Machine Learning for the Deterioration of Patients on Hospital Wards

Farah Shamout
Balliol College
University of Oxford

Supervised by
Dr. David Clifton
Dr. Tingting Zhu

A transfer report submitted to the University of Oxford in accordance with the requirements of the degree of Doctor of Philosophy in the Department of Engineering Science.

August 24, 2017
Part 1: Scientific Paper
Age- and Sex-Based Early Warning Score

Farah Shamout1,2
Supervised by Dr. David Clifton1 and Dr. Tingting Zhu1

Abstract—Early Warning Score Systems assign scores to a patient’s vital signs to detect adverse events in order to secure timely medical response. However, most systems do not account for variations in vital signs due to aging changes or sex differences, which affects the accuracy of current systems for different deteriorating patient sub-groups. We aimed to develop an Age- and Sex-Based Early Warning Score derived from statistical distributions of vital signs across ages and sexes. The data is obtained from four hospitals in Oxford, United Kingdom, which contains a total of 48,433 patient admissions to the wards with documented sex and age, including 3,182 patient admissions with at least one adverse event. Vital sign distributions were compared at each year of age (age, 16 yr to 99 yr) as well as sexes (females and males) using aggregated distributions. The centiles of the vital sign cumulative distribution functions. The centiles of the vital sign distributions modeled the normal ranges of the Age- and Sex-based Early Warning Score. An adverse event was defined as the occurrence of in-hospital mortality, cardiac arrest or ICU admission within the first 24 hours of stay. The area under the receiver operating curve for the Age- and Sex-Based Early Warning Score was higher than that of existing systems, namely Centile-based Early Warning Score (0.828 vs 0.730; \( p < 0.001 \)), the National Early Warning Score (0.828 vs 0.808; \( p < 0.001 \)) and the Modified Centile-based Early Warning Score (0.828 vs. 0.825; \( p < 0.001 \)). The Age- and Sex-Based Early Warning Score also outperformed the current systems over the non-elderly patients (age, 16 yr to 40 yr), and amongst both sexes, females and males. Our results suggest that accounting for age- and sex-related vital sign changes more can accurately detect deterioration of non-elderly patients prior to an adverse event than current methods. Further investigations of age- and sex-specific methods are necessary to decrease preventable deaths.

Keywords: age, sex, early warning scores, adverse events, vital signs, statistical models

I. INTRODUCTION

Despite efforts to improve patient care, in-hospital patients may still suffer from unexpected adverse events such as ICU admission and cardiac arrest, which at times result in death. Recent studies have identified possible causal factors of preventable deaths, which include late detection of serious vital sign abnormalities and lack of timely response to identifying deteriorating patients [1], [2].

Preventable deaths may be avoided using technologies that detect deterioration in a timely manner, by documenting vital signs in electronic health records and processing the data using software algorithms with improved sensitivity and specificity [3]. Digital data collection developments include the System for Electronic Notification and Documentation (SEND), which is designed to collect vital sign observations electronically and is currently being deployed in clinical settings in the Oxford University Hospitals, UK [4].

II. RELATED WORK

A. Early Warning Score Systems

Captured vital sign data in clinical settings are commonly fed to various Early Warning Score (EWS) systems, also known as ‘track-and-trigger’ (T&T) systems. In an EWS system, each vital sign is assigned a score according to the predetermined normality ranges, and alerts are generated according to the patients overall score. The EWS systems are deployed to secure timely and effective responses for patients with deteriorating medical conditions, rather than to predict a particular outcome [5].

To address the multiplicity of EWS systems used in different hospitals across the UK and standardize detection of patients deterioration, the Royal College of Physicians (RCP) recently published a proposal for the National Early Warning Score (NEWS) as the standard EWS to be widely adopted across the National Health Service (NHS) hospitals [5]. Based on the VitalPAC Early Warning Score (ViEWS) system and clinical expertise, the NEWS system assigns scores to six physiological parameters: respiratory rate, oxygen saturations, temperature, systolic blood pressure, pulse rate, and level of consciousness. The aggregate score of those parameters, in addition to the score assigned to patients requiring supplemental oxygen, identify those patients at risk of clinical deterioration [5], [6].

The Centile-based Early Warning Score (CEWS) system is based on a statistical approach to determine the normal ranges of only four vital signs, including respiratory rate, heart rate, systolic blood pressure, and oxygen saturation [7]. It was later modified, namely the Modified Early Warning Score (MCEWS) system, to account for differences between manually-collected and continuously-acquired respiratory rate. MCEWS also takes into account body temperature, whether a patient is being supplemented oxygen, and level of consciousness [8]. While CEWS and MCEWS are developed for adult patients (such as those with age > 16 yr), the RCP also recommends to use NEWS to clinically assess all adult patients of both sexes, females and males.

B. Age and Sex Physiological Differences

Individual variations of a patients physiological responses are a qualitative and quantitative function of age, sex, body composition and other individual characteristics. While the mechanisms are not fully clarified, recent studies have investigated the pathological and physiological changes that

\[1\] Institute of Biomedical Engineering, Department of Engineering Science, University of Oxford, Oxford, UK
\[2\] Corresponding author: farah.shamout@eng.ox.ac.uk
occur with age and sex differences that may alter vital signs, specifically for heart rate, blood pressure, temperature, and respiratory rate.

Resting heart rate is often observed to increase with age due to deconditioning and autonomic dysregulation [9], [10], [11]. However, it has also been shown that in the seated upright position at rest, during routine activities of daily living and during exertion, heart rate decreases with aging due to a decrease in the efficacy of beta-receptors modulation of both the heart and vasculature [9], [10].

Increased vascular loading of the heart and thickening of the left ventricular also reflects in a modest increase in systolic blood pressure at rest [10]. While systolic blood pressure shows a gradual increase between the ages of 30 and 84 years, diastolic blood pressure increases until the fifth decade and then slowly decreases to at least 84 years of age [12]. Females have a higher incidence of hypertension with aging due to menopause, since estrogen deficiency further induces endothelial dysfunction [13]. Pre-menopausal females however have lower blood pressure than age-matched males due to less stiff arteries [14], [15].

Decline in elastic recoil and chest wall compliance with aging leads to a decrease in tidal volume and an increased work of breathing, reflecting on older adults having increased respiratory rate [16]. Respiratory muscle strength decreases much more in men than in women with aging [17].

Aging also alters thermoregulatory responses and immune systems, leading to lower core body temperature among older adults. Rise in body temperature is an important reflection of changes in the immune system to combat microbial infections, and thus early detection of subtle temperature variations in older patient according to individual reference ranges can avoid life-threatening outcomes [9]. As for oxygen saturation, there is very limited literature but one study have found no significant differences amongst different age groups [18].

In previous studies, it has been shown that age has a significant impact on in-hospital mortality and the best performing EWS incorporate age as one of its components [19], [12]. Yet, most EWS systems do not include age as it increases complexity, conveys little benefit and may raise ethical concerns [6], [20]. The various EWS systems fall into three broad categories: (a) objective EWS systems applicable for all patients, regardless of age and sex [6], [7], (b) patients age and/or sex are considered as parameters of the EWS and assigned a score in a heuristic manner [6], [21], [22], [19], [20], [23], or (c) the EWS is designed only for a specific patient population [24].

C. Motivations

Having identified the complexity of physiology between sexes and increasing age, we hypothesize that the incorporation of those two major factors of individual variations (age and sex) in an early warning score may improve performance over the current systems used in clinical practices. We propose a new Age- and Sex-Based Early Warning Score that was developed through a statistical method using cumulative distribution functions.

III. METHODS

A. Dataset

This study used a large dataset comprising of vital sign monitoring from patients collected using the SEND system [4] which was provided by the Hospital Alerting Via Electronic Noticeboard or HAVEN project. The HAVEN project is a large database comprised of anonymous patient-data from adult patients, related to demographics, periodic vital sign measurements on hospital wards and adverse events information. The provided database (havenouh-1.0) includes patient data from four Oxford University Hospitals in the United Kingdom (UK): John Radcliffe Hospital, Churchill Hospital, Horton General Hospital and Nuffield Orthopaedic, between 2014 and 2016. The study protocol was approved by the Oxford Research Ethics Committee.

Independent admissions were considered as independent patient episodes, thereby treated as independent patients. Only patients with a length of stay greater than 24 hours were considered, where the performance of EWS systems were evaluated and compared. The minimum and maximum ages of adult patients considered in this study were 16 and 99, respectively.

