Outlier detection of vital sign trajectories from COVID-19 patients

Sara Summerton[1], Ann Tivey[2], Rohan Shotton[2], Gavin Brown[1], Oliver C Redfern[3], Rachel Oakley[2], John Radford[2,4], and David C. Wong[1,5]

***Abstract*— There is growing interest in continuous wearable vital sign sensors for monitoring patients remotely at home. These monitors are usually coupled to an alerting system, which is triggered when vital sign measurements fall outside a predefined normal range. Trends in vital signs, such as an increasing heart rate, are often indicative of deteriorating health, but are rarely incorporated into alerting systems. In this work, we present a novel outlier detection algorithm to identify abnormal vital sign trends.**

We introduce a distance-based metric to compare vital sign trajectories. For each patient in our dataset, we split vital sign time series into 180 minute, non-overlapping epochs. We then calculated a distance between all pairs of epochs using the dynamic time warp distance. Each epoch was characterised by its mean pairwise distance (average link distance) to all other epochs, with large distances considered as outliers.

We applied this method to a pilot dataset collected over 1561 patient-hours from 8 patients who had recently been discharged from hospital after contracting COVID-19. We show that outlier epochs correspond well with patients who were subsequently readmitted to hospital. We also show, descriptively, how epochs transition from normal to abnormal for one such patient.

I. INTRODUCTION

Monitoring vital signs is commonly used in clinical practice to assess a patient's condition. Abnormal vital signs, those that are outside of a normal range relative to a general population, often precede adverse events. For instance, in hospitalised patients, abnormal respiration and cardiac function commonly occurs in the 24 hours before sudden cardiac arrest [1]. Abnormal vital signs are also known to precede deterioration in patients with COVID-19. In particular, there is strong correlation between low oxygen saturations (SpO2) and severe cases of COVID-19 requiring hospitalisation [2].

Traditionally, detection of abnormal vital signs is assessed using Early Warning Scores (EWS) calculated from intermittent, manually-collected measurements [3]. Non-continuous data collection means that deterioration may go unnoticed between measurements. Improvements in continuous, wearable vital sign monitors aim to address this issue [4], and there is some supporting evidence for their feasibility in both hospital and community settings [5], [6].

Early warning scores are one way of combining information from multiple vital signs into a single score to trigger appropriate clinical intervention. Higher scores identify patients whose vital signs deviate most from normal.

One key limitation of early warning scores is that they are typically calculated from only the most recent set of vital signs [7]. This has traditionally been the case even for continuous monitoring devices [8]. It is possible that earlier detection of deterioration can occur when trends in vital signs are also considered. Previous studies of hospitalised patients have attempted to include information about vital sign trends. However, these have used relatively simplistic summaries, such as difference between current and baseline values [9], [10], and not been used for out-of-hospital monitoring.

In this exploratory analysis, we assessed whether vital sign trends, regardless of the absolute value, are potentially informative for detecting COVID-19 deterioration.

II. METHODS

We propose an approach to outlier detection for vital sign trends, and consider its performance on continuous vital sign data collected from COVID-19 patients that had been discharged from hospital to their own home.

In the following sections we first introduce our overall approach for identifying outlier trends. We then describe the vital sign data in more detail. Finally, we apply the our method to the data and comment on its performance.

A. Outlier Detection

We implement an outlier detection approach based on distance to nearest neighbours. This family of approaches is described in detail by Pimentel et al. [11]. Consider a sequence of data $A = [a_1, a_2, ..., a_m]$, in which $a_i \in R^n$ is an n-dimensional feature vector. We denote the similarity between two such sequences, A and B, as some function $F(A, B)$. One way to define similarity is as a distance. One possible distance is the Dynamic Time Warp (DTW) distance:

$$F(A,B) = \sqrt{\sum_{(i,j)\in\pi} ||a_i - b_j||^2}$$

with $1 \leq i, j \leq m$, and where π is the optimal alignment path, defined as the contiguous path through the matrix of squared element-wise differences between both sequences

[1]S. Summerton, G. Brown and D.C. Wong are with the Department of Computer Science, University of Manchester, Kilburn Building, Oxford Road, Manchester, UK. sara.summerton@manchester.ac.uk, david.wong@manchester.ac.uk

[2]A. Tivey, R. Shotton, R. Oakley, and J. Radford are with the Christie NHS Foundation Trust, Manchester Academic Health Science Centre, Manchester, UK

[3]O. Redfern is with the Nuffield Department of Clinical Neurosciences, University of Oxford, Kadoorie Centre, John Radcliffe Hospital, Oxford, UK.

