An Analysis of the Errors in Recorded Heart Rate and Blood Pressure in the ICU Using a Complex Set of Signal Quality Metrics

CW Hug[1], GD Clifford[2]

[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA, USA

Abstract

We analyzed over 95,000 individual values of heart rate and blood pressure derived from 118,000 hours of electrocardiogram (ECG) and 71,000 hours of Arterial Blood Pressure (ABP) data from 1,071 patients using two methods. One method was a nursing-staff verified automatic measurement transmitted from the bedside monitor to central nursing station at intervals of 5 to 60 minutes. The other method involved re-deriving the estimates from continuous ECG and ABP waveforms using independent algorithms and a set of previously described signal quality metrics to reject noisy and untrustworthy data. Results demonstrate that after the removal of obvious artifactual derived HR and ABP estimates, the two measurement sources disagree, on average, by a clinically insignificant amount. Furthermore, after rejection of data using signal quality metrics, the error distribution curve significantly tightens. The clinically-verified BP values exhibit a small but significant bias towards overestimation, both as a function of time of day and as a function of day of the week. Differences in values between time of day and day of week were small but statistically significant. Inter-nurse differences are also described.

1. Introduction

In modern Intensive Care Units (ICUs) physiological data is derived from bedside monitors and transmitted to central nursing stations every minute and verified once every 5 to 60 minutes. These values are sometimes corrected if the nursing staff believe they are erroneous or unrepresentative. Invalid values are also sometimes accepted as correct. To test the validity of these "clinically-verified" (CV) values we compared a subset of parameters to independently derived values for the same parameters, validated using previously tested signal quality metrics [1, 2].

2. Methods

The database used for this evaluation is the MIMIC II database [3] which currently provides nursing data from over 30,000 patients, together with high resolution waveform data (mainly electrocardiogram (ECG) and arterial blood pressure (ABP)) from over 2,700 of these patients. Heart Rate (HR) and systolic/mean/diastolic ABP have been derived for non-overlapping 10s windows (at 0.1Hz) together with signal quality indices (SQIs) for both the ECG and ABP waveforms [1, 2]. The available data set comprises 95,000 individual values of CV HR and average ABP derived from 118,000 hours of ECG and 71,000 hours of ABP data from 1,071 patients.

The SQIs have been calibrated to allow determination of the error in the HR or ABP for a given SQI value. SQI values higher than 0.9 equate to HR errors less than 2 bpm and ABP errors less than 5 mmHg. We consider these to be clinically insignificant. We therefore chose to compare our re-derived values of HR, mean blood pressure (MBP), systolic blood pressure (SBP) and diastolic blood pressure (DBP) possessing SQI values higher than 0.7 with the corresponding clinically verified values. This provides an average SQI of 0.9.

Before making any comparisons, we removed obviously invalid values. This included cases where the DBP, MBP and SBP values where not monotonically increasing. Additionally, Table 1 provides the thresholds used to screen ABP values and HR values that were outside of a physiologically reasonable range. It also includes the total number of matched observations for each measurement after these thresholds and rules have been applied.

Table 1. Value thresholds and observation count

	Min	Max	Observations
SBP	50	240	83095
DBP	30	130	86048
MBP	30	240	89434
HR	15	220	91948

Using each clinically-verified measurement (henceforth indicated with a "CV" prefix) we calculated the median re-derived value (prefix "RD") for the preceding minute (6 samples). Error values were calculated by subtracting the RD values from the CV values; consequently, a negative error reflects underestimation by the nurse and a positive error reflects overestimation by the nurse. We explored the distribution of these errors, how they change as a function of time of day and day of week and whether large differences exist between the accuracy of different nursing staff.

3. Results

Table 2 presents the first four moments of the distributions for the CV, the RD and the SQIRD (filtered with the SQI) values. While the differences between the mean and variances of each distribution are statistically significant ($P<0.0001$ using the F- and t-tests) the magnitudes of the average differences are clinically insignificant. With no filtering, large errors (± 20 mmHg/bpm) are found in about 20% of the values. This rate decreases to 11% with threshold filtering and drops to 4% when the SQI is used.