For each patient, the dataset included continuous timestamped observation sets of vital signs. Each observation set included heart rate (HR), respiratory rate (RR), diastolic blood pressure (DBP), systolic blood pressure (SBP), peripheral capillary oxygen saturation (SpO₂), fraction of inspired oxygen (FiO₂), temperature (TEMP), Alert, voice, pain, unresponsiveness score (AVPU), and Glasgow Coma Scale score (GCS), as well as demographics information such as date of birth, sex, and date of death, if available. The adverse events considered in this paper are death within 24 hours of hospital discharge, in-hospital ICU admission and cardiac arrest.

The dataset was split into a training and test set through random stratification, with set sizes 80% and 20% of overall dataset, respectively. The stratification variables were age and the number of adverse events per age. The split was performed on the patient level, where each patient is only considered either in the training set or test set.

B. Development of Age- and Sex- Based Early Warning Score (ASEWS)

The development of the Age- and Sex- based Early Warning Score (ASEWS) was developed in a two-fold process. First, the variations of vital signs with age and sex were investigated using cumulative distribution function (cdf). The centiles of the cdfs were then used to produce the normal distribution functions.

The development of the Age- and Sex- based Early Warning Score (ASEWS) was developed in a two-fold process. First, the variations of vital signs with age and sex were investigated using cumulative distribution function (cdf). The centiles of the cdfs were then used to produce the normal distribution functions.

We have used an age-smoothing approach to allow patients to move through sliding ranges as they age, instead of abruptly moving between strictly divided age groups, such as age deciles. At each year of age, we included observation sets
of patients within lower and upper age bounds. The bounds are within a particular number of years from the year of age.

Let $a$ be the year of age, in the age range $[a - e, a + e]$, where $e$ is a user-defined number, then the observation sets considered for patients at year of age $a$ include observation sets of patients aged between $a - e$ and $a + e$, such that, $S_a = \text{Observation sets of patients-aged } [a \pm e]$.

For example, when $e = 10$, the observation set investigated for 30-year-olds includes patients between ages 20 and 40. This age-smoothing approach is performed for females, males and both sexes combined.

To investigate vital sign variations with age, cdfs were plotted for each vital sign (HR, RR, SBP, TEMP, & SPO$_2$) at each year of age using the above age-smoothing approach. The age range for our patient population is [16, 99], therefore for each vital sign there are 83 cdfs. Next, for each vital sign, the cdfs of all years of age were rotated 90$^\circ$ and visualized in a heatmap, such that the x-axis represents age, y-axis represents the vital sign values within physiologically relevant limits and the colour value references the cdf evaluated at each vital sign value. The 1st, 50th and 99th centiles are shown in a white dashed line. This approach allows us to visually observe how vitals may change with increasing age for each sex.

The cut-off values for the normal ranges of vital signs at each age per sex were then obtained using centiles of the cdfs, following the method previously designed for the Centile-based Early Warning Score (CEWS) [7]. For double-sided distributions (HR, RR, SBP, TEMP & TEMP), excluding diastolic blood pressure. The centiles of the vital signs cdfs modeled the normal ranges of MCEWS-

<table>
<thead>
<tr>
<th>Demographics</th>
<th>Overall Dataset</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients (%)</td>
<td>1454565</td>
<td>1162676</td>
<td>291889</td>
</tr>
<tr>
<td>Number of female patients (%)</td>
<td>24325 (50.2)</td>
<td>19393 (50.0)</td>
<td>4932 (51.0)</td>
</tr>
<tr>
<td>Mean Age (SD)</td>
<td>63.3 (19.4)</td>
<td>63.3 (19.4)</td>
<td>63.3 (19.4)</td>
</tr>
<tr>
<td>16–40 year olds (%)</td>
<td>7262 (15.0)</td>
<td>5811 (15.0)</td>
<td>1451 (15.0)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Prevalence of Adverse Events (Death, ICU and cardiac arrest)</th>
<th>Overall Dataset</th>
<th>Training Set</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of patients with at least one adverse event (%)</td>
<td>3182 (6.6)</td>
<td>2547 (6.6)</td>
<td>635 (6.6)</td>
</tr>
<tr>
<td>Females (%)</td>
<td>1396 (2.9)</td>
<td>1119 (2.9)</td>
<td>277 (2.9)</td>
</tr>
<tr>
<td>Males (%)</td>
<td>1786 (3.7)</td>
<td>1428 (3.7)</td>
<td>358 (3.7)</td>
</tr>
<tr>
<td>16–40 year olds (%)</td>
<td>291 (0.6)</td>
<td>231 (0.6)</td>
<td>60 (0.6)</td>
</tr>
<tr>
<td>&gt; 40 year olds (%)</td>
<td>2891 (6.0)</td>
<td>2316 (6.0)</td>
<td>575 (6.0)</td>
</tr>
</tbody>
</table>

C. Model Selection

Using a random k-fold cross-validation on the training set, the normal ranges of the ASEWS were obtained k times, where k = 10. To evaluate the performance of the k ASEWS ranges, the area under receiver-operating characteristics (AUROC) curve is evaluated at 24 hours prior to an adverse event, as defined above, for each fold. Each year of age was then assigned with the ASEWS ranges that were evaluated with the highest AUROC amongst the k folds, for that year of age. Therefore, the final selected overall ASEWS model comprised of ranges from different folds.

D. Retraining the Modified Centile-Based Early Warning Score

In order to compare the performance of ASEWS with MCEWS, we retrained MCEWS on our training set to obtain cdfs for each vital sign (HR, SBP, SPO$_2$, RR & TEMP), excluding diastolic blood pressure. The centiles of the vital signs cdfs modeled the normal ranges of MCEWS-H (Modified Centile-Based Early Warning Score trained on our HAVEN dataset).

E. Performance Assessment

To evaluate the performance of the EWS systems, the area under receiver-operating characteristics (AUROC) curve is evaluated at 24 hours prior to an adverse event. The AUROC curve is computed by plotting the true positive rate on the y-axis against the false positive rate on the x-axis, where the true positive rate is the proportion of positives that are correctly identified and the false positive rate is the proportion of negatives that are incorrectly identified as positives. The AUROC value ranges between 0.5 and 1.0, where the least value of 0.5 represents a randomly
predictive model, and value exceeding 0.80 indicates a good discriminatory model [6].

The selected ASEWS model was then applied to the test set where the AUROC was evaluated using a bootstrapping technique, where the number of bootstraps (n_b) is 100, to obtain the mean AUROC and its standard error. We then compared the overall performance of our model to the performance of existing models on the test set, also using bootstrapping, namely CEWS, NEWS and MCEWS-H. The mean AUROC and confidence intervals were also computed per year of age for all EWS systems through bootstrapping.

For pair-wise comparisons of performance of EWS systems, a t-test with known mean and unknown variance was performed to obtain p-values.

F. Statistical Methods

The statistical methods deployed in this study include normalised histograms for each vital sign, for females, males, and both sexes combined. The heatmaps of the vital signs were generated using cumulative distribution functions as described above.

If AVPU is missing but GCS is available, the AVPU score is imputed from the GCS score. If more than two vital signs (of HR, RR, SBP, TEMP, DBP & SPO₂) at any time point are absent, the observation set at that time point is eliminated entirely. If two or less vital signs at any time point are absent, then the vital signs carried the previous value forward. If the previous value was unavailable, then:

- For HR, RR, SBP, TEMP, DBP & SPO₂ are replaced with the vitals population median value.
- For AVPU, replaced by 1 if GCS is unavailable.
- And for whether supplemental oxygen was provided, it is replaced by 0.

Evaluation on test set was performed by bootstrapping without replacement (n_b = 100) to find the estimated standard error of the computed AUROC statistic. The bootstrapping technique was also used to compute 95% confidence intervals.

For pair-wise comparisons of performance of EWS systems, a t-test with known mean and unknown variance was performed to obtain p-values.