[4]J. Radford is also with the Institute of Cancer Sciences, University of Manchester, Manchester, UK.

[4]D. Wong is also with the Centre for Health Informatics, University of Manchester, Vaughan House, Portsmouth Street, Manchester, UK.

that minimizes the cumulative distance between them [12]. The DTW distance can accommodate time series of unequal length, however all sequences in this work were fixed to the same length.

The advantage of DTW over the euclidean distance is that it allows nonlinear alignments, so similar but non-aligned or out of phase sequences can be meaningfully compared. Of course, other distance measures are possible and their suitability depends on the specific context.

A hierarchical clustering approach can be taken to assess the similarity of a fixed length multivariate time series in the context of a set of multiple time series. Agglomerative clustering will calculate the distance between each time series, then join the pair of time series with the shortest distance into a single cluster in an iterative process until the entire dataset is contained in a single cluster. The distance between clusters is defined as

$$D(U,V) = \frac{1}{(|U| \cdot |V|)} \sum_{u \in U} \sum_{u \in V} F(u,v)$$

where u and v are elements and $|U|$ and $|V|$ are the cardinalities of clusters U and V, respectively, and D is the average-linkage distance. U and V can be a cluster of multiple time series or a single sequence.

Many real-world clinical problems involve detecting abnormal physiological signals in an abundance of normal data. Patients with stable vital signs will comprise the majority of the time series segments, and we expect these series to have low average-linkage distances and thus be clustered together first. The single time series most dissimilar to the rest of the data will have the largest average-linkage distance, and it is these final clusters that may implicate outliers in the data corresponding to abnormal physiological signals.

B. Data and pre-processing

Data were collected as part of the RECAP pilot study. The study is listed on the ISRCTN registry with study ID ISRCTN16601772 (http://www.isrctn.com/ISRCTN16601772). All study participants provided signed written consent.

Data were collected from cancer patients who had contracted COVID-19, and had been considered suitable for outpatient care. All participants wore Isansys™ sensors which recorded their heart rate (HR), respiratory rate (RR) and temperature (Temp) each minute for up to three weeks.

Oxygen Saturations (SpO2) was measured via a standard finger probe and recorded twice daily. Heart rate (PPG-HR) was also derived and recorded from the underlying photoplethysmogram (PPG) signal. Only SpO2 data were used as an intervention to guide clinical care. In total, data were recorded from eight patients.

The time series for each patient were initially processed as follows. First, each channel was was normalized on a per-patient basis to zero mean and unit variance, then low-pass filtered using a 25-point (i.e. 25 minute) median filter to remove short-term fluctuations in heart and respiration rate, likely caused by movement artefacts and sensor noise.

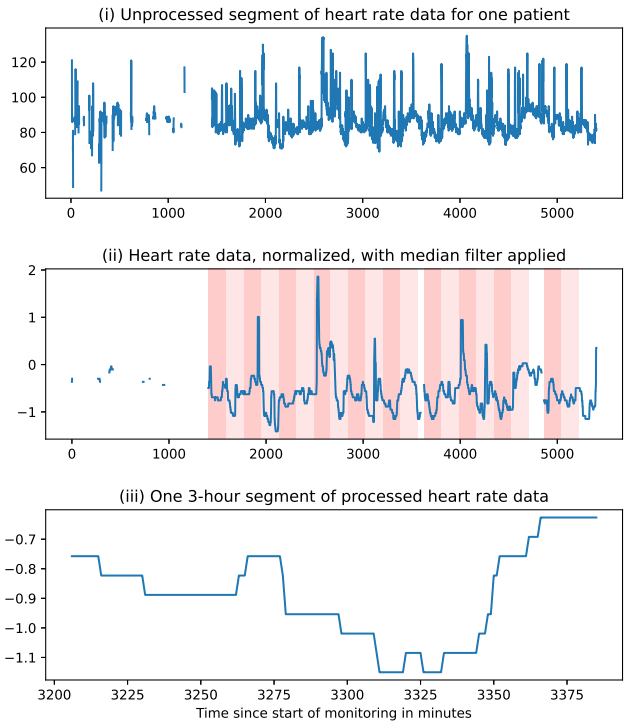

Fig. 1. (i) A subset of the original heart rate data for one patient. (ii) The data was normalized and a 25-minute median filter applied, then segmented into epochs (shaded in red). Data segments that were shorter than the epoch length were discarded. (iii) One 180-minute epoch of smoothed, normalized heart rate data.

Second, the signals were segmented into 180-minute epochs. This process is depicted in Figure 1. The epoch length was chosen so as not to capture medium-term variations in vital sign data, such as transitory increases in heart rate due to short-term physical activity, but still encapsulate overall changes in physiological condition.

