Eye Tracking the Visual Attention of Nurses Interpreting Simulated Vital Signs Scenarios: Mining Metrics to Discriminate Between Performance Level

Jonathan Currie, Raymond R. Bond, Paul McCullagh, Pauline Black, Dewar D. Finlay, Aaron Peace

Abstract— **Nurses welcome innovative training and assessment methods to effectively interpret physiological vital signs. The objective is to determine if eye-tracking technology can be used to develop biometrics for automatically predict the performance of nurses whilst they interact with computer-based simulations. 47 nurses were recruited, 36 nursing students (training group) and 11 coronary care nurses (qualified group). Each nurse interpreted five simulated vital signs scenarios whilst 'thinking-aloud'. The participant's visual attention (eye tracking metrics), verbalisation, heart rate, confidence level (1-10, 10=most confident) and cognitive load (NASA-TLX) were recorded during performance. Scenario performances were scored out of ten. Analysis was used to find patterns between the eye tracking metrics and performance score. Multiple linear regression was used to predict performance score using eye tracking metrics. The qualified group scored higher than the training group (6.85±1.5 vs. 4.59±1.61, p=<0.0001) and reported greater confidence (7.51±1.2 vs. 5.79±1.39, p=<0.0001). Regression using a selection of eye tracking metrics was shown to adequately predict score (adjusted R^2=0.80, p=<0.0001). This shows that eye tracking alone could predict a nurse's performance and can provide insight to the performance of a nurse when interpreting bedside monitors.**

***Index Terms*—Task Performance; Patient Monitoring; Patient Simulation; Simulation-Based Training; Visual Attention; Eye Tracking**

J. Currie, R.B. Bond, P. McCullagh, P. Black and D.D. Finlay are with Ulster University, Northern Ireland. Aaron Peace is with the Clinical Translational Research and Innovation Centre (C-TRIC) at Altnagelvin Area Hospital, Northern Ireland. This research has been funded by the Department of Employment and Learning (DEL), Northern Ireland. Correspondence email: currie-j@email.ulster.ac.uk.

INTRODUCTION

Patient safety is a critical area of concern within healthcare and medical errors are a well-known problem that can have fatal ramifications [1]. Lack of knowledge and skill with clinical tasks and procedures, as well as decision-making can be significant factors with many of the errors that are reported in healthcare [2]. Research into high-level skills and performance attributes a critical role for 'continual practice' and maximizing training time in order to reach a specific performance level [3]. Concepts and techniques that stem from early work on 'working memory' [4] underpin proficiency-based progression using simulation [5]. Proficiency-based progression, removing vulnerable patients from the setting, has proven that tasks can be simulated and those skills measured precisely [6], [7]. This facilitates assessment and feedback, thereby improving the skills of the trainee and can ameliorate lack of training time, facilities and expertise available to trainees. Many healthcare tasks can be simulated using computer and web technology for training purposes and provide trainees (students and practicing) with a way to improve or maintain their knowledge and skills [8], [9], [10]. The area of simulation-based training, especially in the form of screen-based simulation (using computer application/web browser), still requires further research to fully demonstrate how it can provide an adequate training and assessment of skills in comparison to higher-fidelity modalities (e.g. manikin, standardised patients and haptic simulators) [11]. This study involved capturing the visual attention of nurses while interpreting five simulated patient scenarios (representing 'bedside' physiological monitors with accompanying vital signs). The aim was to explore the potential for visual attention, via eye gaze, such that the derived metrics could become a new component of assessment. This assessment could then be used for automatically and non-intrusively measuring the performance of trainees. The technology could have further implications for patient monitoring at the bedside with practicing nurses and facilitate long-term measurement of their competency.

A. *Nursing and Patient Monitoring*

Patient monitoring is a core role of the nurse [12], [13], involving surveillance of the patient and the patient's physiological signals (usually in the form of a vital signs monitor or central monitoring unit).

According to Wheatley [14]:

> *"The interpretation of data from assessments is vital in determining the level of care a patient requires, providing treatment and preventing a patient deteriorating from an otherwise preventable cause."*

This task is imperative to the early identification of the deterioration of a patient's condition [15], [16]. Vital signs monitoring (e.g. temperature, heart rate, blood pressure and respiratory rate) has been integral to care for over 100 years but we still read reports of inadequate monitoring and decision-making [17]. The detection and reporting of these vital signs are critical, as delays with required treatment can have a significant impact on patient outcomes [18]. There is a desire for more consistent practice of high-level patient surveillance [19] and promotion of more effective assessment of patient physiological status [20].

B. *Simulation-Based Training for Nursing*

Simulation-based training can be a valid form of training for nursing students and any practicing nurse seeking skill development [21], [22], [23], [24]. Patient monitoring training via SBT can be provided using standardized patients, manikin based technology or screen-based (computer/web) simulators [11]. Screen-based simulation training may be able to combat a known problem with training in healthcare, i.e. the lack of overall training time and frequency (repeatability) [25]. Web-based simulation is becoming a suitable candidate for nursing educators to deliver training on many scenarios that require decision-making skills and task/device knowledge [8]. Medical and healthcare simulation development has been growing in recent years [7], [26]. However educational institutions do not approve every screen-based simulation-based training solution for a task that a physician or nurse would undertake but they are aware of the gaps that exist [27]. Screen-based training also lacks tactile and haptic components during the simulation [11]. One area of interest in researching performance is the link between visual attention, attentional capacity and task performance (specifically high-level performance). This area could investigate if visual attention metrics, while performing a set task, provide insights to the participant performance level.

C. *Visual Attention and Patient Safety*

The concept of visual attention during a task has been tested in many medical and healthcare studies [28], [29], [30], [31], [32]. The mind-eye hypothesis [33] states that measurements of visual attention may indicate underlying cognitive activity [34], [35], [36]. Put differently, could where someone looks indicate their training level, their current state of awareness, uncertainty and most critically, the likelihood that their future actions could cause harm to patient? An example is a recent study performed with surgical tasks [29] which was able to distinguish between novices and experts using eye tracking metrics alone.

D. *Objective*

The use of eye tracking technology by healthcare practitioners has been recently reported in the research literature. The study of this previous work has led us to hypothesise that eye tracking metrics exclusively have a relationship with task performance and can discriminate between performance level observed when nurses interpret patient vital signs from a monitor. The objective is to determine if eye-tracking technology can be used to develop biometrics for automatically predicting the performance of nurses whilst they interact with computer-based simulations.

Methods and Materials

This exploratory research study captures the visual attention of nurses (with varied experience levels) when reading patient vital signs, interpreting them and making recommendations while using a vital sign monitor at the bedside. A series of simulated vital signs scenarios assess their ability at this task. These were designed and validated by expert nurses at Ulster University. The participant's visual attention was captured using a Tobii X60 Eye-Tracker[1]. This non-intrusively acquires eye gaze fixations using invisible infrared light that reflects off the cornea and employs trigonometric functions to approximate the loci of the eye gaze, providing us with eye tracking metrics. The eye tracker data rate is 60Hz with mean latency of 30-35ms, precision/accuracy of 0.5°, max gaze angles of 35° and the tracking is accomplished by light and dark pupil tracking. The participant's heart rate during the performance was recorded using photo plethysmography via Empatica's E4 wristband[2]. Their responses to the NASA-TLX[3] questions, post-performance, were used to measure their cognitive load during the entire performance. There are no direct comparable studies conducted previously for visual attention during interpretation of patient vital signs (although similar research has been conducted with patient monitoring within the patient room with no particular task focus [36]). As a result, all performance metrics were collected for analysis since no underlying models currently exist in the literature. This facilitates a free exploration of the data to uncover what patterns exist. The research study was submitted to the Faculty of Computing and Engineering Ethics - Filter Committee at Ulster University and was approved before data collection began (ref: 20150901-15.39).

