Current clinical methods of measurement of respiratory rate give imprecise values

Gordon B. Drummond[1], Darius Fischer[2] and D.K. Arvind[2]

Affiliations: [1]Dept of Anaesthesia, Critical Care, and Pain Medicine, University of Edinburgh, Edinburgh UK. [2]Centre for Speckled Computing, School of Informatics, University of Edinburgh, Edinburgh, UK.

Correspondence: Gordon B. Drummond, Dept of Anaesthesia, Critical Care and Pain Medicine, University of Edinburgh, Edinburgh Royal Infirmary, 51 Little France Crescent, Edinburgh, EH16 4SA, UK. E-mail: g.b.drummond@ed.ac.uk

ABSTRACT
Background: Respiratory rate is a basic clinical measurement used for illness assessment. Errors in measuring respiratory rate are attributed to observer and equipment problems. Previous studies commonly report rate differences ranging from 2 to 6 breaths·min^{-1} between observers.
Methods: To study why repeated observations should vary so much, we conducted a virtual experiment, using continuous recordings of breathing from acutely ill patients. These records allowed each breathing cycle to be precisely timed. We made repeated random measures of respiratory rate using different sample durations of 30, 60 and 120 s. We express the variation in these repeated rate measurements for the different sample durations as the interquartile range of the values obtained for each subject. We predicted what values would be found if a single measure, taken from any patient, were repeated and inspected boundary values of 12, 20 or 25 breaths·min^{-1}, used by the UK National Early Warning Score, for possible mis-scoring.
Results: When the sample duration was nominally 30 s, the mean interquartile range of repeated estimates was 3.4 breaths·min^{-1}. For the 60 s samples, the mean interquartile range was 3 breaths·min^{-1}, and for the 120 s samples it was 2.5 breaths·min^{-1}. Thus, repeat clinical counts of respiratory rate often differ by >3 breaths·min^{-1}. For 30 s samples, up to 40% of National Early Warning Scores could be misclassified.
Conclusions: Early warning scores will be unreliable when short sample durations are used to measure respiratory rate. Precision improves with longer sample duration, but this may be impractical unless better measurement methods are used.

@ERSpublications
In acutely ill patients, the usual 30 s to count breathing are insufficient to give a reliable measurement https://bit.ly/31DNhKU

Introduction

Respiratory rate is universally employed in the clinical assessment of ill patients and is now used widely in early warning scores to grade severe illness in acutely ill patients in an emergency setting [1]. Doubts have been raised about the reliability of such observations [2, 3], but measurement error is rarely considered [4].

Studies of respiratory rate tend to assume that breathing is stable and disregard breath-to-breath variation. In fact, variation from breath to breath can be substantial and appears random. When comparing alternative measurement methods, observations can differ unless exactly the same time periods are considered by each method. Thus, discrepancies between devices may not result from measurement error, as previous studies have assumed [5]. We could find no systematic studies of the repeatability of respiratory rate measurements, so we investigated the imprecision of clinical measurements of respiratory rate by simulating repeated measurements.

The obvious factor affecting repeatability is sample size. Random variation exerts a greater effect in small samples (the law of large numbers was first described by Gerolamo Cardano in the 16th century). We studied observations of different durations (*i.e.* sample size) to assess how this affected repeatability.

Methods

We used records made in a previous study in which we assessed a new device to measure respiratory rate. For that study, we had recruited a convenience cohort of adult patients who were admitted to hospital with acute illness.

Patients were studied in the acute admission unit of a 570-bed teaching hospital. All were studied in the first 4–8 h after admission. Nasal cannula pressure was recorded in each patient. This continuous measure allows precise measurement of successive breath durations over a greater time period than would be clinically feasible (figure 1). The sole criterion for admission to the study was that the patient accepted the nasal cannula placement. If possible, respiratory signals were recorded for 1 h, or until the patient was prepared for discharge or transfer from the ward. These data gave us a unique chance to simulate repeated clinical measurements of breathing rate. We designed the current study to present results in terms familiar to clinical workers.

For this previous study, we were granted ethical permission by the Scotland A Research Ethics committee (ref 12/SS/0054) subject to the provisions of section 51 of the Adults with Incapacity (Scotland) Act 2000. This legislation was relevant because we could not be certain that acutely ill patients were capable of giving full informed consent. We were allowed to retain only limited patient data (year of birth, height and weight) but were permitted to process anonymised data recordings and retain them for further use.

