data streams have been described [42]. Examples include Kalman-filters [19], Gaussian processes [15], probabilistic data recovery methods using data from related sensors [43], and neural networks [17]. Furthermore, we only investigated the performance of single imputation techniques, which by definition create bias and neglect variability of the missing values in risk models [35]. Methods that account for imputation uncertainty, such as multiple imputation or maximum likelihood methods, could be valuable to reduce bias in decision models [14, 38, 39].

Although the development and evaluation of these and other advanced imputation methods require in-depth analysis of missing data characteristics and relevant covariates—which was beyond the scope of this study—further investigation is highly recommended in future studies that aim to find the best imputation methods for a specific clinical decision model or for real-time monitoring. Likewise, it is of interest to investigate whether errors introduced by imputation methods can be predicted, for example, using historical signal characteristics, activity level, or prior signal quality. This knowledge may help to indicate the accuracy of imputed data and contribute to safe implementation. To encourage further investigation and development of imputation techniques, the dataset used in the current study is available to other researchers on request.

4.4 Conclusion

Imputation of missing data periods in continuous vital signs recordings can be useful to facilitate data analysis for patient monitoring and risk modelling, but imputation errors vary strongly between cases and increase for larger gap sizes. Mean percentage errors differ between vital parameters and are highest for respiratory rate measurements. Although the studied imputation techniques showed overlapping error ranges, errors were structurally lowest for linear interpolation, followed by the cluster-based prognosis technique. Correspondingly, these techniques had the lowest impact on signal features and calculation of early warning scores, and are therefore recommended for retrospective imputation of vital signs measurements. In contrast, spline interpolation or a mean- or last-observation carried forward technique were associated with larger ranges of signal features bias compared to performing no imputation, and can therefore increase the uncertainty for risk modelling. Further investigation of factors influencing imputation errors and evaluation of (acceptable) risks for clinical decision-making is desired to promote safe implementation in clinical care.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s10877-023-00975-w>.

Authors contribution MCR, PMAS, YW, and HJH were contributors to the methodology, analysis, and writing of the manuscript. MCR and EAK contributed to the collection of patient data. All authors reviewed and approved the final manuscript.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declarations

Conflict of interest M.C. van Rossum, P.M. Alves da Silva, Y. Wang, E.A. Kouwenhoven, and H.J. Hermens declare that they have no conflicts of interest.

Ethical approval The current study was performed retrospectively using an anonymized database of the MoViSign study (NL65885.044.18) that was approved by The Medical Research Ethics Committee Twente. Informed consent was obtained from all individual participants included in the MoViSign study.

Consent to participate All included subjects included in the database provided written informed consent to use their data for current research purposes