A. *Simulated Vital Signs Scenarios*

A series of simulated vital signs scenarios were developed to assess their ability at this task. The scenarios were developed specifically for this study as no standardised scenarios were available which would have been applicable for all participants. The baseline for each scenario was developed from curriculum learning outcomes specified by the Nursing and Midwifery Council standards for pre-registration nurse education in the United Kingdom (UK) [37] and the format based on typical scenarios of patient care that would be encountered by nurses in practice. Each scenario and the score allocation guide was sent to three nurses deemed as expert by their professional and academic qualification (all nurses educated to Master's level), clinical practice and education experience (each at least 15 years across cardiology, neurology and critical care settings) and active involvement in Resuscitation Council UK Immediate and Advanced Life Support training courses which include the assessment of the sick patient [38]. They were asked to comment on the realism and clinical accuracy of the scenarios and the expected performance level as reflected in the score allocation guide. Their feedback indicated that the scenarios reflected the symptomatology of the patients and the expected care response was realistic for current nursing practice.

The information provided in Table I details the scenarios used in this study for each participant, including the vignette and Table II details the assessment criteria for the performance score they received. The criteria assessed participants' verbal responses at three levels and scores were allocated according to:

- Low-level criteria: identification of abnormalities in the presented vital signs.
- Mid-level criteria: identification of why the abnormalities occurred based on their knowledge and understanding of the presenting condition outlined in the case scenarios.
- High-level criteria: identification of and decision-making about the immediate interventions required to stabilise the patient.

[1] http://www.tobii.com

[2] http://www.empatica.com

[3] http://humansystems.arc.nasa.gov

TABLE I. SIMULATED VITAL SIGNS SCENARIOS

Patient	Vignette	*ECG*	*Heart Rate*	*Arterial Blood Pressure*	*Oxygen Saturation*	*Respiratory Rate*	*Temperature*
		Rhythm	**Beats / min**	**(Diastolic/Systolic) mmHg**	**%**	**Breaths / min**	**˚C**
James	"69-year-old man who is being cared for on a medical ward. He was diagnosed with COPD (chronic obstructive pulmonary disease) 10 years ago. He complains of shortness of breath when you are checking his observations as part of his assessment. What do his vital signs suggest to you?"	Sinus Tachycardia	108	140/86	85	25	37.2
Charlie	"45-year-old man admitted to the cardiology ward three hours ago complaining of acute chest pain. You are taking over his care and carry out a set of observations. What do his vital signs suggest to you?"	Atrial Fibrillation	92	100/60	95	18	36.6
Susan	"50-year-old lady who has just returned from theatre to your surgical following abdominal surgery. As you settle her, she complains of severe pain. What sense do you make of her observations?"	Sinus Tachycardia	140	72/46	90	26	36.1
Elizabeth	"65-year-old lady admitted to the respiratory ward with an acute respiratory infection. She started on her third course of intravenous antibiotics yesterday and has called you over as she is experiencing acute coughing. What do the following observations suggest?"	Normal Sinus	92	160/82	90	25	37.8
Joe	"18-year-old student who has been admitted to the Emergency department after being found in a drowsy state by his friends following a party last night. What do his vital signs suggest to you?"	Sinus Bradycardia	45	100/55	90	6	36.2

TABLE II. SCENARIO INTERPRETATION ASSESSMENT CRITERIA

Patient	Simulated Vital Signs Scenario Interpretation Criteria		
	Basic-level (5 points)	*Mid-level (2 points)*	*High-level (3 points)*
James	1. Heart rate is high. 2. Normal blood pressure. 3. High respiratory rate. 4. Low saturation. 5. High temperature.	1. James has COPD so his saturation would normally be low, but his respiratory rate is high. 2. His high temperature (and heart rate) could suggest a chest infection.	1. A sputum sample should be taken to see if infection is present. 2. He should be nursed in an upright position. 3. Oxygen therapy should be considered (with close monitoring).
Charlie	1. Heart rate is high, with an irregular rhythm. 2. Low blood pressure. 3. Respiratory rate higher end of normal. 4. Saturation lower than it should be. 5. Temperature normal.	1. Charlie is young and has chest pain, it could be a sign of an acute MI. 2. He is in an abnormal rhythm which could indicate damage to his heart.	He needs: 1. to have a 12 lead ECG. 2. to be referred for an urgent medical assessment. 3. to have prescribed medication administered for any pain.
Susan	1. Heart rate very high. 2. Blood pressure very low. 3. Respiratory rate high. 4. Saturation low. 5. Temperature low.	1. She is at risk of haemorrhaging after surgery. 2. Her vital signs suggest that she is experiencing hypovolemic shock.	She needs: 1. an emergency assessment by the surgical team to find out where she is bleeding from. 2. IV fluids/blood/fluid resuscitation urgently. to return to theatre as an emergency.
Elizabeth	1. Heart rate high. 2. Blood pressure high end of normal. 3. Respiratory rate high. 4. Oxygen Saturation low. 5. Temperature high.	1. She could have dislodged a plug of sputum which is causing her to cough. 2. She may require oxygen therapy to correct her oxygen saturations.	She needs: 1. a full respiratory assessment with auscultation to assess her air entry. 2. nebulized humidification to help her expectorate effectively. 3. chest physiotherapy.
Joe	1. Slow heart rate. 2. Blood pressure lower end of normal (for his age). 3. Low respiratory rate. 4. Saturations low. 5. Temperature low.	1. He may have taken drugs during the party – either deliberately or by having his drinks spiked. 2. His symptoms need to be treated by oxygen therapy.	1. It is important to find out what substances were taken. 2. Was this recreational drug use or deliberate drug overdose? 3. He may also be intoxicated so an assessment may be difficult.

Their scores out of 10 (referred as *Performance Score*) were then put into categories for *Performance Level* according to expert advice: 0-5 = low-level, 6-7 = mid-level, 8-10 = high-level. A consensus was established that we would categorize their *Performance Score* by the same labels as *Performance Level* (low, mid, high), despite there being no direct relationship between the criteria levels and the total score awarded for each scenario. This would allow us an easier understanding of range of performances within the dataset collected. The lead nurse practitioner led the design process by identifying some commonly encountered nursing assessment scenarios involving a variety of clinical practice settings and a range of symptoms and conditions to include key body systems – respiratory, cardiovascular, neurological and the impact of deterioration on the body. Scenarios covered a range of activities specified in the ABCDE assessment framework recommended by the Resuscitation Council UK and pitched at a level judged suitable for student nurses (who had covered the knowledge and application to practice in their degree programme) and nurses who were actively registered to practice in the UK. Storyboards were initially used to provide an agreed structure and details were incorporated to add clinical and practice realism.

The vital signs monitor in these scenarios displays an electrocardiogram lead, heart rate, the waveform and numeric components for arterial blood pressure, central venous pressure, oxygen saturation and the numeric displays of respiratory rate and temperature. The Tobii software allows us to present media to the participant. The sequence of scenarios provided the participant with a text briefing (vignette), followed by a video recording of a simulated vital signs screen each time. The recording of the simulated vital signs screen was taken from a simulated monitor component of Laerdal's SimMan®[4] and was approximately 1 minute in duration.