Patient recordings

To obtain a precise measure of breathing rate (intended for comparison with the new device) we recorded pressure at the nostrils with a nasal cannula (Sleep Sense 15802–2; Medes Ltd., Radlett, UK). This is a well-established and reliable method often used in sleep studies [6]. A single-use set of nasal cannulae were placed below the nostrils. The cannulae were connected through a bacterial filter to a battery-powered pressure transducer (PTAF2; Philips Respironics, Chichester, UK; www.philips.com/respironics). This was placed beside the patient and the pressure signal was transmitted wirelessly (Bluetooth LE) to an iPod receiver. The pressure signal was digitised at 12.5 Hz. After each study period, the patient recordings were transferred from the receiver to a secure computer for further analysis.

The records were analysed using proprietary software (Spike2, version 5.19; CED, Cambridge, UK). Each breath onset time was identified and recorded automatically using a threshold detection facility in the display software, to give a sequence of times (accuracy >0.1 s) from the start of the record (figure 1a). An overall respiratory rate was calculated for each patient, in breaths per minute, by dividing the total number of complete breath cycles in the record by the total duration of those cycles. This value represents the most exact measure of breathing rate for that patient (figure 1b).

Each patient record of breath times was then randomly sampled, on multiple occasions, to select time segments, each of specified duration (figure 1c). The durations of these random samples were nominally 30, 60 and 120 s: the 30 s duration was chosen because it is common in clinical practice (most data sets show a substantial excess of even values [3]), 60 s because it is an ideal duration [7] and 120 s to assess the effect of a larger sample, although this is rarely used clinically. The respiratory rate was calculated from the whole number of breaths and the overall duration of those breaths, to the nearest 0.1 s, with the total duration as near as possible to the chosen value. The number of random sample periods taken from each patient record was related to the size of the recording: approximately one random sample was taken for each minute of each recording. Since the values were mostly not normally distributed, we expressed variation in repeated measures by the interquartile range of the observed rates. For each patient, and each
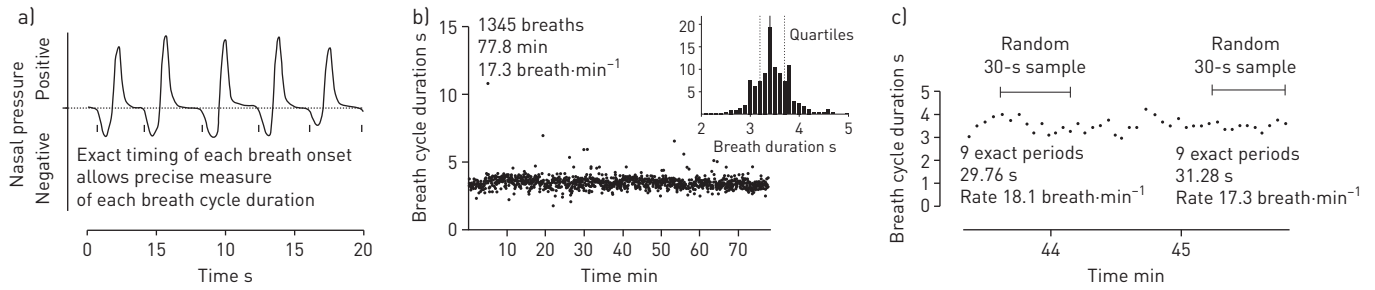
FIGURE 1 a) Example of a nasal pressure trace. Nasal pressure decrease at the onset of each inspiration is detected automatically. Breath duration is calculated from the time between each mark. b) Each patient record consists of successive breath durations. The distribution of breath duration is shown in the insert, with the quartiles shown. c) Random samples are taken from this series, using a whole number of breath cycles, as close to 30 s as possible.

sample duration, the median and interquartile range of all the rate estimates was calculated. Figure 2a shows an example of the rate values obtained, taken from the record of the patient shown in figure 1.

We used the measurements from the entire cohort of patients as a model population. These data allowed us to derive the likelihood of a repeat observation taken from a single patient, drawn from a population of patients similar to those we had studied. The model population allows incorporation of variance within and between patients, and the procedure is described in supplementary appendix 1.