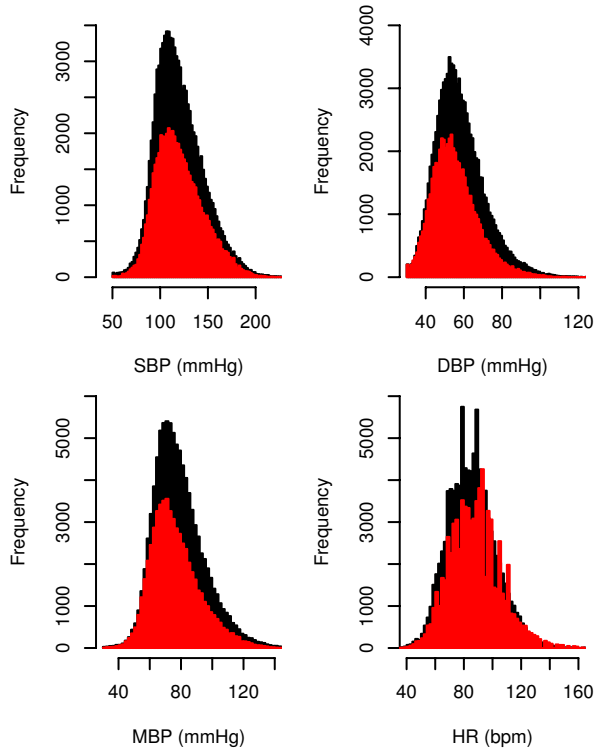

Figure 1. CV and SQIRD Distributions

The SQI threshold (> 0.7) keeps RD values that were derived from clean signals. Figure 1 shows histograms for each signal, where the CV distribution is overlaid with the RD distribution using the SQI filter (SQIRD). While the RD distributions in Figure 1 appear relatively smooth,

Table 2. First four moments of observations: the mean, standard deviation (StdDev), skewness and kurtosis.

	mean	StdDev	skewness	kurtosis
CV_SBP	120.64	25.09	0.61	0.52
RD_SBP	116.78	26.81	0.41	0.26
SQIRD_SBP	120.47	24.76	0.56	0.19
CV_DBP	58.50	13.07	0.88	1.38
RD_DBP	55.41	12.33	0.84	1.54
SQIRD_DBP	55.70	12.15	0.83	1.16
CV_MBP	79.36	16.52	0.92	1.88
RD_MBP	73.64	17.72	0.55	1.11
SQIRD_MBP	76.74	15.75	0.79	0.88
CV_HR	86.17	17.08	0.51	0.97
RD_HR	88.37	21.09	1.18	4.62
SQIRD_HR	87.57	17.81	0.51	0.76

Table 3. Error distribution statistics with and without SQI. The root mean squared error (RMSE), the standard deviation (StdDev), the skewness and the kurtosis.

	RMSE	StdDev	skewness	kurtosis
SBP	19.47	19.02	1.24	7.42
SBP (SQI)	12.34	12.25	0.20	13.24
DBP	9.53	8.99	0.76	8.63
DBP (SQI)	6.52	5.93	0.71	17.20
MBP	16.18	15.07	1.63	7.64
MBP (SQI)	8.74	8.33	0.93	16.55
HR	15.55	15.37	−2.35	23.70
HR (SQI)	7.91	7.77	−0.42	32.34

without the SQI filter the RD distributions contain a variety of abnormal peaks. Consequently, the distributions of the differences between the CV and RD values tighten significantly when the SQI is employed (see Figure 2). Table 3 provides the first four moments of these distributions.

Notice from Table 3 that the bias (skewness) and the randomness (Gaussianity) are reduced when the SQI is used to filter the data. This indicates that the errors are systematic. As a result, analyses of the error distributions require nonparametric methods such as the Mann-Whitney U test to compare the means and the Fligner-Killeen test to compare the variances. It should also be noted that the errors between the two distributions decrease as the SQI threshold is increased (Figure 3). This analysis uses the same 1-minute, 6-sample median for the signal value and for the SQI metric. However, similar results are observed for longer windows and when using the mean values instead of median values. This indicates that the patients are relatively stable and the window over which the measurements are averaged has little impact on the actual physiological parameter estimate.

3.1. Diurnal variations and beyond

To examine the influence of the time of day and day of week on the errors, we looked for consistent errors in the CV values relative to the RD values. As expected from the