B. Recruitment

The study included participants who were either currently in training (undergraduate level) or who had already received a nursing qualification. Recruitment was performed using convenience sampling (including no bias or incentive) from two locations within Northern Ireland: (1) School of Nursing, Ulster University, (2) Clinical Translational Research and Innovation Centre, Altnagelvin Area Hospital. Location 1 provided recruitment of student nurses, referred to as the training group (n=37, mean age=27.31 years, mean experience=0 years), with a prediction that scores would be skewed towards lower performances. Location 2 provided recruitment of coronary care nurses, referred to as the qualified group (n=11, mean age=31.91 years, mean experience=8.73 years), with a prediction that scores would be skewed towards higher performances.

C. Protocol

The protocol was designed to deliver a non-intrusive collection of biometrics (visual attention & heart rate), NASA-TLX, and verbal data (think-aloud) while the participant performed interpretations for five scenarios. (1) The participant was brought into room, they read the study information sheet provided to them and signed the consent form. (2) They were asked to sit in the chair in front of a display monitor and eye tracker (shown in Fig. 1). (3) Then were asked to maintain the same distance from chair to desk. No restrictions were made for whether they should adjust their visual attention or not from the monitor during their performance. (4) They were asked to wear the E4 wristband during the interpretations. Assistance with this was provided if needed. (5) The participant had their eye tracking calibrated (required by Tobii software). Adjustments were made accordingly if the calibration was not able to complete. Calibration is mandatory before recording can begin. (6) The participant was briefed the following:

A. They will perform five interpretations.
B. For each scenario, they will first read a text vignette which is a brief of the patient's context scenarios. Note: The vignette was-provided on the LCD screen to the participant before the vital signs were shown. Participants had as much time as needed to read the short vignette and prime themselves on the scenario. They did not have a version of the vignette on paper. However, the participant gave a signal to the investigator once they felt ready and satisfied to move on and interpret the vital signs.
C. They will indicate to the coordinator they are ready to see the patient vital signs.
D. Once the vital signs are visible on-screen, they should think aloud, interpret and make recommendations for the patient.
E. When each scenario performance (interpretation) is finished, they will provide their confidence for it, 1-10 (1=not confident, 10=very confident).
F. They can ask to stop the study at any point if they feel uncomfortable.

(7) The participant performed their readings and interpretations for the five scenarios. (8) Once finished they removed the E4 wristband (if worn). (9) They were finally asked if they would respond to the NASA-TLX survey before finishing the study and leaving.

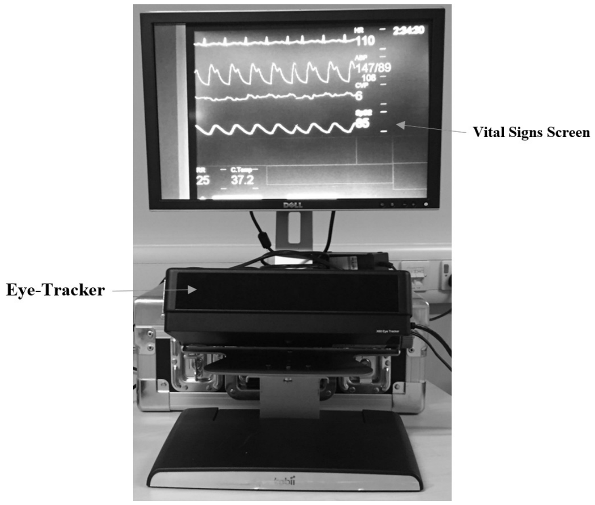

Fig. 1. Participant point of view during scenario performance - vital signs screen presented on display monitor and stationary eye tracker positioned below.

[4] http://www.laerdal.com

D. Data Analysis

The scores attained by the participant and used in the analysis are summations of the points awarded from the three levels of criteria. The performance levels are given similar names but classification analysis using these categories are not reported in this paper (future work will look at this). We are using the total score awarded for each scenario as a continuous variable, in both correlation and statistical model development. The eye tracking data was screened for recording quality provided by the Tobii software. Recording quality is a measurement of total eye gaze data recorded during the participant's performance – the higher percentage, the more data collected. Unfortunately, this means any attention away from the screen results in a low result for this measurement. That does not necessarily mean poor experimental design or mechanical fault with the eye tracking hardware. However, a measurement of 0% does certainly indicate a fault with the recording. The recording quality was tested for correlation to the performance score but a small effect without statistical significance was found (r = -0.09, P = 0.161). We tested the correlation to confirm whether recording quality as a variable could be interpretable to participant behaviour during eye tracking. Only recordings with >50% recording quality were included in this analysis. A cut-off of 50% was used to allow inclusion of participants who did look away but also to maintain a certain amount of visual attention on the screen. Numerous eye tracking metrics, provided by the Tobii X60 eye tracker were measured during the scenario performances. These eye tracking metrics are described in Table III. In addition to those 99 metrics, two further calculations were made to create *Fixation Frequency* and *Visit Frequency* across all on-screen objects during any recorded eye gaze activity. *Fixation Frequency* and *Visit Frequency* were calculated manually from other metrics provided. We did this by using the summation of *Total Visit Duration* (seconds) for all areas of interest (AOIs) and the non AOIs, providing us with suitable total eye gaze activity duration to use for the vital signs portion of each scenario. We then divide this by the summation (all AOIs and Non-AOIs) of *Fixation Count* (for *Fixation Frequency*) and *Visit Count* (for *Visit Frequency*). This should be correct regardless of head movement away from the screen. Delta heart rate *(*ΔHR) is measured as the difference between the participant's final heart rate, once the performance is complete, and the initial recorded heart rate when they began interpreting. NASA-TLX's question responses are quantified and used for an overall *NASA-TLX Score* (out of 600). Then we convert to a percentage to measure the participant's self-reported cognitive load. All correlation tests were performed using a Pearson product-moment correlation (r) assessing the degree of the linear relationship between two variables. An independent t-test was used for significance testing (where $\alpha = 0.05$), using a Welch two-tailed t-test assuming unequal variance. Bonferroni correction to the alpha value was applied given there are many significance tests. Multiple linear regression models were developed using different feature selection methods and their fitness was evaluated using the adjusted R squared value. Multiple linear regression models take the form:

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

The number of eye tracking predictors considered for statistical modelling was 99 of Tobii software produced metrics, in addition to the two of our own. The 99 Tobii metrics are derived from eye gaze data from each of the 10 vital signs screen components/AOIs shown in Fig. 2 (the non-AOI measurement makes 11 in total) with the 9-metrics detailed in Table III (9 x 11 = 99 metrics). The statistical model was evaluated using leave one out validation (k folds = n observations) and we report the mean square error. Means and standard deviations (SDs) were presented as mean ±SD. All analysis presented was carried out using the R[5] programming language through the R Studio IDE[6].

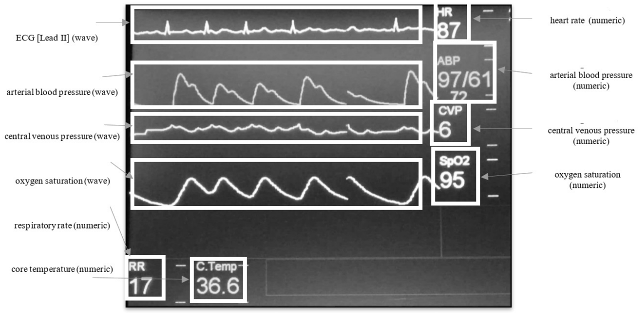
ECG [Lead II] (wave)
arterial blood pressure (wave)
central venous pressure (wave)
oxygen saturation (wave)
respiratory rate (numeric)
core temperature (numeric)
heart rate (numeric)
arterial blood pressure (numeric)
central venous pressure (numeric)
oxygen saturation (numeric)
HR
87
ABP
97/61
72
CVP
6
SpO2
95
RR
17
C.Temp
36.6

Fig. 2. Simulated Vital Signs Scenario – Areas of Interest used in analysis.