Finally, we estimated how variations in the observed rate might affect the respiratory element of the UK National Early Warning Score allocated to a particular patient. We took the overall respiratory rate, based
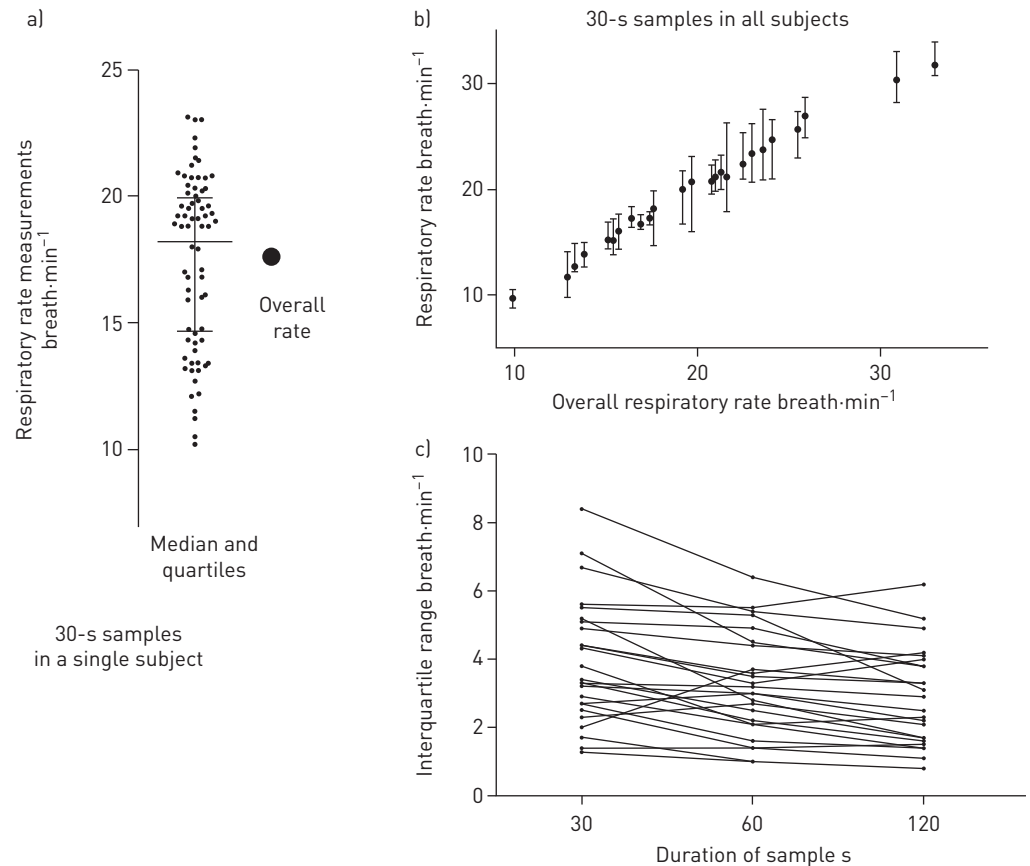
FIGURE 2 a) The distribution of separate rate value measurements (using 30 s samples) for the patient shown in figure 1, with median and quartile values, in relation to overall respiratory rate (shown as ●). b) Measurements from all subjects, with median and interquartile range, in relation to the overall respiratory rate for each subject, using a 30 s sample duration. c) The influence of a greater duration of sample on the interquartile range of the observations. Observations from each subject are linked.

on the entire record, as the true respiratory rate of the patient. We then estimated the proportion of observations made from the patient that would have resulted in the same score value.

Data were processed using Python scripts and GraphPad Prism version 6.05 for Windows (GraphPad Software, La Jolla, CA, USA; www.graphpad.com). Unless otherwise stated, data are summarised as the median (quartiles). The original breath time data used are available as a web appendix (appendix 2).

Results

We used recordings of >30 min duration obtained from 25 patients. These patients (11 female, 14 male) had a mean (SD) age of 66 (15) years, weight 81 (16) kg, height 1.67 (0.12) m and BMI 28 (8) kg·m^{-2}. They were admitted to hospital with a variety of acute medical conditions, the most frequent being respiratory (7), cardiac (4), neurological (4) and urinary (4). Most patients had intercurrent disease, predominantly cardiac and respiratory.