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Areia C, Biggs C, Santos M, Thurley N, Gerry S, Tarassenko L, et al. The impact of wearable continuous vital sign monitoring on deterioration detection and clinical outcomes in hospitalised patients: a systematic review and meta-analysis. *Crit Care*. 2021;25:351. <https://doi.org/10.1186/s13054-021-03766-4>.
2. Michard F, Kalkman CJ. Rethinking patient surveillance on hospital wards. *Anesthesiology*. 2021;135:531–40. <https://doi.org/10.1097/ALN.0000000000003843>.
3. Posthuma LM, Visscher MJ, Hollmann MW, Preckel B. Monitoring of high- and intermediate-risk surgical patients. *Anesth Analg*. 2019;129:1185–90. <https://doi.org/10.1213/ane.0000000000004345>.
4. Downey CL, Chapman S, Randell R, Brown JM, Jayne DG. The impact of continuous versus intermittent vital signs monitoring in hospitals: a systematic review and narrative synthesis. *Int J Nurs Stud*. 2018;84:19–27. <https://doi.org/10.1016/j.ijnurstu.2018.04.013>.
5. Michard F, Saugel B, Vallet B. Rethinking the post-COVID-19 pandemic hospital: more ICU beds or smart monitoring on the wards? *Intensive Care Med*. 2020;46:1792–3. <https://doi.org/10.1007/s00134-020-06163-7>.
6. García-del-Valle S, Arnal-Velasco D, Molina-Mendoza R, Gómez-Arnau JJ. Update on early warning scores. *Best Pract Res Clin Anaesthesiol*. 2021;35:105–13. <https://doi.org/10.1016/j.bpa.2020.12.013>.
7. Petit C, Bezemer R, Atallah L. A review of recent advances in data analytics for post-operative patient deterioration detection. *J Clin Monit Comput*. 2018;32:391–402. <https://doi.org/10.1007/s10877-017-0054-7>.
8. Weenk M, van Goor H, Frietman B, Engelen JL, van Laarhoven JHMC, Smit J, et al. Continuous monitoring of vital signs using wearable devices on the general ward: pilot study. *JMIR Mhealth Uhealth*. 2017;5:e91. <https://doi.org/10.2196/mhealth.7208>.
9. Breteler MJM, KleinJan EJ, Dohmen DAJ, Leenen LPH, van Hillegersberg R, Ruurda JP, et al. Vital signs monitoring with wearable sensors in high-risk surgical patients: a clinical validation study. *Anesthesiology*. 2020;132:424–39. <https://doi.org/10.1097/ALN.0000000000003029>.
10. Breteler MJM, Huizinga E, van Loon K, Leenen LPH, Dohmen DAJ, Kalkman CJ, et al. Reliability of wireless monitoring using a wearable patch sensor in high-risk surgical patients at a step-down unit in the Netherlands: a clinical validation study. *BMJ Open*. 2018;8:e020162. <https://doi.org/10.1136/bmjopen-2017-020162>.
11. Hernandez-Silveira M, Ahmed K, Ang S-S, Zandari F, Mehta T, Weir R, et al. Assessment of the feasibility of an ultra-low power, wireless digital patch for the continuous ambulatory monitoring of vital signs. *BMJ Open*. 2015;5:e006606. <https://doi.org/10.1136/bmjopen-2014-006606>.
12. Bent B, Goldstein BA, Kibbe WA, Dunn JP. Investigating sources of inaccuracy in wearable optical heart rate sensors. *NPJ Digit Med*. 2020;3:1–9. <https://doi.org/10.1038/s41746-020-0226-6>.
13. Hravnak M, Pellathy T, Chen L, Dubrawski A, Wertz A, Clermont G, et al. A call to alarms: current state and future directions in the battle against alarm fatigue. *J Electrocardiol*. 2018;51:44–8. <https://doi.org/10.1016/j.jelectrocard.2018.07.024>.
14. Azimi I, Pahikkala T, Rahmani AM, Niela-Vilén H, Axelin A, Liljeberg P. Missing data resilient decision-making for health-care IoT through personalization: a case study on maternal health. *Futur Gener Comput Syst*. 2019;96:297–308. <https://doi.org/10.1016/j.future.2019.02.015>.
15. Clifton L, Clifton DA, Pimentel MAF, Watkinson PJ, Tarassenko L. Gaussian processes for personalized e-Health monitoring with wearable sensors. *IEEE Trans Biomed Eng*. 2013;60:193–7. <https://doi.org/10.1109/TBME.2012.2208459>.
16. Kim S-H, Yang H-J, Kim S-H, Lee G-S. PhysioCover: recovering the missing values in physiological data of intensive care units. *Int J Contents*. 2014;10:47–58. <https://doi.org/10.5392/IJoC.2014.10.2.047>.
17. Sharma P, Shamout FE, Abrol V, Clifton D. Data pre-processing using neural processes for modelling personalised vital-sign time-series data. *IEEE J Biomed Heal Informatics*. 2021. <https://doi.org/10.1109/JBHI.2021.3107518>.
18. Nickerson P, Baharloo R, Davoudi A, Bihorac A, Rashidi P. (2018). Comparison of gaussian processes methods to linear methods for imputation of sparse physiological time series. 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4106–9. <https://doi.org/10.1109/EMBC.2018.8513303>
19. Gui Q, Jin Z, Xu W. (2014). Exploring missing data prediction in medical monitoring: A performance analysis approach. 2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB), pp. 1–6. <https://doi.org/10.1109/SPMB.2014.7002968>
20. Pimentel MAF, Clifton DA, Clifton L, Watkinson PJ, Tarassenko L. Modelling physiological deterioration in post-operative patient vital-sign data. *Med Biol Eng Comput*. 2013;51:869–77. <https://doi.org/10.1007/s11517-013-1059-0>.
21. Sow D, Biem A, Sun J, Hu J, Ebadollahi S. Real-time prognosis of ICU physiological data streams. *Annu Int Conf IEEE Eng Med Biol*. 2010. <https://doi.org/10.1109/IEMBS.2010.5625983>.
22. Smith GB, Recio-Saucedo A, Griffiths P. The measurement frequency and completeness of vital signs in general hospital wards: an evidence free zone? *Int J Nurs Stud*. 2017;74:A1–4. <https://doi.org/10.1016/j.ijnurstu.2017.07.001>.
23. DeVita MA, Smith GB, Adam SK, Adams-Pizarro I, Buist M, Bellomo R, et al. “Identifying the hospitalised patient in crisis”—a consensus conference on the afferent limb of rapid response systems. *Resuscitation*. 2010;81:375–82. <https://doi.org/10.1016/j.resuscitation.2009.12.008>.
24. Moritz S, Sardá A, Bartz-Beielstein T, Zaefferer M, Stork J. Comparison of different methods for univariate time series imputation in R. *arXiv*. 2015. <https://doi.org/10.48550/arXiv.1510.03924>.
25. Clifton L, Clifton DA, Pimentel MAF, Watkinson PJ, Tarassenko L. Predictive monitoring of mobile patients by combining clinical observations with data from wearable sensors. *IEEE J Biomed*