TABLE III. EXPLANATION OF EYE TRACKING METRICS PROVIDED BY TOBII SOFTWARE

Eye Tracking Metric	Definition	Insight [39]
Time to First Fixation (s)	How long it takes before a participant fixates on an AOI for the first time.	Indicates the search/scan path (e.g. top to bottom) of the participant and generally the shorter time taken, the higher efficiency of finding the AOI.
Fixations Before	Number of times the participant fixates on the media before fixating on an AOI for the first time.	N/A.
First Fixation Duration (s)	Duration of the first fixation on an AOI.	Usually reflects the time taken for recognition and identification of an object.
Fixation Duration (s)	Mean duration of individual fixations within an AOI.	Generally longer fixations equal deeper and effortful processing.
Total Fixation Duration (s)	Sum of the duration for all fixations within an AOI.	
Fixation Count	Number of times the participant fixates on an AOI.	Significantly more fixations land on semantically informative areas.
Visit Duration (s)	Mean duration of individual visits within an AOI.	Insight from this appears to depend on the semantics of the object and the task of participant. Usually sensitive to slow and long-term cognitive processes.
Total Visit Duration (s)	Duration of all visits within an AOI.	
Visit Count	Number of visits within an AOI.	Sensitive to semantic informativeness.

[5] http://www.r-project.org

[6] http://www.rstudio.com

Results

A. Interpretation Performance Summary

Individual interpretations (1 interpretation, 1 observation) were scored and categorised according to criteria (see Tables I & II) and the results are shown in Table IV.

TABLE IV. BREAKDOWN OF PERFORMANCE BY GROUPS AND OVERALL

Performance Level	*Training*	*Qualified*	All
Low-Level (0-5)	129 (72%)	10 (18%)	139 (59%)
Mid-Level (6-7)	45 (25%)	25 (45%)	70 (30%)
High-Level (8-10)	6 (3%)	20 (36%)	26 (11%)

Mean *Performance Score* for all (n = 235) observations = 5.12 ±1.85 with full range of scores recorded (0-10). The *Qualified* group scored significantly higher than the *Training* group (6.85 ±1.5 [n = 55] vs. 4.59 ±1.61 [n = 180], p = <0.0001).

B. Reported Confidence Summary

Mean *Scenario Confidence* for all (n = 235) observations = 6.19 ±1.53. The *Qualified* group reported significantly higher *Scenario Confidence* for interpretations than the *Training* group (7.51 ±1.2 vs. 5.79 ±1.39, p = <0.0001) and a weak but statistically significant correlation to *Performance Score* was found for the *Training* group but not for the *Qualified* group (*r*=0.32, p = <0.0001 vs. *r* = 0.21, p = 0.13).

C. Heart Rate Monitoring Summary

A total of 36 participants (25 training group, 11 qualified group) wore the E4 wristband during interpretations and analysis looks at the heart rate values during the entire performance – all five scenarios, instead of the separate scenario interpretations. Mean ΔHR= +4.10±19.00 bps, with a significant increase found among the *Training* group compared with the *Qualified* group (+9.08±20.15 vs. -3.90±22.33, p=<0.001) but no statistical significance was found in its correlation to *Performance Score* (r=-0.06, p=0.45).

D. NASA-TLX Survey Summary

32 participants (21 training group, 11 qualified group) responded to the *NASA-TLX*. Mean total *NASA-TLX Score* = 248.13±81.35 (41.4% total cognitive load) for all participants with no significant difference between the *Training* and *Qualified* groups (244.76±92.19 vs. 254.55±55.25, p = 0.4) and no statistical significance was found in its correlation to *Performance Score* (r=-0.04, p=0.65).

E. Eye Tracking Metrics Correlating with Performance Score

In total, only 13 eye tracking metrics out of 101 provided a statistically significant correlation to *Performance Score* shown in Table V. However not one of these metrics is a strong or even moderately strong correlation to *Performance Score*. The strongest correlation found was the measured visit count on the *Arterial Blood Pressure (Wave)* (r=0.28, p=<0.01). The highest count of correlations was metrics for any non-AOI: *Total Visit Duration*, *Total Fixation Duration*, *Fixation Count* and *Visit Count*.

TABLE V. SIGNIFICANT CORRELATIONS FOR AOI METRICS TO PERFORMANCE SCORE

AOI	Eye Tracking Metric	*r*	P =
Central Venous Pressure (Wave)	*Fixations Before*	0.15	0.04
	Visit Count	0.26	<0.01
Temperature (Numeric)	*Fixations Before*	0.15	0.03
	Visit Count	0.16	0.02
Any Non-AOI	*First Fixation Duration*	-0.17	0.02
	Fixation Count	0.20	<0.01
	Total Visit Duration	0.16	0.03
	Visit Count	0.28	<0.01
ECG (Wave)	*Fixation Count*	0.17	0.01
	Visit Count	0.21	<0.01
Arterial Blood Pressure (Wave)	*Visit Count*	0.29	<0.01
Heart Rate (Numeric)	*Visit Count*	0.15	0.04
Respiratory Rate (Numeric)	*Visit Count*	0.17	0.02

F. Discriminating between Performance Level using Eye Tracking Metrics

Participant eye tracking metrics were selected by *Performance Level* categories (low/mid/high) and significance testing (*t*-test) is performed on all areas of interest to find those metrics that best discriminate between performance level. The most notable results (using standard α = 0.05), are shown in Table VI. Due to the number of hypothesis tests (101 metrics), an adjustment is required for statistical significance. Thus, with the Bonferroni correction applied (α = 0.0005), no metrics were statistically significant. The two most notable results between low and high *Performance Level* have been presented in Fig. 4. It is likely that individual metrics on each AOI are not enough by themselves to discriminate. However, the combination of variables is more effective in predicting *Performance Score* or *Performance Level*. Supplementary visualisation of the data comparison between performance levels are provided by visiting the web addresses given in Fig. 3.

For a more complete look at the differences between the performance levels, two supplementary figures can be accessed online. These contain boxplot comparisons between the three levels for: (1) *First Fixation Duration*, *Total Fixation Duration* and *Total Visit Duration* at http://tinyurl.com/z888ya6. (2) *Fixation Count*, *Visit Duration* and *Visit Count* at http://tinyurl.com/gtpused.

The stars in these figures represent statistical significance like mentioned in Table V (note: stars placed centrally/below 'med' represent the statistical significance between low and high). The x-axis contains the levels (e.g. low, high) and the y-axis the eye tracking metric on an AOI – including the measured non-areas of interest. Metrics have been abbreviated in these figures (e.g. time to first fixation = ttff, arterial blood pressure (numeric) = abp.num).

Fig. 3. Supplementary dataset visualisations

TABLE VI. MOST NOTABLE DIFFERENCES BETWEEN METRICS COMPARED BETWEEN THE PERFORMANCE LEVELS. WITH BONFERRONI CORRECTION APPLIED, α = 0.0005, NONE ARE STATISTICALLY SIGNIFICANT.