Plots of the breath durations against elapsed time (as in figure 1b) did not show a trend in any of the patients that could cause variation of the measured rates. The median number of rate estimates made from the patients was 62 (quartiles 54, 81). The scatter of rate values around the median varied from patient to patient. As would be expected, the median of the rates from a patient was always close to the overall respiratory rate, measured from the entire sample from that patient. Each patient generated three sets of rate measurements, making 75 sets of rate measurements in total.

Figure 2b shows the median and interquartile range of all the rate measurements in each patient, based on samples of 30 s, plotted in relation to the overall respiratory rate. The distribution of rates was normal in only 35 of the 75 sets of rates considered (D'Agostino & Pearson omnibus test).

The interquartile ranges for samples of 30, 60 and 120 s are shown in figure 2c. For samples of 30 s, the mean interquartile range of the rate estimates was 3.4 breaths·min^{-1}. For samples of 60 s, the mean interquartile range was 3.0 breaths·min^{-1} and for 120 s samples was 2·5 breaths·min^{-1}.

Using the model described in appendix 1, we predicted respiratory rates if a specific rate observation, taken from our studied population, were to be repeated. We selected observation values of 12, 20 and 25 breaths·min^{-1}, which are threshold values in the UK National Early Warning Score. The predictions are shown in figure 3. Particularly for the 30 s samples, the possibility that a repeat observation would be within 2 breaths·min^{-1} of the previous value is <50%.

Overall, the likelihood of a repeat measure yielding the correct respiratory component of the UK National Early Warning score was 70%. However, this proportion varied considerably, from 27% to 100%, depending on the particular value of the overall respiratory rate. For patients with rates in the middle of the normal range, between 15 and 17 breaths·min^{-1}, a large proportion of scores were correctly allocated. However, for patients with overall respiratory rates between 19 and 25 breaths·min^{-1} (i.e. 11 out of the 25 patients), correct classification was on average 47%. If the sample duration was increased to 120 s, this proportion increased to 54% (figure 4b).

Discussion

We have shown that routine clinical measurements of respiratory rate taken in unwell patients vary substantially and do not necessarily match the breathing rate averaged over a longer time. Many clinicians already suspect that small samples of a variable feature such as breathing can give imprecise results. Variation in repeat observations, within patients, has already been described [5, 7] but has been attributed to inter-observer or inter-device variation. For example, in a widely cited study of 140 emergency patients, two trained observers measured respiratory rate, sequentially, each over a 1 min period [5]. The limits of agreement between pairs of measures was 5.4 breaths·min^{-1}, and was attributed to "interobserver variability". The authors did not consider the possibility that two measures of a patient's respiratory rate, within a 5 min time frame, could be very different. When the inherent variation of the signal is recognised, the need for much longer periods of observation to obtain reproducible values becomes evident, as noted by Tobin and co-workers [8] who showed that 15 min averages were required for consistent results.

Our measurements of respiratory rate used an integer number of breaths and a precisely measured time and are more precise than usually possible clinically by a trained observer. At the bedside, it is often the peak of an inspiration that is counted as a breath, which may be counted as the first in the observation period, whereas it is really the zero. However systematic errors of this type will not affect the outcome measure of interest, which is variation between repeated measures.