IV. RESULTS

A. Patient Cohort and Vitals Characteristics

The provided patient population included 98,919 patients. Extracting patients who were discharged from hospital in less than 24 hours from admission yielded 48,482 patients who were admitted between March 2014 and October 2016. Among those, 36 patients were not aged between 16 and 99 and were thus excluded. Finally, patients with more than 2 missing vital signs in the observations sets were also excluded and the final dataset size is 48,433 patients. The dataset was randomly divided into a training set and test set. The characteristics of the patient population is summarized in Table I including demographics and prevalence of

<table>
<thead>
<tr>
<th>TABLE II</th>
</tr>
</thead>
<tbody>
<tr>
<td>COMPARISONS OF MEAN AND STANDARD DEVIATION OF CONTINUOUS VITAL SIGNS AND DISTRIBUTION OF DISCRETE VARIABLES BETWEEN THE OVERALL DATASET, TRAINING SET AND TEST SET</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Mean (SD)</td>
</tr>
<tr>
<td>Heart rate, beats/min</td>
</tr>
<tr>
<td>Systolic blood pressure, mm Hg</td>
</tr>
<tr>
<td>Diastolic blood pressure, mm Hg</td>
</tr>
<tr>
<td>Temperature °C</td>
</tr>
<tr>
<td>Respiratory rate, breaths/min</td>
</tr>
<tr>
<td>Oxygen Saturation, %</td>
</tr>
<tr>
<td>Median (IQR)</td>
</tr>
<tr>
<td>Alert, Voice, Pain, Unresponsive</td>
</tr>
<tr>
<td>Supplemental Oxygen Provided</td>
</tr>
</tbody>
</table>

TABLE III

DIFFERENCES IN AREA UNDER THE RECEIVER OPERATOR CHARACTERISTICS CURVE (95% CONFIDENCE INTERVALS) BETWEEN THE AGE- AND SEX- BASED EARLY WARNING SCORE AND EXISTING SYSTEMS EVALUATED FOR FEMALES, MALES AND BOTH COMBINED, WITHIN 24 HOURS OF THE FIRST ADVERSE EVENT *

<table>
<thead>
<tr>
<th>All Patients</th>
<th>Female Patients</th>
<th>Male Patients</th>
</tr>
</thead>
<tbody>
<tr>
<td>(n= 9679)</td>
<td>(n= 4932)</td>
<td>(n= 4747)</td>
</tr>
<tr>
<td>System</td>
<td>Mean AUROC (95% CI)</td>
<td>p</td>
</tr>
<tr>
<td>CEWS</td>
<td>0.730 (0.717 - 0.746)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>MCEWS-H</td>
<td>0.825 (0.818 - 0.837)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>NEWS</td>
<td>0.808 (0.799 - 0.824)</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>ASEWS</td>
<td>0.828 (0.818 - 0.842)</td>
<td>-</td>
</tr>
</tbody>
</table>

* p value for comparison with the same patient cohort of ASEWS

The statistical methods deployed in this study include normalised histograms for each vital sign, for females, males, and both sexes combined. The heatmaps of the vital signs were generated using cumulative distribution functions as described above.

If AVPU is missing but GCS is available, the AVPU score is imputed from the GCS score. If more than two vital signs (of HR, RR, SBP, TEMP, DBP & SPO₂) at any time point are absent, the observation set at that time point is eliminated entirely. If two or less vital signs at any time point are absent, then the vital signs carried the previous value forward. If the previous value was unavailable, then:
Fig. 1. Heatmap of the aggregated cumulative distribution functions at each year of age ($a$ to $e$, where $e = 5$) for each sex, for each vital sign: (a) Heart Rate, (b) Systolic Blood Pressure, (c) Diastolic Blood Pressure, (d) Temperature, (e) Respiratory Rate, and (f) Oxygen Saturation. The 1st, 50th and 99th centiles are shown in a white dashed line.
adverse outcomes. It is observed that the mean ages and the proportions of females and males in the training and test set are approximately equal in both sets. The prevalence of an adverse outcome in the overall dataset, training set and test set amongst the patient subgroups by sex and age-group (below and above 40 year-olds) are equal.

The mean (SD) for each vital sign per sex and age group is computed in Table II. Using the overall dataset, the heatmaps of the cumulative distribution functions per year of age for each vital are plotted as shown in Figure 1. The observations at each year of age were grouped using $e = 5$. The white dashed lines indicate the 1st (bottom) and 99th (top) centiles for the consecutive cumulative distribution functions. The centiles for each age interval were obtained from their respective cumulative distribution functions.

B. Performance Evaluation

Evaluating the overall test set, the ROC curves are shown in Figure 2. The AUROC in Figure 2 suggests that our proposed method outperforms other EWS. Through the bootstrapping method, the mean AUROC and 95% confidence intervals are shown in Table III.

The mean AUROC and confidence intervals at each year of age are plotted in Figure 3. ASEWS performs better than existing EWS systems, noticeably among the non-elderly patients (age, 16 yr - 40 yr). The AUROC performance per sex is summarized in Table III, and the results suggest that ASEWS outperforms other EWS in the case when considering each individual sex patients, as well as those when both sexes were combined.

V. DISCUSSION

Our results show how vital signs vary with age (Figure 1). Heart rate (HR) is observed to decrease with increasing age similarly in both sexes, which may be physiologically explained by the decrease in the efficacy of the beta-receptors in modulating the heart and vasculature [10]. Yet, this physiological explanation only refers to measuring resting heart rate in the seated upright position, while other studies have shown that resting heart rate increases with age due to deconditioning and dysregulation [9].

Systolic blood pressure (SBP) is observed to rise with age, with a higher increase in older females than males. This is explained by the thickening of the left ventricular in both sexes with age, and the effects of menopause on estrogen deficiency in females [13]. Diastolic blood pressure gradually increases until the age of 50 and then begins to decrease, with a more steep decrease in males, which agrees with the results of the Framingham Heart Study [12].

There are no obvious changes in temperature (TEMP) with aging. It is important to note that physiological studies have investigated changes in core temperature with age, and it is difficult to obtain an estimate of core temperature using available instruments. Oxygen saturation (SPO$_2$) is observed to decrease with aging, which has an important implication on alerting for deviating oxygen saturation levels for younger patients.

The changes of vital signs with age emphasize the importance of considering age as an individual variation for female and male patients when designing the EWS. We provide a novel approach that allows patients to transition through normal ranges of the EWS smoothly as they age. The ASEWS performs better than existing EWS overall, especially for the non-elderly population and per sex comparisons. It is important to note that ASEWS only uses a small fraction of the overall data that were used to model MCEWS-H, and
yet performs better over the non-elderly patients.

Our study has several limitations. First, there is a low prevalence of adverse outcomes amongst the non-elderly patients. Our solution to limited data for certain ages is the grouping of patients to years of age and to randomly split and stratify the training and test set according to age and prevalence of adverse events. This can explain the similar fluctuations in the mean AUROC among the different EWS (Figure 3).

We also assume that the vital sign observation sets of each patient are independent and identically distributed random sets and that there is no correlation between vital signs, which may not be the case in reality.

We also assumed that $c = 5$ is an appropriate choice to group the years of age. The choice of this value heavily relies on the quantity of data available at each age and we find that this is the minimal value to smoothen the vital signs heatmaps.

It is also important to note other factors that may alter vital signs with aging. These include disease, its length of duration, use of medication and lifestyle. For example, in some individuals, blood pressure decreases with aging due to Alzheimers, dementia, cancer or impaired ventricular function [15]. Interventions through medication may also alter the underlying physiology of the patient. Data extraction of patients on certain medication or with certain diseases will be investigated further.

VI. CONCLUSION

Our study suggests that incorporating age and sex in the design of an early warning score system improves performance in comparison to current systems.

REFERENCES


Part 2: Literature Review & Research

Proposal
Contents

1 Literature Review 3

1.1 Introduction .......................................................... 3

1.2 Clinical Deterioration on Hospital Wards ....................... 4

1.2.1 Preventable Adverse Events ................................... 4

1.2.2 Antecedent Instability to Adverse Event ................... 6

1.2.3 Overall Patient-, Disease- and Organizational- Factors .... 6

1.3 Data for Machine Learning: Electronic Health Records ...... 7

1.3.1 What are Electronic Health Records? ......................... 7

1.3.2 Electronic Health Records-derived Phenotyping .......... 7

1.3.3 Challenges of Using Electronic Health Records .......... 8

1.3.4 Data Preparation for Machine Learning .................... 10

1.4 Machine Learning Approaches to Clinical Deterioration .... 11

1.4.1 Clinical Deterioration as a Classification Problem ....... 12

1.4.2 Clustering Analysis ............................................. 15

1.4.3 Gaussian Processes for Modeling & Forecasting Physiological Trajectory 16

2 Research Proposal 18

2.1 Thesis Contributions to the Field ................................ 18

2.1.1 Personalized Patient Time-Series Modeling ................ 18

2.1.2 Novelty Detection for Clinical Deterioration .............. 18

2.2 Access to ‘Hospital Alerting via Electronic Noticeboard’ Database 19

2.2.1 HAVEN-OUH Version 10 ........................................ 19

2.2.2 HAVEN-OUH Version 14 ........................................ 19

2.3 Work to Date: Age- and Sex- Based Early Warning Score .... 23
1 Literature Review

1.1 Introduction

Over the last few decades, medical informatics has witnessed an increased integration of routinely acquired data in clinical settings, formulating extremely large medical datasets, also known as healthcare ‘big data’. Electronic Health Records (EHRs), for instance, securely store a patient’s medical history, providing automated access to information and efficient clinical work flow. Despite the challenges imposed on the secondary use of EHRs due to its heterogeneous nature, EHRs are favored as a rich data source [1]. The increased availability of EHRs has motivated efforts in medical computing to develop diagnostic and treatment tools that aim to improve patient care, such as through applied machine learning.