AOI	Eye Tracking Metric	Test	P =
Heart Rate (Numeric)	*Time to First Fixation*	Low vs. Mid	0.03
	First Fixation Duration	Mid vs. High	0.04
Oxygen Saturation (Numeric)	*Time to First Fixation*	Mid vs. High	0.03
	Fixations Before	Mid vs. High	0.02
Any Non-AOI	*First Fixation Duration*	Low vs. High	0.005
	Fixation Count	Low vs. Mid	0.03
	Visit Count	Low vs. Mid	0.003
Respiratory Rate (Numeric)	*Total Fixation Duration*	Mid vs. High	0.005
	Visit Duration	Mid vs. High	0.04
	Visit Duration	Low vs. High	0.03
	Total Visit Duration	Mid vs. High	0.006
Arterial Blood Pressure (Numeric)	*Visit Duration*	Low vs. Mid	0.02
Temperature (Numeric)	*Visit Duration*	Mid vs. High	0.03
	Visit Duration	Low vs. High	0.006
	Visit Count	Low vs. Mid	0.03
Central Venous Pressure (Numeric)	*Total Visit Duration*	Mid vs. High	0.04
	Total Visit Duration	Low vs. High	0.02
Arterial Blood Pressure (Wave)	*Visit Count*	Low vs. Mid	0.002
Central Venous Pressure (Wave)	*Visit Count*	Low vs. Mid	0.004
ECG (Wave)	*Visit Count*	Low vs. Mid	0.02

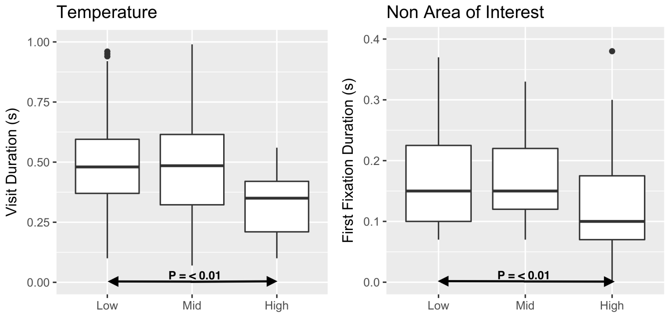

Fig. 4. Boxplot Comparisons: *Visit Duration* (s) for *Temperature* and *First Fixation Duration* (s) for Non-AOI.

G. Prediction of Performance Score Using Eye Tracking Metrics

A multiple linear regression model was built using eye tracking metrics alone to predict the *Performance Score* (0-10) of a participant. After initial experimenting, it was decided to only use observations that had complete data records (no missing data) for attempts to predict performance score. This reduced the total number of observations for training data to 112 observations. This included 66 low-level, 37 mid-level and 9 high-level performances – not significantly different from the full dataset (n=235) performance level breakdown seen in Table III. The range of scores in the training dataset used ranged from 2/10 to 9/10. There is no direct background to this type of research study and we began with rudimentary methods for selecting independent variables to predict the dependent variable. The most elementary method is the entry method [40] (entering all available predictors). When attempted there was not a significant or a strong explanatory linear model found (adjusted R-squared=0.29, p=0.27). The next step involved removal of the least significant IV (by highest p-value) and then re-calculating the model. Repeating this process until satisfied is a method known as backwards elimination [40]. We continued this step by step until only statistically significant IVs remained in the model, with a significant model overall (adjusted R-squared=0.80, p=<0.01). The summary results of this final regression model are shown in Table VI. The R code and model summary which details the predictors (both those included and those removed during the selection process) can be viewed at **http://tinyurl.com/j8lxtew**.

TABLE VII. MODEL PREDICTING PERFORMANCE SCORE (1-10)

Statistical Model	
Model Choice	Multiple Linear Regression
Total Predictors Considered	101
Predictor Selection	Backwards Elimination
No. Model Predictors	62
Total Training Observations	112 (66 Low, 37 Mid, 9 High)
Residual Standard Error	0.75 on 49 Degrees of Freedom
Multiple R-squared	0.91
Adjusted R-squared	0.80
F-statistic	7.94
P value	3.712e-12
Cross Validation	
Validation Method	Leave One Out Validation
K Folds	112
Mean Square Error	1.2

DISCUSSION

The number of participants recruited (n=47) in this study, given the complexity of the data recording was sufficient, with a good range of scores collected (range = 0 to 10). Therefore, a full range of scores facilitated adequate analysis of eye tracking metrics for discriminating between different participant *Performance Score* and *Performance Level*.

The two groups of participants fulfilled their predicted roles of providing scores at the opposite ends of the spectrum. Most the low-level scores were collected from *Training* group and most of the mid-level and high-level scores were collected from *Qualified* group. Although the reader is reminded when reading the analysis that the participant count for the *Qualified* group was substantially lower in comparison to the *Training* group. With Δ*HR*, the results suggest that on average the *Qualified* group began to relax as they went through the task, whereas on average the *Training* group seemed to demonstrate a slight rise. The *NASA-TLX* score did not provide any insight as cognitive load did not correlate with *Performance Score*. Future work may include analysis of the individual question responses to the NASA-TLX survey. No strong correlations with any single eye tracking metric or other metric/measurement with the *Performance Score* was found. However, given the complexity of human decision-making, it is already unclear if a large body of eye tracking metrics would correlate collectively with performance – so it is unlikely that single measurements on single AOIs would provide insight. Instead we see a collection of eye tracking metrics that discriminate between performance levels.

A. Eye Tracking Metrics to Discriminate

The most discriminating eye tracking metrics are those that measure visual attention taking place away from on-screen vital signs (i.e. non-AOI). An example of this is that *First Fixation Duration* on a non-AOI, seen in Fig. 4, decreases as the *Performance Level* rises from low to high. Table III tells us that *First Fixation Duration* is generally indicative of identification or recognition of objects. High performers are measurably fixating less with their initial fixation on non-important objects – something that might seem insignificant but does discriminate between low and high performances. If we view the supplementary plots provided by Fig. 3, we see that no other AOI discriminates this way between low and high performers. We could speculate that a high performer is more capable, intentionally or not, of disregarding visual distraction over the key objects that they need to retrieve information for the task given. Fig. 5 also shows us that the mean *Visit Duration* (described in Table II as rising when task objects have longer cognitive demand) on *Temperature* was lower with high performers compared with low performers. The supplementary plots provided by the second url link in Fig. 3 show us that only *Respiratory Rate* shared this discriminatory metric between low and high performers. These two AOIs share a common position on the vital signs screen – bottom of the picture (see Fig. 2) and don't require a waveform. We could speculate that high performers do not require a long duration to recognize these vital signs and retrieve their value. The four AOIs that discriminate between low and high scores are: (1) *Respiratory Rate*, (2) *Central Venous Pressure* (numeric), (3) *Temperature* and (4) any non-AOI. Following on from this, we were then able to use a large amount of the 101 eye tracking metrics to produce a statistically significant multiple linear regression model. 62 eye tracking metrics were selected through backward elimination and then used to produce a predictive model. We can reduce the predictors using further statistical techniques on the same dataset in future. With more interpretations from a good distribution of *Performance Level* and a reduction of the dimensionality regarding the number of predictors, we could possibly develop an improved predictive model. Even so, this result is promising and shows that eye tracking metrics are connected to the performance (knowledge, skills and decision-making) with nurses performing this task. This has several consequences if researched further.