We studied a convenience cohort of clinically unwell patients with a wide range of respiratory rates (see figure 2b), and our findings should be generalisable to many emergency patients. However, clinical
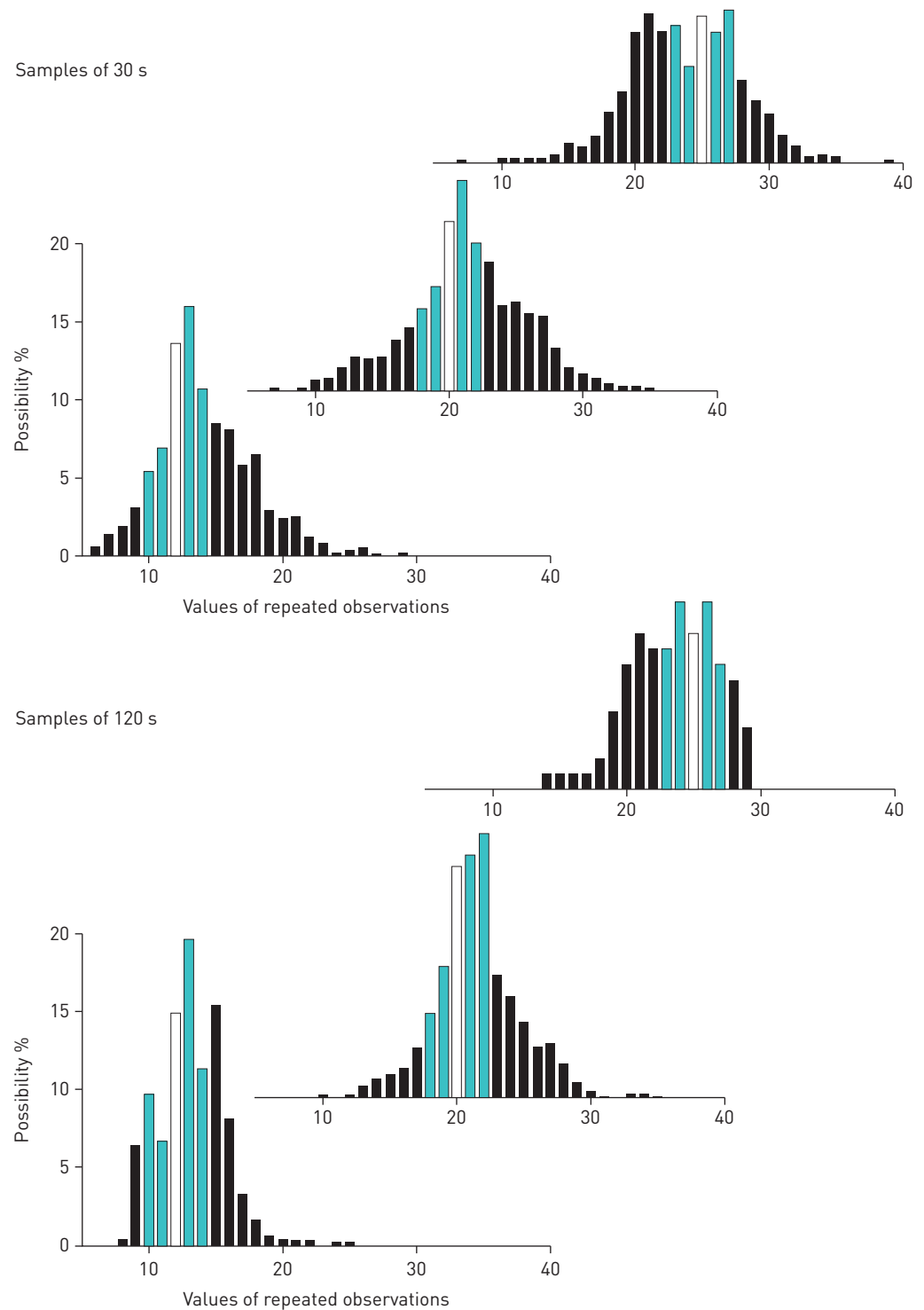
FIGURE 3 Likelihood of repeat observations based on the entire observed population. The upper three distributions are based on 30 s samples, the lower three on 120 s samples. The histograms show the distribution of values that might be found if an initial observation of 12, 20 or 25 breaths·min^{-1} had been made, from any subject in the entire population, and this observation had been repeated. The open column indicates the chance that the first observation would be found again in a subsequent sample, and the blue columns the chances of observing values within 2 breaths min^{-1} of the initial observation. The variation of repeat measures is reduced when 120 s samples are used.