Estimating antibiotic resistance to treat infectious disease in a timely manner, processing vast intensive care unit (ICU) data to relief staff stress and stratifying risk amongst post-operative patients are just a few examples of machine learning applications in medicine [2]. Machine learning models for early detection of patient deterioration prior to a defined clinical complication have been recently advancing in particular. From modeling time series physiology data to predicting the probability of an adverse event, previous machine learning-based studies have extracted specific patient cohorts based on inclusion and exclusion criteria of the EHRs’ data elements [3].

This literature review aims to assess the challenges of extracting data from EHRs, and to present the current state-of-the-art machine learning techniques developed to detect clinical deterioration on hospital wards using uniquely rich, and complex, medical ‘big data’.
1.2 Clinical Deterioration on Hospital Wards

The concept of clinical deterioration refers to the worsening of a patient’s condition on hospital wards, and is assessed by medical staff through routine observations and protocols. However, the definition of what exactly constitutes clinical deterioration has evolved over time based on the data collection and processing techniques. Early attempts to define clinical deterioration focused on medical neglect and its end result of clinical complications [4]. Subsequent studies were based on more discrete clinical complications, such as severe sepsis, unexpected cardiac arrest, ICU admission or death. Such models assessed data retrospectively and did not reliably inform clinicians of prospective insight, however they highlighted antecedent instability, or serious physiological abnormalities, prior to the complication [5–8].

Current systems aim to detect deterioration using an objective criteria to assist clinicians in real time, such as early warning scores (EWS) systems. EWS systems, also known as ‘track-and-trigger’ (T&T) systems, assign a score to each vital sign according to predetermined normality ranges, and alerts are generated according to the patient’s overall score. Examples of such systems include the CEWS, VIEWS, and MEWS [9, 10]. Other systems have also been developed to specific clinical conditions such as cardiac arrest, EuroScore, and sepsis, SIRS score [11, 12].

The reliability and accuracy of current detection systems are highly dependent on the clinical complications evaluated during modeling, or adverse events, and must consider other patient-, disease- and organizational- factors in the future.

1.2.1 Preventable Adverse Events

Definition, Types and Causes

An adverse event (AE) is defined as unintended injury or complication, which results in disability, death or prolongation of hospital stay, and is caused by health care management rather than the patient’s underlying disease [13]. Routinely defined adverse events in the literature include, or are limited to, any of: unexpected ICU transfer, mortality or cardiac arrest [14].

Preventable adverse events are avoidable adverse events where there was enough knowledge and accepted practices available to have avoided the event [15]. Such events incur high costs of prolonged hospital stays, litigation, staff time, impact on patients and staff, and broader economic consequences.
It has been shown that preventable adverse events mainly occur due to incorrect or delayed drug therapy (mainly associated with antibiotics and antithrombotic agents), surgery, procedures, diagnosis and system issues (hospital-acquired infections or equipment malfunction) \[17\]. Ward management is another causal factor, and its major contributory factor is failure to identify at-risk patients and to undertake preventative measures \[16\].

**Evaluation of Preventable Adverse Event Selection**

Most studies that have developed models for early identification of patients at risk, tend to select one or more adverse events as the end-point measures of clinical deterioration. Since studies vary in terms of data types and underlying techniques, it is important to consider the advantages and disadvantages of the adverse events selection and thus assess the subsequent generalization of the models, especially for the most commonly evaluated adverse events: mortality, cardiac arrest and ICU admission.

Mortality can be reliably retrieved from hospital administrative data, and has the same unambiguous definition across healthcare providers. However, the difficulty lies in differentiating between natural deaths and potentially preventable deaths, especially among medical patients and patients whose medical care was actively withdrawn by the primary team \[14,18\]. Additionally, mortality or readmission soon after hospital discharge due to unexpected medical complications may often occur due to inadequate assessment of the patient’s condition prior to discharge or inadequate home support \[16\]. Such instances must be assessed, and mortality within 24 hours of hospital discharge could be considered as in-hospital mortality. Readmission criteria must also be assessed.

In-hospital cardiac arrest is similarly generalizable, and is explicitly defined as loss of a pulse with attempted resuscitation \[14\]. However, this definition relies on the availability of Do-Not-Attempt Resuscitation (DNR) orders in the data, which are often difficult to obtain. Moreover, the study design must consider cases of deaths following unsuccessful resuscitation attempts, as such instances could either count for cardiac arrest, death or both.

While cardiac arrest is the least commonly occurring event amongst patients, transfer to ICU is the most frequent one \[19\]. However, this event is the least generalizable adverse event of the three, since the criteria for ICU admission varies from one hospital to another and it relies on location- and timing-related factors such as bed availability \[14\]. It is also important to differentiate between unplanned and planned ICU transfers, such as planned postoperative admission after major surgery, and to consider
Table 1.1: Primary Antecedent Instability within \( n \) hours to Adverse Events

<table>
<thead>
<tr>
<th>Study</th>
<th>Clinical Conditions</th>
<th>( n ) (hours)</th>
<th>Adverse Event Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>[5]</td>
<td>Respiratory and metabolic derangements</td>
<td>8</td>
<td>Cardiac Arrest</td>
</tr>
<tr>
<td>[6]</td>
<td>Hypotension, tachypnoea, &amp; tachycardia</td>
<td>8</td>
<td>Death not preceded by cardiac arrest or ICU admission</td>
</tr>
<tr>
<td>[22]</td>
<td>Hypotension, tachycardia, tachypnoea, and sudden change in level of consciousness</td>
<td>8</td>
<td>ICU Admission</td>
</tr>
<tr>
<td>[7]</td>
<td>Low systolic blood pressure (&lt; 90 mmHg), fall in GCS, and low heart rate &amp; respiratory rate</td>
<td>24</td>
<td>Death, cardiac arrest or unplanned ICU admission</td>
</tr>
</tbody>
</table>

excluding neonatal, maternal or ICU admissions directly from the emergency department [17].

### 1.2.2 Antecedent Instability to Adverse Event

Previous studies have investigated antecedent instability to adverse events mainly in terms of physiological instability. Table 1.1 provides a summary of such findings, noting that a standardization of findings is not possible due to variations in patient cohorts and adverse events examined in each study. There has also been interest in changes in laboratory variables. Increased levels in digoxin and serum creatinine, for example, have been investigated prior to iatrogenic cardiac arrest and alongside congestive heart failure [20]. Instability and recovery following adverse events or surgical operations have also been investigated [21].