The outlier detection approach was then applied to all epochs, ranking each epoch by hierarchical average linkage distance using the DTW distance measure. The 2% of epochs with the greatest average linkage were considered outliers.

To visualise similarity between epochs, we used multidimensional scaling (MDS), which is a dimensionality reduction approach that seeks to preserve the distance between data in the original high-dimensional space, in this case, the matrix of DTW distances between epochs [13]. We used this to describe and examine the sequence of contiguous epochs for representative patients from the dataset.

All data processing was undertaken in Python using the *scipy*, *sklearn*, and *tslearn* libraries. Code supporting this article is available at https://github.com/sara-es/outlier-detection-RECAP-data.

III. RESULTS

Table I shows the duration of vital sign data recorded for each patient in the data set, as well as whether the patient was readmitted to hospital. In total, there were 1561 patient-

TABLE I

PATIENT-LEVEL OVERVIEW OF AMOUNT OF VITAL SIGN DATA RECORDED (HOURS) AND CLINICAL EVENTS (HOSPITAL READMISSION)

ID	HR	RR	Temp.	SpO2	PPG-HR	Events
1	146	146	154	10	10	Hospital
2	2	2	2	0	0	None
3	2	2	17	0	0	None
4	300	300	107	1.7	1.7	None
5	28	28	41	1	1	Hospital
6	280	280	305	1	1	None
7	369	369	402	7	7	None
8	156	156	257	1	1	None

hours of data for the 8 patients; the mean length of data recording was 230 hours (range: 2.3 to 527 hours)

We excluded SpO2 and PPG-HR from the analysis, as these variables were sampled infrequently compared to HR, RR and Temp. We checked the data quality of remaining three variables by plotting their distributions. Based on this, we further excluded temperature, as the data contained a high proportion of physiologically implausible values (20.8% lower than 34 C). Poor data quality from skin temperature sensors in wearable devices is a known issue [14].

After segmenting data from each patient into epochs, we applied our outlier detection approach, using the DTW distance to hierarchically cluster all epochs. Those with the greatest average linkage distance are considered outliers, and are labelled with patient ID in the dendogram in Figure 2. Of the seven outliers, two belong to patient 5, and one belongs to patient 1 - the two patients that were readmitted to hospital.

Figure 3 shows the MDS map of time series epochs from all patients. The sequence of contiguous epochs for Patient 1 has been highlighted in red. Patient 1's initial epoch lies towards the centre of the MDS map, indicating that

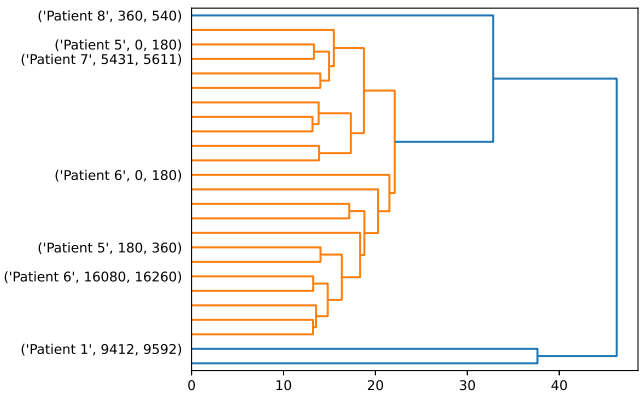

Fig. 2. A truncated dendrogram of vital sign time series epochs using agglomerative hierarchical clustering with average-linkage on the DTW distance matrix. The epochs to be clustered last, that is, those considered the greatest outliers, are labelled with the patient ID and (start, end) indices of the epoch. Notably, the final epoch belonging to patient 1 is the last to be clustered; two epochs from patient 5 are also visible, despite the relatively short duration of recorded data (28 hours, or 9 epochs) for that patient. Both patients 1 and 5 were readmitted to hospital.