conditions may differ. For example, rapid shallow breathing in pneumonia, where respiratory rate may exceed 40 [9], may limit the potential for variation. Conversely, bradypnoea caused by opioids increases variation in breath duration [10]. The breathing records we used are likely to be typical of acutely ill
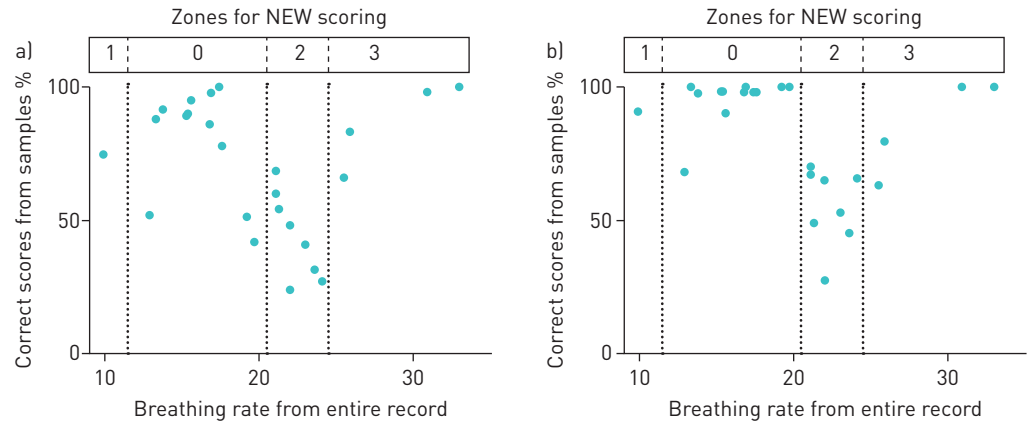
FIGURE 4 The proportion of measurements from each patient that would give a respiratory rate score the same as the score gained by the overall respiratory rate (the most precise measure) for that patient. The vertical dotted lines are the score cut-off values and the numbers above are the score values that are allocated for a rate in each zone. Panel a) uses measurements made over 30 s, Panel b) measurements made over 120 s. Scoring is much less consistent when the overall respiratory rate is between 20 and 25 breaths·min^{-1}. Consistency improves when samples of longer duration are used. NEW: UK National Early Warning Score.

patients with a variety of illnesses. The patients we studied were of a "sick but stable" group, in a well-staffed environment and not giving rise to immediate concern, and the median values we report are clinically plausible. Breath-to-breath variation appears to be greater in older subjects [8]

Our observations suggest that previous reports have drawn inappropriate conclusions. BIANCHI et al. [11] compared 1 min observation periods using two methods of measurement, but their samples were not synchronous. They found a negligible overall difference, which would be expected on statistical grounds, but the 95% limits of agreement were wide, ±5 breaths·min^{-1}. BRETELER et al. [12] compared five commercial devices with similar results. Clearly, comparisons that use short observation periods should only be made using exactly the same time period for each measurement. If not, discrepancies between different devices or methods will be inferred, because small samples have been taken from a number of different breaths (figure 1c).

Most automatic devices process the derived signal to display a "smoothed" breathing rate using methods such as median filtering. Although the output value is a continuous "average", it may never represent a specific value from an identifiable time period. If respiratory rate is derived from other physiological signals, such as ECG or pulse plethysmography, the temporal relationship becomes even less direct.

Such variability in clinical observations has been overlooked by research into physiological scoring and warning systems, which do not consider imprecision of the input data. The first obvious problem is that the measure itself is imprecise. Since recognition of this problem, and attempts to reduce it, have been limited, we should consider the possibility that not only would greater precision increase trust in respiratory measurement for patient assessment, but also that better information might improve management. Consider assessing cardiac rhythm using an electrocardiogram compared with palpation of the pulse!

The second problem is that imprecise measures introduce bias in statistical relationships. Random variation introduces "noise", which can obscure the relationship between an input measure and an outcome such as admission to intensive care. Random variation blurs what might otherwise be a clear association [13]. For example, the present UK scoring system sets a steep relationship between respiratory rate and the score value for respiratory rates between 20 and 25 (figure 4). In contrast, a machine-based decision tree generated a more gradual fit [14]. With the decision tree, a rate of up to 18 scores 0, 19 to 20 scores 1, 21 to 25 scores 2, and 25 or more is scored at 3, thus fitting a gradual effect to the clinical respiratory rate measurements. Similarly, a logistic regression process can more readily classify poor outcome [15]. Simply put, converting a continuous measure (such as respiratory rate) into categories (such as a score) by using cut-off values is statistically inefficient [16].

Another statistical problem is that scanty and imprecise data from individuals, often a single score, are used for categorical predictions, such as "will be admitted to ITU, or die, within 24 h". Although the proportion of patients with an adverse outcome increases as the score increases, the precision of prediction for an individual patient may be limited. Problems in transforming individual categorical data into predictive tools are acknowledged in other biological analyses [17, 18] and are now becoming relevant in outcome

prediction [19, 20]. The phenomenon has been imaginatively described as the "giant rat" approach: data from many individuals can be combined to yield a quantitative description that may well be valid when applied to an entire population [14] but may not reliably predict an individual outcome [21].