B. Visual Attention and Simulation-Based Training

One of the most valuable uses of eye tracking metrics, that prove to have predictive ability into *Performance Level*, would be in a computer-based simulator like those already being evaluated in healthcare [22], [36], [41], [42]. This would allow for automatically classifying a nurse's performance. With more affordable and unobtrusive eye tracking technology being developed each year, it is likely to become ubiquitous in our everyday lives. It may be included within various technological devices we use (e.g. smartphones, tablet devices). As a result, the visual attention of a user would become rudimentary input for many screen-based devices and could therefore become a valid assessment tool for simulation-based training tasks like the simulated scenarios in this study.

C. Visual Attention and Patient Safety

A potential opportunity to use eye tracking metrics is using them as a feature of patient safety on vital signs monitors. It could become a real-time measurement of performance when interpreting vital signs, essentially allowing the monitor itself to become sensitive to the viewer's level of uncertainty. They could then use these measurements to alert supervisors (or ideally a dedicated taskforce to monitor healthcare practitioner performance). Any nurses that appear to be struggling with the task of patient monitoring would be highlighted for immediate assistance from a supervising nurse and could be supported by further education and training in this aspect of their practice. It could offer targeted appraisal and professional development opportunities through the provision of real-time monitoring of nursing performance to augment those other markers that are currently used to detect failing competency with these skills.

D. Limitations

The design of the study and how we perform analysis to draw conclusions has a limitation, as participants do not just interpret the scenarios based on the vital signs portion. It must be acknowledged that the vignette primes their interpretation (before they even set eyes on the vital signs) and some participants may have a rule-based process that they already have set in motion regardless of what the vital signs show. Participant numbers were adequate to provide exploratory data analysis however more subjects would allow for more accurate results and perhaps more statistically significant eye tracking metrics. More participants could uncover additional patterns that exist within two distinct groups or the performance levels we define here. The number of participants' who demonstrated high-level performance could ideally be more numerous but as we see, not all the *Qualified* group provided high-level performances despite their superior experience and expertise in comparison to the *Training* group. As a result, the dataset has many more low-level performances than the other two levels of performance combined. This limits our ability to find all distinct patterns that exist in the top performers at this task and we may not therefore, be able to make any recommendations for optimal performance or present 'expert level' metrics to be used as assessment criteria. As noted in Methods and Materials, we excluded recordings that had less than 50% recording quality and this could arguably be lower. This is an automatic measurement by the Tobii hardware and software. It quantifies eye gaze activity recorded on screen against the time spent during the eye tracking session. We used a conservative threshold of 50% in recording quality as a cut-off point – to allow enough observations into the pool of analysis while removing some (likely) poor recordings. The initial judgement and assumption was that if the participant is looking away from the screen more than 50% of the time that it would not be as insightful when analyzed is potentially not true. As when these lower recording quality recordings have been manually replayed in review, it's clear that a lot do include eye gaze data and importantly, high-level performances. It could be that eye tracking recording quality is not as critical as first thought for this task – given the real task would likely not see nurses watching the screen without altering their visual attention to elsewhere in the room or ward. One further factor for low measured recording quality is a technical/hardware fault. This is when you see recording sessions with a measurement of anything from 0% (definite fault) to perhaps as high as 30%. To confirm this, we can replay the recordings (as we did for assessment). If no eye gaze activity is shown on the vital signs portion, despite clear evidence that the participant is responding verbally to the information on screen, then you can categorise that performance as having a technical failure. In our review of recordings, 2 of the 47 participant recordings were categorised as such. The eye tracking metrics themselves are limited to the temporal form – which is the statistics provided by the proprietary software. Other eye tracking solutions provide spatial metrics as well, for example the dispersion of fixations or the scan path length in pixels. These could provide potentially different forms of insight to these tasks that we are unable to gather without them. Another potential issue is regarding the assessment method which asked the participant to concurrently think aloud. This is a traditional and commonly used method and widely accepted [43]. It receives some criticism as verbal processing is thought to require a certain level of attention and can cause distraction to the participant [44]. It's also known that a common side effect of a research coordinator being present in the room is that they will talk to the coordinator whilst explaining themselves and can lose eye tracking data (as we technically did) [45]. However, the other two options aren't that suitable despite the drawbacks for the concurrent method mentioned. (1) Retrospective think aloud, which would involve the participant saying nothing and then providing their interpretation as they watch their eye tracking recording post-viewing or a hybrid between the two. We take the view that the concurrent talk aloud method is the most like for like scenario for what a nurse would do in the real-life event. They might not verbally speak but they wouldn't necessarily wait a certain time before deciding – which is what we would be doing with retrospective think aloud. It also could be argued that if we wish to detect any discrete differences between high performers and low performers at the task, it must be in real time. One argument that could be made is that some qualified participants might have habits of thinking quietly (without verbally stating anything) aloud during their thought process and so might have kept components of their interpretation from being captured by ourselves. This could have had a limitation for high performances in the qualified group and may explain some mid-level and low-level performances.

(2) The hybrid method, which allows them to think aloud concurrently but gives them an opportunity to provide their thoughts retrospectively is superior is some scenarios, especially in usability but again, it's not necessarily appropriate for this task we assess in the study. One final limitation is the lack of the vignette (on-screen) during the vital signs interpretation. However, the vignettes represented quite short cases, and a key observation is that no participant requested to see the vignette again – instead relying on their memory (verbally reminding themselves) when needed. Statistically, some limitations must be acknowledged. Firstly, the number of significance tests performed in Table VI presents a problem of Type I error inflation. The Bonferroni Correction[7] has been applied to adjust the alpha value for inferring statistical significance (101 metrics: $\alpha = 0.0005$) and is acknowledged/presented in the results. However, arguments have been made against applying corrections in these scenarios [46] – that it is too conservative, increases risk of Type II errors (which may be a more at risk scenario depending on the experiment/hypothesis) and more generally, that providing transparency to the reader of the statistical methods used (allowing them to make a clear conclusion) is more important than providing a correction for significance testing. Also for the statistical model there is a potential problem if attempting to interpret the model [47]. Due to the nature of the derived eye tracking metrics, which are continuous values that are made from 9 metrics on 10 areas of interest and 1 category for non-areas of interest, many of the independent variables would not be considered truly independent (problem of collinearity). However, the model presented in the results is not to be interpreted, it is instead (as stated) an attempt to show the value of eye tracking metrics in predicting *Performance Score* (0-10) of the participant.

E. Future Work

Eye tracking data collection and analysis performed in this study could be expanded to include an array of ward objects, including a bed with either an actor or manikin like a similar recent study [36]. Using mobile eye tracking solutions (glasses) to capture not only the verbal output but the actions taken by nurses would be a more complete study to undertake. The use of mobile tracking is becoming feasible from a technological and affordability perspective and could indeed enhance the state of the art in human computer interaction. Further statistical analysis will be performed on the dataset that we have collected in the study. This includes principal component analysis to reduce the dimensionality on both the entire feature set collected from all scenarios and the individual scenarios interpreted. Qualitative analysis will also take place in the form of thematic analysis on the verbal recordings of each participant. Future work can also include using more machine learning classification algorithms such as decision trees, deep learning, neural networks, support vector machines, k-nearest neighbor along with better evaluation of these models using 10-fold cross validation. We could also increase the size of the subset of observations (interpretations) by removing the recording quality criteria and simply relying on patterns to persist despite missing eye tracking data. Another opportunity

[7] https://en.wikipedia.org/wiki/Bonferroni_correction

is to analyse the potential differences between performance levels within shorter time windows. Specifically, to identify if there are clearer discriminatory metrics when analysing only the first 5s, 15s or 30s instead of the full 60s they were allowed. This may reduce the amount of 'noise' in the data collected towards the end of the scenario interpretation. For example, when quite a few high-level performers reached the 30-45s mark, they were finished with their interpretation – the remaining time may in fact be washing out some of their distinguishing behaviour.