### 1.2.3 Overall Patient-, Disease- and Organizational- Factors

Overall, a clinical deterioration framework must account for other patient-, disease- and organizational-based factors [23], as summarized in Table 1.2.
### Table 1.2: Factors Related to Clinical Deterioration

<table>
<thead>
<tr>
<th>Patient-related</th>
<th>Disease-related</th>
<th>Organization-related</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Degree of severity</td>
<td>Hospital type</td>
</tr>
<tr>
<td>Lifestyle Factors</td>
<td>Medications &amp; Therapy</td>
<td>Time of day</td>
</tr>
<tr>
<td>Admission Type</td>
<td>Rate of Progression</td>
<td>Staff Ratio</td>
</tr>
<tr>
<td>Coexisting medical conditions</td>
<td>Effects on vital signs</td>
<td>Time between ward transfers</td>
</tr>
<tr>
<td>Baseline factors</td>
<td>Interventions</td>
<td>Staff training</td>
</tr>
<tr>
<td>DNS orders</td>
<td>Surgeities &amp; Procedures</td>
<td>-</td>
</tr>
</tbody>
</table>

1.3  **Data for Machine Learning: Electronic Health Records**

1.3.1  **What are Electronic Health Records?**

In the 1990s, the Institute of Medicine (IOM) argued for the promising use of electronic health records (EHRs), previously known as ‘clinical information systems’ or ‘computer-based patient records’, in improving the quality of health care through R&D focused on developing real-time clinical alerts [24]. EHRs are now increasingly implemented across hospitals worldwide such as in the United States and United Kingdom [25, 26]. The Healthcare Information and Management Systems Society (HIMSS) provides a concise definition of EHRs:

“The Electronic Health Record (EHR) is a longitudinal electronic record of patient health information generated by one or more encounters in any care delivery setting. Included in this information are patient demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports. The EHR automates and streamlines the clinician’s workflow. The EHR has the ability to generate a complete record of a clinical patient encounter - as well as supporting other care-related activities directly or indirectly via interface - including evidence-based decision support, quality management, and outcomes reporting.” [27]

1.3.2  **Electronic Health Records-derived Phenotyping**

The development of machine learning techniques relies on the extraction of specific patient cohorts in the EHR, also referred to as phenotypes. According to the Encyclopedia of Epidemiology, a phenotype “comprises the characteristics, traits, values, or abnormalities that we observe, measure, test, or evaluate in an individual”, which may include “behavioral, biochemical, clinical, molecular, morpho-
logical, physical, and physiological characteristics, as well as the presence or absence of disease” [28]. Based on the availability and quality of data elements, EHR-derived phenotyping can be used to extract specific patient cohorts by defining inclusion and exclusion criteria [3].

Several phenotyping algorithms have been developed in the past. A chronic pain identification algorithm was developed by combining multiple data elements including diagnostic codes related to chronic pain (categorized into ‘highly likely’ and ‘likely’), pain score and medications belonging to the opioid family (including duration and dosage), which proved to be more accurate in identifying patients than using individual data elements [29]. An extraction algorithm that combined structured EHR data with unstructured clinical notes better identified patients who were potentially undergoing dialysis, compared to using any individual source [30]. A more advanced machine learning study deployed active learning methods to identify patient cohorts for rheumatoid arthritis, colorectal cancer, and venous thromboembolism, outperforming passive learning methods [31]. Overall, the portability of phenotyping algorithms requires robust validation methods for adaptive implementation across different institutions [32].

Extracted phenotypes are often used as data for machine learning algorithms.

1.3.3 Challenges of Using Electronic Health Records

Information contained in EHRs can be categorized to four different data types: (1) Categorical data, (2) Continuous data, (3) Textual data and (4) Visual data. In addition to its heterogeneous nature, EHRs reflect the recording process inherent in healthcare, rather than being a direct reflection of the patient and physiology [33]. Thus, this imposes challenges on the usability of EHRs.

1. Complexity

EHRs are complex, as they include structured and unstructured data. Free textual data is an example of the latter, and they are usually synthesized using Natural Language Processing (NLP) techniques [34]. These techniques may retrieve, structure and code encounter notes, symptoms or thought processes to be reliably accessed for further research [35].

Categorical data, such as diagnostic coding, may adopt different coding systems across different institutions, and thus may require manual interventions. Variation in definition may also exist locally within an institution over time. National and international efforts have tried to unify the EHRs by formulating documentation standards, such as the Health Level Seven International (HL7) standards [36].
2. **Completeness**

Data completeness is defined as “the proportion of observations that are actually recorded in the system” [1]. Incompleteness of EHRs can be a result of health service fragmentation due to inefficient exchange following patient transfer among institutions, recording of data only during healthcare episodes that correspond to illness, or the increased personalization of attributes per patient or patient cohort [33,37]. Completeness has been assessed by identifying blanks in the dataset using a computer algorithm [38], and several approaches have been proposed to handle missing data [37]. Completeness may also vary across institutions based on adopted protocols and staff compliance.

3. **Correctness**

Correctness, or “the proportion of recorded observations in the system that are correct” [1], is another dimension of data quality. Errors can occur while clinical staff observe a patient or record data, influenced by random and systematic errors such as influence of billing requirements or avoidance of liability [33]. Accuracy of EHRS can be assessed by looking at agreement between elements within the EHR and verifying whether values are within expected ranges or compliant with general medical information [38].

4. **Currency**

Currency investigates if data is recorded within a reasonable period of time [38]. For example, a patient’s discharge date and time may precede the patient’s date of death due to timeliness inefficiencies among clinical staff. Although this aspect of data quality is highly dependent on the efficiency of the clinical staff, it also depends on the workflow protocols adopted at different institutions.

Timeliness of data must be assessed to evaluate the chronology of data elements in relation to admission or discharge decisions, for example laboratory results prior to admission may be considered as part of subsequent admission, or death within 24 hours of discharge can be considered as in-hospital mortality.

Thus, the quality of data must be assessed through different dimensions prior to pre-processing for further analysis.
1.3.4 Data Preparation for Machine Learning

Medical data from EHRs, or clinical data collected on a phenotypic level in broad terms, are preprocessed as inputs for machine learning models. The inputs can be raw data, derived from parameters or obtained as features from time series analysis.

1. **Raw Data** is readily available data in EHRs such as vital sign measurements, laboratory variables or other categorical data types. This may also include waveform data such as electrocardiogram (ECG) or photoplethysmogram (PPG) signals from wearable sensors [39].

2. **Derived Parameters** can be computed from available clinical data. Examples of derived parameters included in previous studies are pulse pressure [40,41], shock index [40,42,43], mean arterial pressure (MAP) [40,44], oxygen delivery index [43], absolute successive difference of heart rate, estimated cardiac output, slope of fitted regression lines, or slope projections [42]. Binary numerical data variables can also be derived from categorical data.

3. **Statistical Measures** are obtained from the distributions of available data, such as minimum and maximum extremes, moments (mean, standard deviation, and skewness), percentiles or the difference between two percentiles, also referred to as inter-percentile ranges [42].

4. **Time Series Features** can be obtained from time series analysis of waveforms. Time-series representations are categorized into four types: data adaptive, non-data adaptive, model-based and data dictated approaches [45]. Fourier and wavelet transforms, for instance, decompose raw signals into frequency and wavelets respectively [46,47]. Time domain, Poincaré nonlinear, cross-correlation analysis and geometric measures have also been used to investigate variability of vital signs [41,48]. Nonuniform data may often be standardized as time series data for further analysis, and a commonly used technique is Gaussian Processes [49].

The representation of those features have also been evolving. Principle component analysis, for example, is often used to project high-dimensional data to uncorrelated lower-dimensional vectors. A novel unsupervised representation of EHRs has also been recently deployed using deep feature learning [50]. Overall, a variety of pre-processed data can be combined and represented as inputs for machine learning models.
1.4 Machine Learning Approaches to Clinical Deterioration

Machine learning algorithms aim to predict or recognize patterns within unknown data, by training the model on existing data to learn parameters of the underlying function. The generalization of a model depends on the variability of the training data, which only comprises a fraction of the possible input data [51]. Different learning techniques can be characterized based on various attributes:

1. **Type of Learning**: Supervised learning relies on the availability of input data and the corresponding target output. It can broadly be categorized as *regression* and *classification* problems, where regression cases occur when the target output is a continuous variable, and the latter when the target variable is a finite set of discrete categories. Unsupervised learning tackles problems when the target value of the input is not available. Thus, the goal may be to *cluster* unknown groups, *estimate* the probability density function within the data, or *visualize* high-dimensional input spaces [51].

2. **Dimensionality of Parameters**: Training a model on input data, whether it is through a supervised or unsupervised approach, is performed by learning the parameters of the underlying function. A *parametric model* refers to a model trained for *finite-dimensional* parameters, whereas a *non-parametric model* describes its parameters in an *infinite-dimensional* space, where the number of parameters can grow as the amount of data grows. A *semi-nonparametric model* constitutes of a *mixture* of finite-dimensional and infinite-dimensional parameters. The complexity of the model is therefore bounded by the dimensionality choices, even if the data quantity is unbounded.

3. **Quantification of Uncertainty**: Models can also quantify the uncertainty of predicted target variables using probability theory. *Probabilistic modelling* also accounts for possible noise present in the system.