Many factors affect clinical respiratory rate measurement and recording. The topic is poorly covered in medical textbooks [22] and hospital staff generally do it badly [2]. A retrospective study of 2 500 000 nursing records showed substantial preference for respiratory rates of 14, 16, 18 and 20 [3]. If doctors are trained to count respiratory rate over a 60 s time period, the rates are relatively evenly distributed, whereas values recorded routinely showed marked preference for 16, 18 and 20 [23]. Our findings suggest that a device to acquire breath counts over longer time periods would be needed.

Automatic devices to measure respiratory rate are available, but most are cumbersome, restrictive, not well tolerated, prone to interference and expensive, and are generally only used in specialised applications. The clinical assessment of such monitoring devices is limited: they are often only tested in healthy volunteers or stable patients with moderate illness, or in special circumstances such as during sedation or after anaesthesia. Few studies validate devices in appropriate populations, such as unwell general medical and surgical patients. A systematic review concluded that suitable devices for measuring respiratory rate in general medical wards, *i.e.* more general surveillance monitoring, are not yet available [24]. Another review compared studies of continuous monitoring with those that used intermittent charting [25]. Although nine studies were classified as "continuous", only four used a continuous process to assess respiratory rate, and these appeared no better in preventing adverse events.

A future approach would be to use a device that allowed staff to make reliable measurements of respiratory rate over a suitable length of time [26–29]. Small wearable devices that directly measure respiratory movements and allow precise counts to be made over longer periods could improve the precision of rate measures [30], although such devices require adequate clinical validation. Precision will increase in proportion to the square root of the sample duration, so changing from a 30 s to a 15 min observation could improve precision about five times. At present, we cannot predict the impact that improved precision could have on assessment and early detection of deterioration, but it is this part of the "afferent limb" of the control loop (illness recognition and treatment response) that requires improvement [31].

Limited consideration of measurement error is commonplace in modern medical research [4], as "big data" uses material not originally intended for research. However, despite its limited precision, respiratory rate still contributes as much predictive information to illness scores as do other measures [32]. With more precise measurements, made more frequently, value trends might contribute even more information to outcome prediction [32]. Understanding the reason for imprecision in our current procedures should stimulate developing better methods, and suitable sensors, to count and record appropriate samples of breathing.

References

1   Smith MEB, Chiovaro JC, Neil MO, *et al.* early warning system scores for clinical deterioration in hospitalized patients: a systematic review. *Ann Am Thorac Soc* 2014; 11: 1454–1465.
2   Badawy J, Nguyen OK, Clark C, *et al.* Is everyone really breathing 20 times a minute? Assessing epidemiology and variation in recorded respiratory rate in hospitalised adults. *BMJ Qual Saf* 2017; 26: 832–836.
3   Pedersen NE, Rasmussen LS, Petersen JA, *et al.* A critical assessment of early warning score records in 168,000 patients. *J Clin Monit Comput* 2018; 32: 109–116.
4   Brakenhoff TB, Mitroiu M, Keogh RH, *et al.* Measurement error is often neglected in medical literature: a systematic review. *J Clin Epidemiol* 2018; 98: 89–97.
5   Edmonds Z V, Mower WR, Lovato LM, *et al.* The reliability of vital sign measurements. *Ann Emerg Med* 2002; 39: 233–237.
6   Thurnheer R, Xie X, Bloch KE. Accuracy of nasal cannula pressure recordings for assessment of ventilation during sleep. *Am J Respir Crit Care Med* 2001; 164: 1914–1919.
7   Lim WS, Carty SM, Macfarlane JT, *et al.* Respiratory rate measurement in adults: how reliable is it? *Respir Med* 2002; 96: 31–33.
8   Tobin MJ, Mador MJ, Guenther SM, *et al.* Variability of resting respiratory drive and timing in healthy-subjects. *J Appl Physiol* 1988; 65: 309–317.

9   Binger CAL, Davis JS Jr. The relation of anoxaemia to the type of breathing in pneumonia: a study of respiration by means of a body plethysmograph. *J Clin Invest* 1928; 6: 171–185.