CONCLUSION

The study conducts the first ever capturing of visual attention measurements, from a set of nurses (with varying experience and expertise) while reading and interpreting from a simulated vital signs monitor. The data collected, specifically the data analysis of eye tracking metrics, has shown that visual attention and the *Performance Level* for this specific task (measured by *Performance Score*) are not independent of each other. Put differently, eye tracking metrics exclusively could be used to predict a person's performance when reading vital signs. Further research, including statistical techniques like principal component analysis are required to refine the regression models and the optimal level of accuracy for predicting *Performance Score* and then *Performance Level* using the eye tracking metrics. At present, we can conclude that there is a relationship between eye tracking metrics and performance that can be seen but it is unclear to what extent and what reliable accuracy they can predict performances.

REFERENCES

[1] L. Kohn, J. Corrigan, and M. Donaldson, *To err is human: building a safer health system*. 2000.

[2] J. Zhang, V. L. Patel, and T. R. Johnson, "Medical error: is the solution medical or cognitive?," *J. Am. Med. Inform. Assoc.*, vol. 9, no. 6 Suppl, pp. S75-7, 2002.

[3] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer, "The role of deliberate practice in the acquisition of expert performance.," *Psychol. Rev.*, vol. 100, no. 3, pp. 363–406, 1993.

[4] A. Baddeley, "Working Memory Alan Baddeley," *Science (80-.).*, vol. 255, no. 5044, pp. 556–559, 1992.

[5] R. L. Angelo, R. K. N. Ryu, R. A. Pedowitz, W. Beach, J. Burns, J. Dodds, L. Field, M. Getelman, R. Hobgood, L. McIntyre, and A. G. Gallagher, "A Proficiency-Based Progression Training Curriculum Coupled With a Model Simulator Results in the Acquisition of a Superior Arthroscopic Bankart Skill??Set," *Arthroscopy*, vol. 31, no. 10, pp. 1854–1871, 2015.

[6] R. A. Pedowitz, G. T. Nicandri, R. L. Angelo, R. K. N. Ryu, and A. G. Gallagher, "Objective Assessment of Knot-Tying Proficiency With the Fundamentals of Arthroscopic Surgery Training Program Workstation and Knot Tester," *Arthroscopy*, vol. 31, no. 10, pp. 1872–1879, 2015.

[7] P. C. Smith and B. K. Hamilton, "The effects of virtual reality simulation as a teaching strategy for skills preparation in nursing students," *Clin. Simul. Nurs.*, vol. 11, no. 1, pp. 52–58, 2015.

[8] R. P. Cant and S. J. Cooper, "Simulation in the Internet age: The place of Web-based simulation in nursing education: An integrative review," *Nurse Educ. Today*, vol. 34, no. 12, pp. 1435–1442, 2014.

[9] J. Persson, E. H. Dalholm, M. Wallergård, and G. Johansson, "Evaluating interactive computer-based scenarios designed for learning medical technology," *Nurse Educ. Pract.*, vol. 14, no. 6, pp. 579–585, 2014.

[10] A. Sliney and D. Murphy, "JDoc: A serious game for medical learning," *Proc. 1st Int. Conf. Adv. Comput. Interact. ACHI 2008*, pp. 131–136, 2008.

[11] A. I. Levine, S. DeMaria, A. D. Schwartz, and A. J. Sim, Eds., *The Comprehensive Textbook of Healthcare Simulation*. New York, NY: Springer New York, 2013.

[12] J. Hogan, "Why don't nurses monitor the respiratory rates of patients?," *Br. J. Nurs.*, vol. 15, no. 9, pp. 489–492, May 2006.

[13] M. Elliott and A. Coventry, "Critical care: the eight vital signs of patient monitoring," *Br. J. Nurs.*, vol. 21, no. 10, pp. 621–625, 2012.

[14] I. Wheatley, "The nursing practice of taking level 1 patient observations," *Intensive Crit. Care Nurs.*, vol. 22, no. 2, pp. 115–121, 2006.

[15] A. E. Rogers, G. E. Dean, W.-T. Hwang, and L. D. Scott, "Role of registered nurses in error prevention, discovery and correction.," *Qual. Saf. Health Care*, vol. 17, no. 2, pp. 117–121, 2008.

[16] W. Q. Mok, W. Wang, and S. Y. Liaw, "Vital signs monitoring to detect patient deterioration: An integrative literature review," *Int. J. Nurs. Pract.*, vol. 21, no. S2, pp. 91–98, 2015.

[17] T. Ahrens, "The most important vital signs are not being measured," *Australian Critical Care*, vol. 21, no. 1, pp. 3–5, 2008.

[18] D. B. Chalfin, S. Trzeciak, A. Likourezos, B. M. Baumann, and R. P. Dellinger, "Impact of delayed transfer of critically ill patients from the emergency department to the intensive care unit," *Crit Care Med*, vol. 35, no. 6, pp. 1477–1483, 2007.

[19] S. Osborne, C. Douglas, C. Reid, L. Jones, and G. Gardner, "The primacy of vital signs - Acute care nurses' and midwives' use of physical assessment skills: A cross sectional study," *Int. J. Nurs. Stud.*, vol. 52, no. 5, pp. 951–962, 2015.

[20] M. A. F. Pimentel, D. A. Clifton, L. Clifton, P. J. Watkinson, and L. Tarassenko, "Modelling physiological deterioration in post-operative patient vital-sign data," *Med. Biol. Eng. Comput.*, vol. 51, no. 8, pp. 869–877, 2013.

[21] F. E. Bogossian, S. J. Cooper, R. Cant, J. Porter, H. Forbes, L. McKenna, L. Kinsmen, R. Endacott, B. Devries, N. M. Philips, T. Bucknall, S. Young, and V. Kain, "A trial of e-simulation of sudden patient deterioration (FIRST2ACT WEB???) on student learning," *Nurse Educ. Today*, vol. 35, no. 10, pp. e36–e42, 2015.

[22] P. Bradley, "The history of simulation in medical education and possible future directions," *Med. Educ.*, vol. 40, no. 3, pp. 254–262, 2006.

[23] S. Cooper, A. Beauchamp, F. Bogossian, T. Bucknall, R. Cant, B. Devries, R. Endacott, H. Forbes, R. Hill, L. Kinsman, V. J. Kain, L. McKenna, J. Porter, N. Phillips, and S. Young, *Managing patient deterioration: a protocol for enhancing undergraduate nursing students' competence through web-based simulation and feedback techniques.*, vol. 11, no. 1. 2012.

[24] M. Aebersold, D. Tschannen, M. Stephens, P. Anderson, and X. Lei, "Second Life??: A New Strategy in Educating Nursing Students," *Clin. Simul. Nurs.*, vol. 8, no. 9, pp. e469–e475, 2012.

[25] C. A. Rhodes, D. Grimm, K. Kerber, C. Bradas, B. Halliday, S. McClendon, J. Medas, T. P. Noeller, and M. McNett, "Evaluation of Nurse-Specific and Multidisciplinary Simulation for Nurse Residency Programs," *Clin. Simul. Nurs.*, vol. 12, no. 7, pp. 243–250, 2016.