Machine learning models of various characterizations have been employed to analyze, detect or predict clinical deterioration, and are often compared to traditional approaches (i.e. EWS). It is important to note that the **performance and generalizability** of techniques relies on the **contextual choices** of adverse events, patient cohorts, input variables and evaluation techniques.
1.4.1 Clinical Deterioration as a Classification Problem

**Logistic Regression** Loekito et. al. [52] modeled a multivariate logistic regression to investigate the power of key laboratory measurements in predicting imminent death in ward patients. Nine laboratory variables were selected as the most statistically significant from 30 routinely collected laboratory variables, and the area under the receiver operating characteristic curve (AUROC) obtained a value of 0.872, which was considered as a good value in the study.

Another study used Lasso logistic regression to classify two clinical outcomes: in-hospital mortality and 1-year post-discharge mortality. The inclusion of clinical notes decomposed into a time series inferred by hyperparameters of a Multitask Gaussian Process improved the classification performance. The study also constructed a linear kernel support vector machine, which obtained results that were not statistically different from the logistic regression model [49].

Cao et. al. [42] constructed a multivariate logistic regression to predict hemodynamic instability at least 2 hours prior to its occurrence to avoid life-threatening conditions. The possible input space initially included 220 features obtained from vital signs (heart rate, systolic arterial blood pressure, diastolic arterial blood pressure & mean arterial blood pressure), derived parameters and statistical measures of vital signs distributions. After filtering the 13 main features that were highly significant and weakly correlated, the best-performing classification model resulted with an AUROC of 0.82 using a combination subset of the filtered features.

**Support Vector Machines** Ong et. al. [46] investigated the binary prediction of cardiac arrest within 72 hours of presentation to the Emergency Department as the primary outcome, and in-hospital death as the secondary outcome, using multivariate support vector machines with different kernel functions. The predictors were heart rate variability (HRV) variables obtained from time domain, geometric and frequency domain measures of short-term ECG recordings, vital signs (heart rate, blood pressure, oxygen saturation, tympanic temperature, respiratory rate, GCS and AVPU), and age. The linear SVM provided the highest confidence in classifying death and survival in comparison to MEWS.

Kennedy et. al. [43] utilized four data classes (multivariate laboratory variables, vital signs, derived clinical parameters, and time series trend analysis) and their different combinations to predict cardiac arrest in a pediatric ICU. The support vector machine that used a combination of only multivariate variables and time series trend analysis performed best in terms of accuracy and AUROC in comparison to other data combinations and models, namely linear regression, decision tree, and a
neural network.

**Tree-based Models** Churpek et. al. [53] found that random forest trees were more accurate in predicting the composite of cardiac arrest, ward to ICU transfer or death on wards, in comparison to eight other models: two logistic regression models (the first using linear predictor terms and the other using restricted cubic splines), bagged tree model, boosted tree model, K-nearest neighbors, support vector machine using the radial basis function, neural network and the Modified Early Warning Score (MEWS). The models’ input data included age, time since ward admission, number of previous ICU stays, vital signs, and laboratory values (electrolytes, creatinine, liver function tests, and blood counts).

Jarvis et. al. [54] developed an early warning score only using common laboratory data obtained through blood tests (Hb, WCC, U, Alb, Cr, Na, and K). The model was based on Decision Tree (DT) analysis that partitioned the data based on associations of the features with the primary outcome of in-hospital death. The evaluation on 22 test sets obtained AUROC values ranging between 0.755 and 0.801, emphasizing the power of laboratory variables in predicting in-hospital mortality.

Another study constructed an ensemble model using gradient tree boosting and adaptive boosting with age, heart rate, oxygen saturation, respiratory rate, temperature, diastolic blood pressure, systolic blood pressure, pulse pressure, approximate mean arterial pressure, and shock index as inputs, to predict the likelihood of pediatric transfer to pediatric ICU, within 2-8 hours preceding the transfer. The ensemble model performed better than the Pediatric Early Warning Score (PEWS), in terms of accuracy, sensitivity, specificity and AUROC [40].

**Neural Networks** Hu et al. [55] presented a binary classifier to predict ICU transfer or cardiac arrest by constructing a neural network using routine vital signs (heart rate, respiratory rate, temperature, diastolic blood pressure and systolic blood pressure) and laboratory values (white blood cell count, hemoglobin, platelet count, sodium, potassium, chloride, total CO$_2$, blood urea nitrogen, creatinine and glucose). The model performed better than ViEWS by evaluating the AUROC and F1 score within 4 hours to an adverse event.

Aiming to predict ventricular tachycardia one hour before its occurrence, Lee et. al. [48] modeled an artificial neural network trained through back propagation learning rule and a perceptron structure. The model was evaluated using 14 different parameters obtained from heart rate variability (time domain analysis (4), frequency domain analysis (4) and Poincaré nonlinear analysis (3)) and respiratory rate variability (time domain analysis (3)), and it performed best when all 14 parameters were included.
One study investigated hemodynamic instability that may lead to hypotensive events, and constructed a neural network to classify normotensive and hypotensive episodes with a log-sigmoid activation function. The training features were obtained from statistical distributions, cross-correlation and discrete wavelet decomposition of 6 time-series data (heart rate, systolic blood pressure, diastolic blood pressure, pulse pressure, relative cardiac output, mean arterial blood pressure), in addition to clinical features such as age and medications: (1) vaso-constrictors and positive inotropic drugs and (2) vasodilators, diuretics, and sedatives, that tend to raise and decrease blood pressure respectively. The best classifier resulted in a mean AUROC of 0.918 within 1-hour of predicting hypotensive events. The same study constructed a hyperbolic tangent sigmoid and linear function in hidden and output layers, respectively, of the same neural network classification architecture to estimate median arterial blood pressure in a neural network regression [41].

Donald et. al. [44] constructed a Bayesian Artificial Neural Network (BANN) classifier based on the average, slope and variance of heart rate, systolic blood pressure and mean arterial pressure, age and gender to predict hypotensive events 15 minutes prior to its occurrence. The study emphasized the need to maintain a low number of false positives in an intensive care environment.

Futoma et. al. [56] constructed a Long-Short Term Memory deep recurrent neural network to classify whether or not a patient will become or is already septic. The RNN is fed with baseline factors (age, gender and comorbidities), medications (class type and time of administration) and standardized vital signs and laboratory values using Multitask Gaussian Processes (MGP). The MGP uses the Ornstein-Uhlenbeck kernel function to standardize the data, since RNNs typically require uniformly-spaced inputs. This framework yielded the best performance of the AUROC when compared to clinical scores (NEWS, MEWS and SIRS) and other baseline methods, which is explained by the model’s ability to retain uncertainty and noise in the data, and can learn correlations amongst different variables.

Gaussian Processes Classifiers A study used Gaussian Processes Classifiers to predict clinical stability after coronary bypass surgery using different time-series techniques, which generally outperformed logistic regression. The time-series analysis was applied on 14 physiological parameters to obtain signal average, multivariate autoregressive models (MAR), cepstral coefficients, autoregressive exogenous (ARX) coefficients and cepstral coefficients from ARX models. Different combinations of those inputs were evaluated using the Gaussian Processes, with an underlying squared exponential with automatic relevance determination kernel, which were then transformed to classifiers using a
logistic function. Expectation propagation is used to analytically compute the posterior. Evaluated using different input combinations, the Gaussian Process Classifier generally outperformed logistic regression [57].

Also using the automatic relevance determination as the underlying kernel, Skolidis et al. [47] utilized Gaussian Processes Classifiers to distinguish between normal and pre-mature ventricular arrhythmic beats, since cardiac arrhythmias cause morbidity and mortality. ECG signal provided input features through the Fast Fourier Transform and the Wavelet Transform and the most relevant features were selected using automatic relevance determination. Since the noise model is a logistic function, the Laplace approximation is deployed to approximate the posterior, and the final model performed with high precision and average accuracy above 90%.

Another study used Gaussian Processes Classifiers to predict the probability of ICU discharge after cardiac surgery, and the evaluated model resulted with an AUROC significantly higher than classification performance by nurses. It also proved to be the best calibrated model, with less overfitting than nurses, physicians and the EuroSCORE [58].

1.4.2 Clustering Analysis

Partitioning Clustering To identify patient subgroups in a large heterogeneous ICU that may require targeted care efforts, Vranas et al. [59] applied a consensus clustering algorithm using 16 clinical features. The model resulted with 6 significantly different clusters in terms of patient and hospitalization characteristics: relatively healthy and short stay patient, older patient with catastrophic critical illness, post-surgical or procedural patient, older patient discharged with long term care needs, previously healthy patient with prolonged ICU course and finally, the elderly patient with severe illness.