- Heal Informatics. 2014;18:722–30. <https://doi.org/10.1109/JBHI.2013.2293059>.
26. Khalid S, Clifton DA, Clifton L, Tarassenko L. A two-class approach to the detection of physiological deterioration in patient vital signs, with clinical label refinement. *IEEE Trans Inf Technol Biomed.* 2012;16:1231–8. <https://doi.org/10.1109/TTTB.2012.2212202>.
 27. Fang AH, Sen, Lim WT, Balakrishnan T. Early warning score validation methodologies and performance metrics: a systematic review. *BMC Med Inform Decis Mak.* 2020;20:1–7. <https://doi.org/10.1186/s12911-020-01144-8>.
 28. Clifton L, Clifton DA, Pimentel MAF, Watkinson PJ, Tarassenko L. Gaussian process regression in vital-sign early warning systems. *Annu Int Conf IEEE Eng Med Biol Soc.* 2012. <https://doi.org/10.1109/EMBC.2012.6347400>.
 29. Tarassenko L, Hann A, Young D. Integrated monitoring and analysis for early warning of patient deterioration. *BJA Br J Anaesth.* 2006;97:64–8.
 30. Morelli D, Rossi A, Cairo M, Clifton DA. Analysis of the impact of interpolation methods of missing RR-intervals caused by motion artifacts on HRV features estimations. *Sensors.* 2019;19:3163. <https://doi.org/10.3390/s19143163>.
 31. Sun J, Sow D, Hu J, Ebadollahi S. A system for mining temporal physiological data streams for advanced prognostic decision support. *IEEE Int Conf Data Min.* 2010. <https://doi.org/10.1109/ICDM.2010.102>.
 32. Mok WQ, Wang W, Liaw SY. Vital signs monitoring to detect patient deterioration: an integrative literature review. *Int J Nurs Pract.* 2015;21:91–8. <https://doi.org/10.1111/ijn.12329>.
 33. Brekke IJ, Puntervoll LH, Pedersen PB, Kellett J, Brabrand M. The value of vital sign trends in predicting and monitoring clinical deterioration: a systematic review. *PLoS One.* 2019;14:e0210875. <https://doi.org/10.1371/journal.pone.0210875>.
 34. Zhu Y, Chiu Y-D, Villar SS, Brand JW, Patteril MV, Morrice DJ, et al. Dynamic individual vital sign trajectory early warning score (DyniEWS) versus snapshot national early warning score (NEWS) for predicting postoperative deterioration. *Resuscitation.* 2020;157:176–84. <https://doi.org/10.1016/j.resuscitation.2020.10.037>.
 35. Little RJA, Rubin DB. *Statistical analysis with missing data.* Hoboken: John Wiley & Sons; 2019.
 36. Dong X, Chen C, Geng Q, Cao Z, Chen X, Lin J, et al. An improved method of handling missing values in the analysis of sample entropy for continuous monitoring of physiological signals. *Entropy.* 2019;21:274. <https://doi.org/10.3390/e21030274>.
 37. Rubin DB. Inference and missing data. *Biometrika.* 1976;63:581–92. <https://doi.org/10.1093/biomet/63.3.581>.
 38. Baraldi AN, Enders CK. An introduction to modern missing data analyses. *J Sch Psychol.* 2010;48:5–37. <https://doi.org/10.1016/j.jsp.2009.10.001>.
 39. Sunny JS, Patro CPK, Karnani K, Pingle SC, Lin F, Anekoji M, et al. Anomaly Detection framework for wearables data: a perspective review on data concepts, data analysis algorithms and prospects. *Sensors.* 2022;22:756. <https://doi.org/10.3390/s22030756>.
 40. Leenen JPL, Leerentveld C, van Dijk JD, van Westreenen HL, Schoonhoven L, Patijn GA. Current evidence for continuous vital signs monitoring by wearable wireless devices in hospitalized adults: systematic review. *J Med Internet Res.* 2020;22:e18636.
 41. Haveman ME, van Rossum MC, Vaseur RME, van der Riet C, Schuurmann RCL, Hermens HJ, et al. Continuous monitoring of vital signs with wearable sensors during daily life activities: validation study. *JMIR Form Res.* 2022;6:e30863. <https://doi.org/10.2196/30863>.
 42. Moody GB. (2010). The PhysioNet/computing in cardiology challenge 2010: Mind the gap. 2010 Computing in Cardiology, pp. 305–8.
 43. Fekade B, Maksymyuk T, Kyrk M, Jo M. Probabilistic recovery of Incomplete sensed data in IoT. *IEEE Internet Things J.* 2018;5:2282–92. <https://doi.org/10.1109/JIOT.2017.2730360>.

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