[26] J. Green, A. Wyllie, and D. Jackson, "Virtual worlds: A new frontier for nurse education?," *Collegian*, vol. 21, no. 2, pp. 135–141, 2014.

[27] B. Mariani and J. Doolen, "Nursing Simulation Research: What Are the Perceived Gaps?," *Clin. Simul. Nurs.*, vol. 12, no. 1, pp. 30–36, 2016.

[28] A. Fong, D. J. Hoffman, A. Zachary Hettinger, R. J. Fairbanks, and A. M. Bisantz, "Identifying visual search patterns in eye gaze data; gaining insights into physician visual workflow," *J. Am. Med. Informatics Assoc.*, p. ocv196, 2016.

[29] B. Zheng, G. Tien, S. M. Atkins, C. Swindells, H. Tanin, A. Meneghetti, K. A. Qayumi, and O. N. M. Panton, "Surgeon's vigilance in the operating room," *Am. J. Surg.*, vol. 201, no. 5, pp. 673–677, May 2011.

[30] S. Zhou, R. Gali, M. Paasche-Orlow, and T. W. Bickmore, "Afraid to ask: proactive assistance with healthcare documents using eye tracking," *Proc. Ext. Abstr. 32nd Annu. ACM Conf. Hum. factors Comput. Syst. - CHI EA '14*, pp. 1669–1674, 2014.

[31] R. R. Bond, D. D. Finlay, C. Breen, K. Boyd, C. D. Nugent, N. D. Black, P. W. Macfarlane, and D. Guldenring, "Eye Tracking in the Assessment of Electrocardiogram Interpretation Techniques University of Ulster , Jordanstown , Northern Ireland , United Kingdom University of Glasgow , Glasgow , Scotland , United

Kingdom Eye tracking system," no. Figure 1, pp. 2–5, 2012.

[32] P. O'Meara, G. Munro, B. Williams, S. Cooper, F. Bogossian, L. Ross, L. Sparkes, M. Browning, and M. McClounan, "Developing situation awareness amongst nursing and paramedicine students utilizing eye tracking technology and video debriefing techniques: A proof of concept paper," *Int. Emerg. Nurs.*, vol. 23, no. 2, pp. 94–99, 2015.

[33] M. a Just and P. a Carpenter, "A theory of reading: From eye fixations to comprehension.," *Psychol. Rev.*, vol. 87, no. 4, pp. 329–354, 1980.

[34] M. P. Stiegler and D. M. Gaba, "Eye Tracking to Acquire Insight Into the Cognitive Processes of Clinicians," *Simul. Healthc. J. Soc. Simul. Healthc.*, vol. 10, no. 5, pp. 329–330, 2015.

[35] O. Asan and Y. Yang, "Using Eye Trackers for Usability Evaluation of Health Information Technology: A Systematic Literature Review," *JMIR Hum. Factors*, vol. 2, no. 1, p. e5, 2015.

[36] N. Suetsugu, M. Ohki, and T. Kaku, "Quantitative Analysis of Nursing Observation Employing a Portable Eye-Tracker," *Open J. Nurs.*, vol. 6, no. 1, pp. 53–61, 2016.

[37] NMC, "NMC Standards for pre-registration nursing education," 2010.

[38] Resuscitation Council UK, "ABCDE approach," *ABCDE approach*, 2016. [Online]. Available: https://www.resus.org.uk/resuscitation-guidelines/abcde-approach/. [Accessed: 19-Apr-2017].

[39] K. Holmqvist, M. Nyström, R. Andersson, and R. Dewhurst, *Eye tracking: A comprehensive guide to methods and measures*. 2011.

[40] L. Kuo and B. Mallick, "Variable Selection for Regression Models," *Sankhyā Indian J. Stat. Ser. B*, vol. 60, no. 1, pp. 65–81, 1998.

[41] R. R. Bond, D. D. Finlay, C. D. Nugent, G. Moore, and D. Guldenring, "A simulation tool for visualizing and studying the effects of electrode misplacement on the 12-lead electrocardiogram," *J. Electrocardiol.*, vol. 44, no. 4, pp. 439–444, 2011.

[42] W. M. Nehring and F. R. Lashley, "Nursing simulation: A review of the past 40 years," *Simul. Gaming*, vol. 40, no. 4, pp. 528–552, 2009.

[43] S. Elling, L. Lentz, and M. de Jong, "Retrospective think-aloud method: Using eye movements as an extra cue for participants' verbalizations," *Proc. 29th SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 1161–1170, 2011.

[44] Z. Guan, S. Lee, E. Cuddihy, and J. Ramey, "The validity of the stimulated retrospective think-aloud method as measured by eye tracking," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '06*, p. 1253, 2006.

[45] J. B. Bavelas, L. Coates, and T. Johnson, "Listener Responses as a Collaborative Process: the Role of eye Gaze," *J. Commun.*, vol. 52, no. September, pp. 566–580, 2002.

[46] T. V Perneger, "What's wrong with Bonferroni adjustments.," *BMJ*, vol. 316, no. 7139, pp. 1236–8, Apr. 1998.

[47] R. I. Kabacoff, "Regression Diagnostics," in *R in Action: Data Analysis and Graphics with R*, 2nd ed., Manning, 2015, p. 193.

Jonathan Currie obtained his BSc in 2014, in the School of Computing and Mathematics at Ulster University. He is currently completing his PhD research by the end of 2017. This is within the broad area of computer-based simulation for medical training – specifically the value of eye tracking analysis for automated assessment.

Raymond Bond obtained his BSc and PhD in the School of Computing and Mathematics (Ulster University). Raymond has research interests within the broad area of biomedical and health informatics. He has research interests in simulation-based training for medicine, usability engineering methods to improve medical devices (which include eye tracking and other psychophysiology metrics).

Paul McCullagh received a BSc (1979) and a PhD (1983) in Electrical Engineering from Queen's University of Belfast. He is a Reader in the School of Computing & Mathematics at the University of Ulster. He has worked on EU FP7 BRAIN project for e-inclusion using the brain-computer interface, EPSRC SMART 2 for self management of chronic disease, TSB NOCTURNAL for assisting people with dementia, and ESRC New Dynamics of Ageing, Design for Ageing Well funded projects.

Pauline Black graduated from the University of Ulster in 1994 with a first class honours degree in Nursing. She completed a Post Graduate Diploma (with distinction) in 1997 and a PhD in 2005. She is currently a Lecturer at the School of Nursing. Prior to this, she worked for eight years as a registered nurse in Critical Care in the Belfast Health and Social Care Trust.

Dewar Finlay is a Reader in Electronic Engineering in the School of Engineering and a member of the Engineering Research Institute at Ulster University. Dewar's main area of research interest is in healthcare technology. He is a member of the Editorial Board of the Journal of Electrocardiology. Dewar holds a BEng Degree in Electronic Systems and a PhD in Computing.

Aaron Peace is a graduate of Queen's University, Belfast. After working in Royal Northshore, Sydney he undertook his Cardiology training on the Irish Higher Medical Training Programme for Cardiology, completing his basic training in Cardiology in 2006. He took up the position of consultant cardiologist at Altnagelvin Hospital, Derry in 2013. He was appointed as Assistant Medical Director, Research & Development and Chief Executive Officer of C-TRIC (Clinical Translational Research and Innovation Centre) in 2016.