10 Drummond GB, Bates A, Mann J, *et al.* Characterization of breathing patterns during patient-controlled opioid analgesia. *Br J Anaesth* 2013; 111: 971–978.

11 Bianchi W, Dugas AF, Hsieh YH, *et al.* Revitalizing a vital sign: improving detection of tachypnea at primary triage. *Ann Emerg Med* 2013; 61: 37–43.

12 Breteler MJM, KleinJan EJ, Dohmen DAJ, *et al.* Vital signs monitoring with wearable sensors in high-risk surgical patients: a clinical validation study. *Anesthesiology* 2020; 132: 424–439.

13 Hutcheon JA, Chiolero A, Hanley JA. Random measurement error and regression dilution bias. *BMJ* 2010; 340: 1402–1406.

14 Badriyah T, Briggs JS, Meredith P, *et al.* Decision-tree early warning score (DTEWS) validates the design of the National Early Warning Score (NEWS). *Resuscitation* 2014; 85: 418–423.

15 Ghosh E, Eshelman L, Yang L, *et al.* Early Deterioration Indicator: data-driven approach to detecting deterioration in general ward. *Resuscitation* 2018; 122: 99–105.

16 Selvin S. Two issues concerning the analysis of grouped data. *Eur J Epidemiol* 1987; 3: 284–287.

17 Waud DR. On biological assays involving quantal responses. *J Pharmacol Exp Ther* 1972; 183: 577–607.

18 Lu W, Ramsay JG, Bailey JM. Reliability of pharmacodynamic analysis by logistic regression: mixed-effects modeling. *Anesthesiology* 2003; 99: 1255–1262.

19 Steyerberg EW, Vickers AJ, Cook NR, *et al.* Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology* 2010; 21: 128–138.

20 Steyerberg EW, Uno H, Ioannidis JPA, *et al.* Poor performance of clinical prediction models: the harm of commonly applied methods. *J Clin Epidemiol* 2018; 98: 133–143.

21 Sani O, Shafer SL. MAC attack? *Anesthesiology* 2003; 99: 1249–1250.

22 Cook CJ, Smith GB. Do textbooks of clinical examination contain information regarding the assessment of critically ill patients? *Resuscitation* 2004; 60: 129–136.

23 Semler MW, Stover DG, Copland AP, *et al.* Flash mob research: a single-day, multicenter, resident-directed study of respiratory rate. *Chest* 2013; 143: 1740–1744.

24 Cardona-Morrell M, Prgomet M, Turner RM, *et al.* Effectiveness of continuous or intermittent vital signs monitoring in preventing adverse events on general wards: a systematic review and meta-analysis. *Int J Clin Pract* 2016; 70: 806–824.

25 Buist M, Bernard S, Nguyen TV, *et al.* Association between clinically abnormal observations and subsequent in-hospital mortality: a prospective study. *Resuscitation* 2004; 62: 137–141.

26 McGaughey J, Alderdice F, Fowler R, *et al.* Outreach and Early Warning Systems (EWS) for the prevention of intensive care admission and death of critically ill adult patients on general hospital wards. *Cochrane Database Syst Rev* 2007; 3: CD005529.

27 Odell M, Victor C, Oliver D. Nurses' role in detecting deterioration in ward patients: systematic literature review. *J Adv Nurs* 2009; 65: 1992–2006.

28 Tirkkonen J, Ylä-Mattila J, Olkkola KT, *et al.* Factors associated with delayed activation of medical emergency team and excess mortality: an Utstein-style analysis. *Resuscitation* 2013; 84: 173–178.

29 Joshi M, Ashrafian H, Aufegger L, *et al.* Wearable sensors to improve detection of patient deterioration. *Expert Rev Med Devices* 2019; 16: 145–154.

30 Kellett J, Sebat F. Make vital signs great again: a call for action. *Eur J Intern Med* 2017; 45: 13–19.

31 Churpek MM, Adhikari R, Edelson DP. The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation* 2016; 102: 1–5.

32 Johan IJ, Puntervoll LH, Pedersen PB, *et al.* The value of vital sign trends in predicting and monitoring clinical deterioration: a systematic review. *PLoS One* 2019; 14: 1–13.