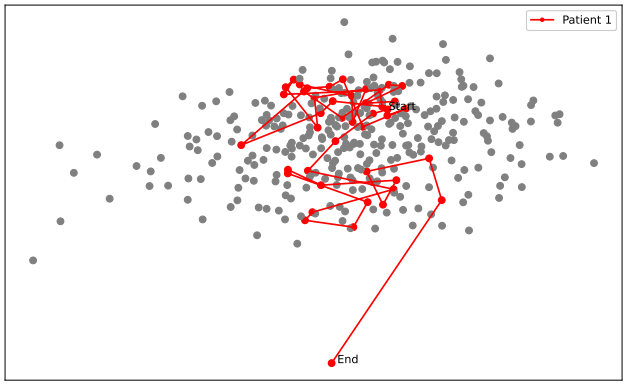

Fig. 3. Distances between epochs are visualised in 2D via a MDS map, on which an individual point represents a 180-minute segment of heart rate and respiratory rate data. Epochs from patient 1, who was readmitted to hospital, have been highlighted in red and connected in consecutive order. The initial 180-minute epoch is denoted by the text 'start', and the final epoch is denoted by the text 'end'. The trajectory can be shown to end outside the central cluster of time series epochs.

it is similar to multiple other epochs. Towards the end of the monitoring period, the epochs progress away from the starting location on the map. The final epoch is far away from all other points on the MDS map, indicating a highly unusual trajectory.

The raw time series epochs corresponding to the start and end points of patient 1, as well as one intermediate epoch, are shown in Figure 4. We observe that the 'start' epoch contains HR and RR trajectories that are both relatively flat. In contrast, the 'end' epoch contains vital signs that have deviated from their baseline average, and trajectories for both increase across the epoch. This trajectory is visually very different to the start and intermediate epochs in the figure, confirming the validity of the outlier detection approach.

IV. DISCUSSION

We developed a novel method to identify abnormal multivariate vital sign time series. Unlike previous methods, which use categorical variables or change scores to summarise a trend, our method considers the entire shape of a time series epoch via the DTW distance. By clustering based on this distance, we can determine outlying, unusual epochs.

In our small patient cohort, this approach yielded promising initial results. Of the 2% of most outlying epochs, 4/7 belonged to patients who went on to require hospital admission. Furthermore, our per-patient visualisation showed how epochs became progressively more abnormal for a patient who required readmission to hospital. These results therefore provide descriptive early evidence suggesting our approach for assessing vital sign trends may be useful for predicting COVID-19 deterioration.

While our approach to detect abnormal vital sign trajectories shows promise, there are several limitations. First, we chose epoch lengths of 180 minutes, based on clinical judgement. However, there is no guarantee that this epoch length is optimum. Second, we used DTW distances to

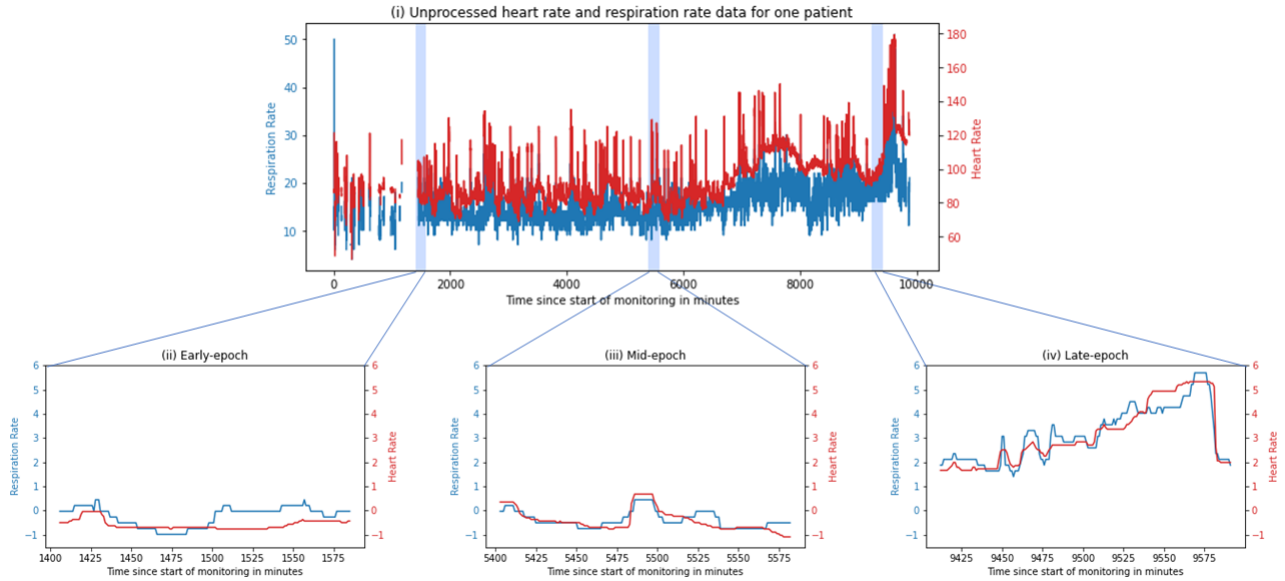

Fig. 4. (i) The raw HR and RR data for patient 1 over the entire duration of monitoring. (ii) The *start* epoch of normalized and smoothed HR and RR data for patient 1. (iii) An epoch taken from the approximate mid point of monitoring data. (iv) The *end* epoch for patient 1, showing visible deviation from previous baseline measurements.

compare epochs, when other distance measures may be more appropriate.Both the epoch length and distance measure can be optimised via cross-validation. The current data set was insufficient to attempt this, as data were collected from 8 patients and only 2 were readmitted to hospital (positive events).