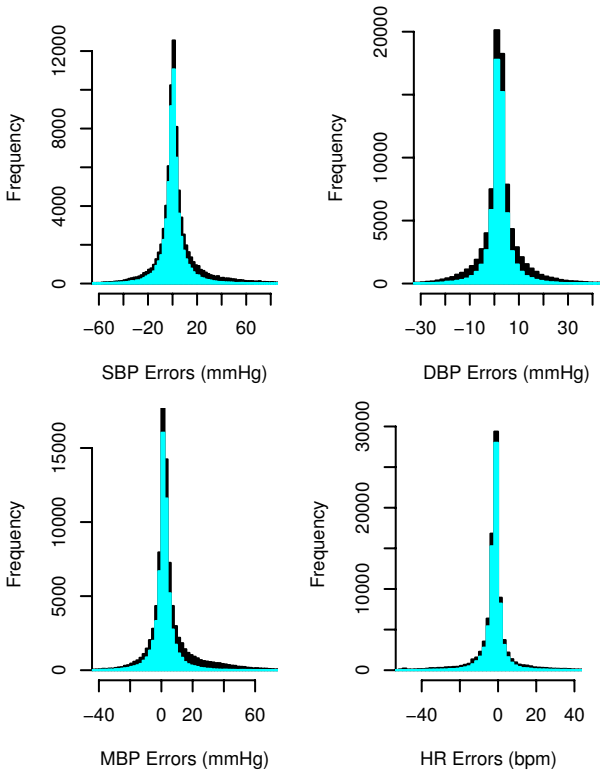

Figure 2. Error distributions (CV - SQIRD)

tight error distributions, most of the CV values are close to the SQIRD values with a difference that is unlikely to be clinically significant on average. However, large outliers do exist. Figure 4 shows the errors between $+60$ and -60 mmHg for SBP as a function of hour and as a function of day (The MBP, DBP and HR have similar temporal signatures). The center line represents the mean of the error distribution and the surrounding lines represent one and two standard deviations from this mean. Each of the ABP variables demonstrate consistent overestimation by the nursing staff. The recorded HR observations exhibit a small underestimation bias.

We conducted a Mann-Whitney U test to examine if the average errors on a particular day or at a particular hour deviated significantly from the average errors for the other days or hours. Similarly, we performed a Fligner-Killeen test to look for significant changes in variance between particular days or hours and the other days or hours.

The average error and the variance in this error decrease during the middle of the week. On Wednesday the average error for each of the four measurements is significantly lower ($p < 0.05$) than the other six days in the week. The variance on Wednesday is also significantly lower ($p < 0.0001$). Similarly, the errors increase on the weekend. The errors on Friday are significantly higher ($p < 0.005$ for ABP, $p = 0.038$ for HR). The variance on

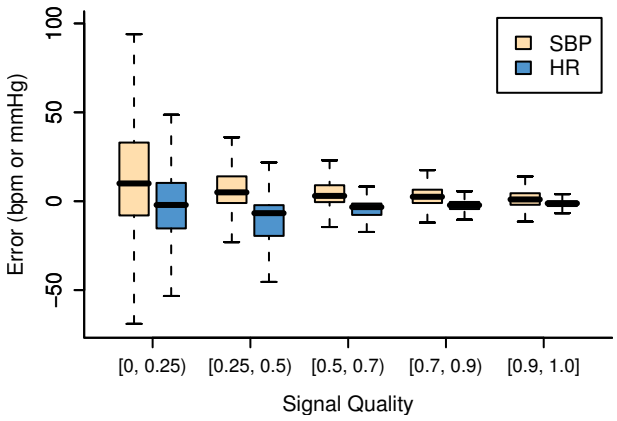

Figure 3. Interquartile range (IQR) box plot of HR and SBP errors for different SQI intervals (DBP and MBP demonstrate similar behavior). Note that the bars extend $1.5 \times IQR$ from each box and that outliers are not shown.

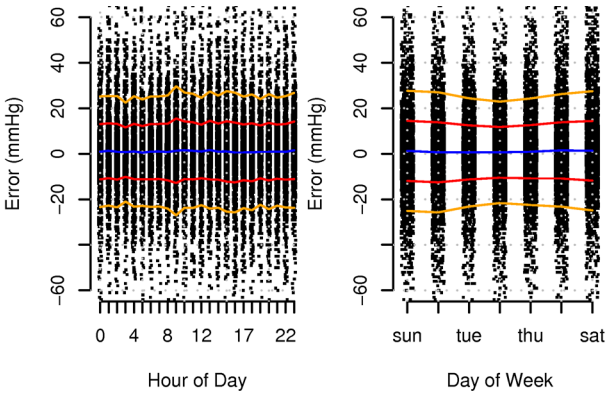

Figure 4. SBP errors vs time

Friday is not significantly different from the other days but on Sunday and Monday, the variance is significantly higher ($p < 0.0001$). Diurnal differences are also present. There is a noticeable tightening of the ABP errors at around 5 am ($p < 0.005$) and 1 pm ($p < 0.05$). The HR errors are only significantly different from the other measurements at 11 pm ($p = 0.025$) where they appear to tighten. The variance in HR and ABP errors decreases at 3 am ($p < 0.005$) and increases for ABP errors at 8 am ($p < 0.005$) and 11 pm ($p < 0.05$).