Ghassempour et al. [60] clustered time series trajectories of continuous and categorical data by first associating the time series data with probability densities belonging to the class of Hidden Markov Models (HMMs). Next, they calculated the distances between the probability densities using the symmetrized Kullback-Leibler (KL) divergence, which were later clustered using the Partition Around Medoids (PAM). Parameters of the model were estimated using the Expectation Maximization (EM) algorithm, and the number of hidden states and clusters is chosen by maximizing the Dunn and Silhouette indexes.
Hierarchical Clustering

Gaussian Process Regression (GPR) was used to generate the mean trajectories and associated uncertainties of unevenly sampled heart rates and breathing rates of each post-operative patient in the training dataset during their last recorded 48 hours, using a covariance function that combines squared exponential and periodic kernels. Hierarchical clustering then categorized the training trajectories according to their global similarity. Each cluster is composed of averaged mean functions and variance values of GPs included in that cluster. For a test patient, the local likelihood is computed with respect to the training set, which is then used to compute the global likelihood with respect to the clusters. Overall, four clusters that can identify abnormal trajectories, and the hierarchical clustering model performed better than its Dynamic Time Warping matching [61].

Hensman et. al [62] introduced hierarchical Gaussian Processes to model and capture the structure of gene expression data, by proposing a novel covariance function. The proposed model accounts for biological variation and accounts for random noise. A Bayesian hierarchical approach draws the mean of each additional layer from as the Gaussian Process function of its preceding layer, in order to combine replicates of a particular gene’s time series. The covariance functions of the layers may be different with independent parameters. This method performed well in imputing entirely missing data and clustering.

Clustering of time series data, in particular, relies on the time series representation, similarity or distance measures between clusters, clustering prototypes and clustering methods. Aghabozorgi et. al. [45] discusses the different choices of this process and further elaborates on model-based, grid-based and density-based clustering techniques, in addition to partitioning and hierarchical clustering.

1.4.3 Gaussian Processes for Modeling & Forecasting Physiological Trajectory

Clifton et. al. [63] adopted a patient-specific approach and constructed independent Gaussian Processes Regressions for the patient’s each vital sign (heart rate, blood pressure, respiratory rate and oxygen saturation) with a square-exponential kernel function. While this method resulted with similar performance to support vector machine regressor, it offers a probabilistic approach while providing confidence intervals of the estimated values at an unknown test point. The novelty scores of the personalized GPs exceeded the alerting threshold within 21.5 hours before ICU readmission.

A study used Multitask Gaussian Processes (MTGP) to simultaneously model multivariate physiological time series, including respiratory rate, heart rate and systolic blood pressure, using a convoluted-
kernel approach such that each vital sign is modeled with independent hyperparameters. The MTGP provided significant improvement in estimating unknown values of vital signs compared to Single Task Gaussian Processes (STGP) applied to each vital sign time series independently. Both methods were optimized by minimizing the negative log marginal likelihood as a cost function [64].

Similarly, Ghassemi et. al. [49] also used MTGP to predict clinical instability. In comparison to STGPs, MTGP provided significant improvements in interpolating missing values of intracranial pressure and mean arterial blood pressure of Traumatic Brain Injury patients, to estimate the Pressure-Reactivity Index (PRx), an indicator of cerebrovascular autoregulation which, if impaired, may lead to secondary brain damage and mortality.
2  |  Research Proposal

2.1  Thesis Contributions to the Field

2.1.1  Personalized Patient Time-Series Modeling

This thesis aims to develop novel modeling techniques for patients on a personalized level. The time-series model aims to aggregate various data types included in a patient’s EHR: vital signs, laboratory variables, diagnostics, medications and other procedures. Thus, our work adds the advantages of (i) massive scale, and (ii) therefore the possibility of true patient-specific analysis through having a sufficiently large database to construct meaningful covariance functions, using Gaussian Processes for example, that are tailored to the individual.

2.1.2  Novelty Detection for Clinical Deterioration

The modeling of time-series data will be used to develop supervised and unsupervised novelty detection methods for clinical deterioration. This includes classification, clustering and forecasting analysis. Clinical deterioration will be primarily defined as preventable mortality, and thus unexpected ICU admission, and patient cohorts from the admitted patients and accident and emergency patients will be investigated separately. This thesis aims to the first fully-predictive means of performing detection for clinical deterioration, rather than merely identifying deterioration as it begins.
2.2 Access to ‘Hospital Alerting via Electronic Noticeboard’ Database

The *Hospital Alerting via Electronic Noticeboard* (HAVEN) project is a collaboration between the University of Oxford, the University of Portsmouth, Portsmouth Hospitals NHS Trust, and Oxford University Hospitals NHS Foundation Trust. HAVEN aims to produce a hospital-wide IT system to continuously assess risk of hospitalized patients, and at its core is the HAVEN Database, an integrated database of EHRs collected in Oxford and Portsmouth Hospitals [65].

Upon completing all ethics requirements, we have been granted approval to access to the Oxford University Hospitals (OUH) and Portsmouth Hospitals (PH) data to conduct our research. We have only been provided the HAVEN-OUH data so far.

2.2.1 HAVEN-OUH Version 10

The first version of the database contained data related to vital signs observations, cardiac arrest calls, ICU stays and dates of death for adult patients admitted between 11 January 2014 to 24 October 2016. Patient demographics included date of birth and sex. This database was used for the development of the Age- and Sex- Based Early Warning Score, explained in Section 2.3.

2.2.2 HAVEN-OUH Version 14

The most updated version of the database includes additional data that is crucial to more specific patient cohort extraction. As shown in Figure 2.1, the database includes all information related a patient’s flow in the hospital: patient admission method, ward stays and ward transfers, corresponding consultant episodes and specialty codes. The database also includes Accident & Emergency Attendances.

The updated database contains a wide variety of data collected from the patient during their hospital stay:

1. **Anonymized Patient Demographic Information**

   For each admission, the date of birth, sex, ethnicity, and Multiple Deprivation Index are included. It is important to note that a patient may have multiple admissions recorded. The Multiple Deprivation Index is developed by the UK to summarize the status of deprived areas in English local councils, calculated using several factors such as income, employment
2. Patient In-hospital Journey

- **Admission** data includes admission date and time, dementia assessment, patient classification and type of admission. HAVEN only considers elective and emergency admissions, and excludes other types of admissions such as maternal admissions or transfers from other hospitals. Elective admissions occur when a doctor schedules a hospital admission for the patient, while emergent admissions occur when patients are transferred from the Accident & Emergency Department or are immediately referred by a General Practitioner (GP).

Each admission also includes the Summary Hospital-level Mortality Indicator (SHMI), or “the ratio between the actual number of patients who die following hospitalisation at the trust and the number that would be expected to die on the basis of average England figures, given the characteristics of the patients treated there” [66]. It also includes derived Charlson Comorbidity Index (CCI) and Clinical Classifica-
tion Software (CCS). The CCI contains 19 categories of comorbidity and predicts ten-year mortality for patients by assigning a score for each co-morbid condition, while CCS clusters patient diagnoses and procedures into clinically meaningful categories [67,68].

• **Consultant Episodes** data includes episode start and ending time, treatment function code and the consultant’s specialty code, where consultant refers to the specialty doctor. Specialty codes can be categorized as *surgical* (elective and emergency patients who undergo surgery) or *medical* (patients requiring treatment due to an underlying disease or illness). Some patients may have consecutive consultant episodes per admission or several consultant episodes at a time, but only the primary episode is considered.

  Each consultant episode has a [clinical classification of a diagnosis](#) and its corresponding position, since a consultant episode may have multiple diagnoses. The diagnoses are classified according to the International Classification of Diseases (ICD-10).

• **Ward Transfers** data includes the start and ending time of the patient’s stay in a particular ward.

• **Theatre Visits** refer to operation rooms, and related data includes start and ending time of session, start and ending time of the operation, time of anesthesia and start and ending time of recovery.

• **Discharge** data includes discharge date, time, type and destination. Date of death can be determined from the type of discharge.

3. **Vital Signs Observations**

The vital signs set contains timestamped sets of observations for each admission, and the corresponding location of where the observations were recorded. Each observation set includes heart rate, respiratory rate, diastolic blood pressure, systolic blood pressure, oxygen saturation, temperature, Alert Voice Pain Unresponsive (AVPU) Score, Glasgow Coma Scale (GCS) Score, fraction of inspired oxygen (units: percentage and liters per minute), and type of oxygen support.
4. **Laboratory Measurements**

The laboratory measurements include results, date and time of when each test was performed and date and time of when the test’s results were obtained.