Furthermore, we arbitrarily set a threshold of 2% of segments to be outliers. While this is suitable as a proof of principle, there may be more principled approaches to determine thresholds that indicate abnormal trajectories (see for instance, Clifton et al. [15]).

In conclusion, this work shows how vital sign trajectories may contain clinically relevant information, predictive of patient deterioration. Future work should apply our method to larger data sets with more positive clinical events. Future iterations of early warning scores, coupled to continuous wearable monitors, could benefit from including temporal trends alongside the most recent vital sign measurements.

ACKNOWLEDGMENT

We are grateful for technical support provided by Isansys. We are grateful for the support of the Christie digital services department in facilitating this study. The authors have no conflicts of interest to declare.

REFERENCES

[1] R. M. Schein, N. Hazday, M. Pena, B. H. Ruben, and C. L. Sprung, "Clinical antecedents to in-hospital cardiopulmonary arrest," *Chest*, vol. 98, no. 6, pp. 1388–1392, 1990.

[2] T. Greenhalgh, M. Knight, M. Inda-Kim, N. J. Fulop, J. Leach, and C. Vindrola-Padros, "Remote management of covid-19 using home pulse oximetry and virtual ward support," *bmj*, vol. 372, 2021.

[3] C. Downey, W. Tahir, R. Randell, J. Brown, and D. Jayne, "Strengths and limitations of early warning scores: a systematic review and narrative synthesis," *International Journal of Nursing Studies*, vol. 76, pp. 106–119, 2017.

[4] D. Dias and J. Paulo Silva Cunha, "Wearable health devices—vital sign monitoring, systems and technologies," *Sensors*, vol. 18, no. 8, p. 2414, 2018.

[5] C. Downey, R. Randell, J. Brown, and D. G. Jayne, "Continuous versus intermittent vital signs monitoring using a wearable, wireless patch in patients admitted to surgical wards: pilot cluster randomized controlled trial," *Journal of medical Internet research*, vol. 20, no. 12, p. e10802, 2018.

[6] F. M. Iqbal, M. Joshi, G. Davies, S. Khan, H. Ashrafian, and A. Darzi, "The pilot, proof of concept remote-covid trial: remote monitoring use in suspected cases of covid-19 (sars-cov 2)," *BMC public health*, vol. 21, no. 1, pp. 1–8, 2021.

[7] S. Gerry, T. Bonnici, J. Birks, S. Kirtley, P. S. Virdee, P. J. Watkinson, and G. S. Collins, "Early warning scores for detecting deterioration in adult hospital patients: systematic review and critical appraisal of methodology," *bmj*, vol. 369, 2020.

[8] L. Tarassenko, A. Hann, and D. Young, "Integrated monitoring and analysis for early warning of patient deterioration," *BJA: British Journal of Anaesthesia*, vol. 97, no. 1, pp. 64–68, 2006.

[9] S. Charbonnier and S. Gentil, "On-line adaptive trend extraction of multiple physiological signals for alarm filtering in intensive care units," *International Journal of Adaptive Control and Signal Processing*, vol. 24, no. 5, pp. 382–408, 2010.

[10] D. Bell, J. Baker, C. Williams, and L. Bassin, "A trend-based early warning score can be implemented in a hospital electronic medical record to effectively predict inpatient deterioration," *Critical care medicine*, vol. 49, no. 10, p. e961, 2021.

[11] M. A. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[12] C. A. Ratanamahatana and E. Keogh, "Making time-series classification more accurate using learned constraints," *SIAM Proceedings Series*, pp. 11–22, 2004.

[13] J. B. Kruskal, *Multidimensional scaling*. Sage, 1978, no. 11.

[14] C. Downey, S. Ng, D. Jayne, and D. Wong, "Reliability of a wearable wireless patch for continuous remote monitoring of vital signs in patients recovering from major surgery: a clinical validation study from the tracing trial," *BMJ open*, vol. 9, no. 8, p. e031150, 2019.

[15] D. A. Clifton, S. Hugueny, and L. Tarassenko, "Novelty detection with multivariate extreme value statistics," *Journal of signal processing systems*, vol. 65, no. 3, pp. 371–389, 2011.