3.2. Care giver comparisons

Finally, we examined the average error of care givers that have at least 10 recorded observations. Histograms constructed from the results are shown in Figure 5. One striking observation from these distributions is the number of care givers who consistently overestimate or underestimate by a significant value.

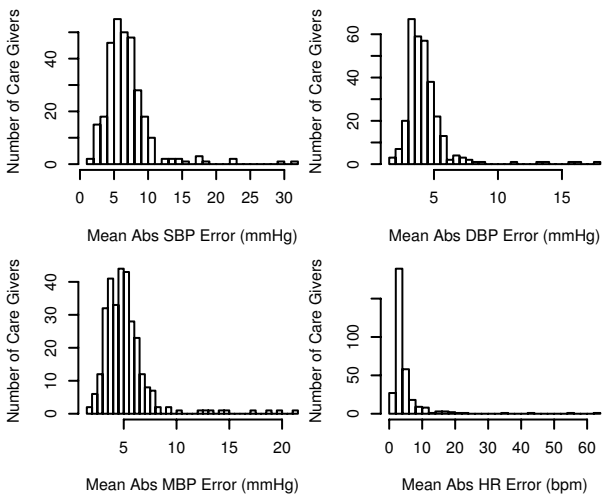

Figure 5. Average absolute error by care giver

Approximately 20% of the observations are from anonymous care givers who failed to enter their identification number. Table 4 shows the number of identified and anonymous observations after filtering using the SQI metric. It also includes the number of unique care givers recorded in the data. A comparison of the errors by identified care givers and anonymous care givers is given in Table 5. With the exception of DBP, the average errors from this anonymous group of care givers were significantly higher than the identified group.

4. Conclusions

Automated SQIs can lead to lower errors in clinically-verified data, potentially leading to improved patient care, where a significant number of HR and ABP readings are verified with errors that are clinically significant.

These errors vary with the hour of the day and day of the week, with significantly lower errors presenting mid-week and twice a day (early morning and early afternoon). This may be a function of changing work load and staffing levels, shift changes and perhaps even circadian and diurnal variations in alertness. Such performance factors require further investigation.

Finally, we detected a significant variation in error levels (mean and variance) between identified care givers and those that did not identify themselves. Furthermore, nursing staff sometimes leave themselves 'logged-on' and therefore those that may not normally identify themselves could be entering data as an identified care giver. Such cases should reduce the difference we observe between errors in the identified and anonymous care giver groups from its actual level. This may indicate that care givers who identify themselves are more diligent and compliant with procedures that may seem irrelevant to nursing care, yet can have an impact on data accuracy. That is, either the care giver identification procedure encourages a more accurate reporting of data, or those care givers that identify themselves are more diligent in checking the accuracy of data. Such information could be used to improve patient care.

Table 4. Identified vs anonymous observations

	SBP	DBP	MBP	HR
Identified	48726	48535	48666	61791
Anonymous	9973	9899	9977	12739
Unique care givers	494	494	494	495

Table 5. Average error for identified (ID) versus anonymous (An) care givers. † = statistical significance.

	SBP	DBP	MBP	HR
ID RMSE	11.6	6.2	8.2	7.8
An RMSE	15.6	7.8	11.0	8.4
P value	$< 10^{-2}$†	0.058	10^{-2}†	$< 10^{-2}$†
ID Var	132.1	31.6	60.1	58.9
An Var	238.6	52.8	112.8	67.6
P value	$< 10^{-3}$†	$< 10^{-3}$†	$< 10^{-3}$†	$< 10^{-3}$†

Acknowledgments

This work was supported in part by the National Library of Medicine (NLM) Medical Informatics Traineeship (LM 07092), the U.S. National Institute of Biomedical Imaging and Bioengineering (NIBIB) and the National Institutes of Health (NIH) under Grant Number R01 EB001659, Philips Medical Systems and the Information and Communication University (ICU), Korea. The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the NLM, the NIBIB, the NIH, Philips Medical Systems, or ICU Korea.

References

[1] Li Q, Mark RG, Clifford GD. Robust heart rate estimation from multiple asynchronous noisy sources using signal quality indices and a Kalman Filter. IOP Physiol Meas July 2007; In Submission.

[2] Li Q, Mark RG, Clifford GD. Artificial arterial blood pressure artifact models and an evaluation of a robust blood pressure and heart rate estimator. Biomedical Engineering Online 2007;In Submission.

[3] Saeed M, Lieu C, Raber G, Mark RG. MIMIC II: a massive temporal ICU patient database to support research in intelligent patient monitoring. Computers in Cardiology 2002; 29:641–644.

Address for correspondence:

Caleb Hug
MIT CSAIL 32 Vassar St #257
Cambridge, MA 02139 USA
hug@mit.edu