5. **Procedures**

Procedures are coded according to the Classification of Interventions and Procedures (OPCS) version 4, used by the National Health Services (NHS). This data element includes date of intervention and order of administration, for situations when a patient undergoes multiple interventions.

6. **Adverse Events**

- **Mortality** can be obtained from date of death.
- **ICU Stays** related data includes admission and discharge time, admission type, planned or unplanned, and delay between ICU decision to admit and actual admission time (in hours).
- **Cardiac Arrest Calls** related data includes time and location of call. DNR orders are not available.
- Other adverse events can be interpolated by searching for corresponding physiological instability, diagnosis, medical interventions or procedures.

7. **Accident & Emergency Department Attendances**

Accident & Emergency (A&E) attendance refers to individual patients visits to the A&E department to receive treatment from the A&E service. Related data includes diagnostics, treatments and whether the patient was later transferred to the hospital as an emergency admission.

More information about the data elements can also be found on the NHS Data Dictionary Website [www.datadictionary.nhs.uk/](http://www.datadictionary.nhs.uk/).
2.3 Work to Date: Age- and Sex- Based Early Warning Score

The first year of my DPhil focused on understanding the characteristics of clinical settings and the current methods deployed to detect clinical deterioration, primarily Early Warning Scores (EWS) Systems. Due to the lack of personalized EWS systems, the Age- and Sex- Based Early Warning Score (ASEWS) was trained using the first version of the HAVEN-OUH Database described in Section 2.2.1. Using the vital signs and the occurrence of the first adverse event, defined as cardiac arrest, ICU transfer or death, ASEWS was developed using a statistical method (Part 1 of the Transfer Report).

2.4 Immediate & Future Work: Machine Learning Pipeline

The current development of the ASEWS assumed that the training vital signs are independent and identically distributed random variables, and no corrections between them were considered. Upcoming work will further investigate the variables as time series data, accounting for correlations within one variable and across different variables. The Work Plan (WP) for the next two years is summarized in the Gantt Chart in Figure 2.3.

WP1: Develop ASEWS

Since the ASEWS assumes that the vital signs are independent and have equal weighting in the patient’s overall risk score, the immediate work to develop ASEWS includes combining systolic and diastolic blood pressure into one derived parameter, such as mean arterial pressure. This would account for correlations between those two variables. Diagnostics and in-hospital administered medications could further enhance patient extraction, such that patients with abnormal vital signs due to underlying comorbidities (hypotension, hypertension, etc...) or therapy are excluded or accounted for.

Moreover, the current model is tested on patient data also recorded at the Oxford University Hospitals, since the dataset was randomly split into a training and test set. This does not assess the model’s generalizability, since some of the patient- and organization-related factors are constant amongst the training and test set. Thus, the model performance may vary with demographics and institutional variations. To verify this, the model will be tested on the Portsmouth Data, which will also be retrieved from the HAVEN database.
WP1: ASEWS
Test on Portsmouth Data
Review & Submit Article

WP2: EHR-derived Phenotyping
Annotate & Structure variables
Inclusion & Exclusion Criteria

WP3: Time-Series Modeling
Fusing Continuous & Categorical Data
Personalized kernels

WP4: Classification Methods
Features Representation
Apply Classification Techniques

WP5: Clustering Techniques
Laboratory Variables Clustering
Phenotypes Clustering
Hierarchical Learning

WP6: Causality Inference
Effect of Interventions
Model Underlying Physiology

WP7: Submissions
Write-up for conferences
Machine Learning Summer School
Confirmation of Status
Thesis Write-up
Thesis Submission
Viva

Figure 2.2: Gantt chart describing the Work Plans (WPs) for the next two years, including analysis and submissions.
WP2: EHR-derived Phenotyping

The first step for laboratory, prescriptions, procedures and diagnostics variables is to create a structured ontology of their clinical use and significance, expected values and extreme values. The Open Biomedical Annotator will be investigated for its use to structure and annotate the codes [69]. This would help in extracting patient cohorts by defining specific inclusion and exclusion criteria.

WP3: Time-Series Modeling

Time-series modeling will primarily focus on fusing the different data elements, to model the patient’s physiology. Gaussian Processes will be used to investigate different kernel combinations for different patients, aiming for a true-patient specific analysis.

WP4: Classification Methods

Using the modeled time-series data of the patient cohort, other feature representation approaches will be investigated, such as the ‘deep patient’ approach or waveform analysis [50]. The different representations will be used in classification models, primarily comparing support vector machines, random forests and neural networks. This would investigate supervised machine learning approaches.

Figure 2.3: Machine learning pipeline to construct models that predict, analyze or classify clinical deterioration

WP5: Clustering Techniques

An unsupervised approach will also be investigated, primarily to cluster the time-series data. This will be performed for the overall time-series representation of patients, but also independently for laboratory variables, medications and phenotypes. Figure 2.3 represents the data extraction and analysis process for each machine learning model. Hierarchical clustering of the personalized Gaussian Processes will also be investigated.
WP6: Causality Inference

In the later stages of the thesis, the work will focus on the effect of interventions, especially medications and procedures, on physiology, i.e. vital signs and laboratory variables, in order to model the true underlying physiology without medical intervention. The challenge in this investigation is that the ground truth of the underlying physiology is not available because the actual effect is unknown and could vary from one patient to another, and thus it must be inferred from the data.

WP7: Thesis Writing and Other Submissions

Write-up for DPhil requirements and conferences will occur simultaneously with the analysis. The main submissions include the confirmation of status and final thesis.

Thesis Expected List of Chapters

1. Background of Clinical Deterioration on Hospital Wards
2. Electronic Health Records: Phenotyping Platform and Features Representation
3. Personalized Time-Series Modeling
4. Hierarchical Learning
5. Causality Inference

2.5 Risk Assessment

The upcoming research plan has low anticipated risk. Access to HAVEN data pertaining to OUH and PH has already been granted through a relational database server set up by senior researchers at the Institute of Biomedical Engineering (IBME). Due to the ethics requirements and sensitivity of the medical data provided by the HAVEN project, it is important to note that any published works using the Portsmouth data must be first ratified by colleagues at Portsmouth, and any journal or conference papers or presentations generally pertaining to HAVEN must first be reviewed by the Wellcome Trust prior to publication. Thus, timeliness of publications must be borne in mind prior to deadlines.
In addition to trying to mitigate the challenges of dealing with EHR data, such as dealing with missingness or combining structured and unstructured data, the following challenges may arise:

1. **Generalisability of Models**

Since machine learning models learn the parameters based on the training set of our patient cohort, the problem of overfitting may arise. The performance on the test set will be assessed using commonly used measures: area under receiver operating curve, sensitivity, specificity, F1 score, etc ... [70]

This would also limit the performance of our models on other datasets collected at different clinical settings, given that the scope of this research mainly focuses on patient- and disease-related factors to clinical deterioration, but not on organization-related factors. One way to validate our models is to evaluate the performance on independent datasets, such as training on OUH data and testing on PH data, keeping in mind differences between the two. For example, the PH dataset does not collect data in the A&E Department.

In terms of Gaussian Processes, generalisability can be explored by choosing different kernels for different patient cohorts. On a more personalized level, we seek to investigate using different kernel combinations for different subjects, accounting for more specific patient-related factors.

2. **Machine Learning ~ ‘Black Box’**

There has been an increasing tendency to view machine learning models, such as neural networks as well as Gaussian processes, as black-box models that are difficult to interpret [71–73]. This has important implications on the medical community’s acceptance of such advancements, as it requires achieving trust from patients and medical staff. We hope to mitigate this risk in the future by working closely with medical staff and to explain the mathematical background of our methods to the best of our knowledge and capabilities.

To conclude, working with a rich and extremely large dataset of real-world hospital data presents an exciting opportunity to apply novel machine learning techniques to potentially improve patient care, in the years to come of my DPhil in Engineering Science.
Bibliography


[26] Ann Robertson, Kathrin Cresswell, Amirhossein Takian, Dimitra Petrakaki, Sarah Crowe, Tony Cornford, Nicholas Barber, Anthony Avery, Bernard Fernado, Ann Jacklin, Robin Prescott, Ela Klecun, James Paton, Valentina Lichtner, Casey Quinn, Maryam Ali, Zoe Morrison, Yogini Jani,

[27] Electronic Health Records | HIMSS.


[67] Charlson Index Online Calculator.

[68] Clinical Classifications Software (CCS) for ICD-9-CM Fact Sheet.


