# Consciousness assessment in prehospital deterioration risk stratification

## A foundation for automated AVPU evaluation

Filipa Maria Carmo Barros

Supervisor:

João Carlos Amaro Ferreira, Assistant Professor with aggregation

ISCTE - Instituto Universitário de Lisboa

Date: 27th of February 2026

Co-funded by the European Union

# Consciousness assessment in prehospital deterioration risk stratification

## A foundation for automated AVPU evaluation

Filipa Maria Carmo Barros

Supervisor:

João Carlos Amaro Ferreira, Assistant Professor with aggregation

ISCTE - Instituto Universitário de Lisboa

Date: 27th of February 2026

Co-funded by the European Union

**Dedication / Acknowledgements**

This thesis represents a journey that extended well beyond my clinical comfort zone. As a medical doctor, my professional training is rooted in direct patient care and clinical decision-making; engaging in a project centered on artificial intelligence, computer vision, and data modelling required me to step into unfamiliar territory. For that reason, I am deeply grateful to those who supported and challenged me throughout this process.

I would like to sincerely thank Pedro Faria for his encouragement and for contributing thoughtful discussions that helped refine the practical and translational dimensions of this work. His perspective and support were important in bridging clinical reasoning with technological implementation.

I extend my profound gratitude to Professor João Ferreira for believing in and supporting a research topic situated in a demanding and relatively unexplored domain, marked by limited prior academic and industrial development and restricted access to real-world data. His guidance was essential in structuring the scope of this thesis, identifying adjacent knowledge domains, and fostering collaboration among students with complementary expertise. His clear direction, constructive feedback, and consistent availability were instrumental in ensuring progress under significant time and logistical constraints.

I am also especially thankful to my colleagues who embraced the practical challenge of translating a clinical problem into technological experimentation. Marco Esperança accepted the challenge of testing computer vision–based AI models, exploring how observable behavioural cues could be algorithmically interpreted. Tomás Mendes contributed through pattern data analysis, applying analytical methodologies to structure and interpret physiological and behavioural datasets. Their willingness to engage with this interdisciplinary problem was crucial in transforming a clinical intuition into a structured, testable framework.

Finally, I extend my appreciation to ISCTE for the opportunity to participate in the ManagiDiTH Summer Week as an external industry representative. Presenting and discussing the AVPU challenge with a broader audience, particularly professionals with strong medical backgrounds, provided valuable validation of the proposed use cases and reinforced the relevance of this work for both clinical practice and future healthcare innovation.

**Abstract**

Early recognition of clinical deterioration in prehospital settings is essential to ensure timely escalation, reduce adverse events, and optimize emergency medical resources. The National Early Warning Score (NEWS) is a widely adopted framework to support deterioration risk stratification based on routinely collected physiological parameters.

Within NEWS, level of consciousness, commonly assessed using the AVPU scale (Alert, Verbal, Pain, Unresponsive), has a high impact because any deviation from alertness carries a strong escalation weight. Nonetheless, AVPU assessment is observational, influenced by operational constraints, particularly in the dynamic context of Emergency Medical Services (EMS).

This dissertation presents a proof of concept designed to evaluate the role of AVPU in prehospital deterioration risk stratification, using NEWS as the reference clinical framework. Two complementary approaches were developed. A video-based methodology defined through clinically grounded scenarios with model patient interactions. The videos were created and annotated under medical rationale to enable assessment of computer vision methods for recognizing responsiveness patterns aligned with AVPU categories. Also, a sensor-data methodology was developed to analyse physiological measurements, their association with consciousness and the NEWS escalation thresholds.

Building on these complementary models, the dissertation proposes a modular multimodal architecture integrating video-derived responsiveness cues and physiological analytics within existing EMS workflows.

Rather than claiming full automation of consciousness assessment, this work provides a structured methodological and architectural foundation for future prospective validation and responsible integration of multimodal decision-support systems in prehospital emergency care.

Keywords

AVPU scale; NEWS Score; EMS; computer vision; clinical validation.

JEL Classification: I10, I18, C45, O31

## Index

**Dedication / Acknowledgements** ..... **II**

**Abstract** ..... **III**

**Glossary (if applicable)** ..... **VII**

**Introduction** ..... **1**

**Thesis Organization** ..... **3**

**Chapter 1** ..... **4**

1.1 Background and Rationale ..... 4

1.2 Role of level of consciousness within the NEWS framework ..... 6

1.3 Importance of risk stratification in Prehospital settings ..... 7

1.4 Scope and objectives ..... 9

**Chapter 2** ..... **10**

2.1 Chapter Overview ..... 10

2.2 Clinical Rationale ..... 10

2.2.1 Video use for face detection ..... 10

2.2.2 Drowsiness ..... 11

2.2.3 Consciousness Assessment in Healthcare ..... 12

2.2.4 Video to access consciousness ..... 13

2.3 Proposing an AVPU video-observable classifier ..... 15

2.4 Dataset Design Principles ..... 17

2.5 Controlled Environment Dataset Construction ..... 17

2.6 Scenario Specification and Scripted Interaction Design ..... 18

2.7 Ground Truth Definition ..... 19

2.8 Annotation Strategy ..... 21

2.9 Validation framework ..... 22

2.10 AI Video Pipeline and Collaboration Context ..... 23

2.11 Summary and Link to the Sensor Data-Based Approach ..... 24

**Chapter 3** ..... **25**

3.1 Overview ..... 25

3.2 Current Evidence for Vital Signs Monitoring by wearable devices ..... 26

3.3 Sensor-Derived Physiological Data and Dataset Characterization ..... 27

3.4 Vital Signs Monitoring and Clinical Risk Prediction ..... 27

3.5 Consciousness assessment in risk scores ..... 28

3.6 AVPU in the NEWS Score ..... 29
3.7 Vital signs to access Consciousness ..... 31
3.8 From Threshold-Based Scores to Data-Driven Escalation Models ..... 32
3.9 Data models and Collaboration Context ..... 32
3.10 Implications for Clinical Risk Stratification and Decision Support ..... 32
3.11 Research Gaps ..... 33
**Chapter 4 ..... 34**
4.1 Integration of both approaches ..... 34
4.2 Relevance from a clinical perspective ..... 35
4.3 System architecture and deployment in EMS ..... 36
4.4 Ethical, privacy and governance considerations ..... 41
4.6 Operational constraints and implementation Challenges ..... 42
**Chapter 5 ..... 44**
5.1 Summary of the main achievements ..... 44
5.2 Contributions to clinical practice and digital innovation ..... 45
5.3 Study limitations ..... 46
5.4 Recommendations for future research ..... 47
5.5 Conclusion ..... 49
**Bibliographic References and Sources ..... 51**
**Annexes ..... 57**
Annex A - AVPU scenario video recording (in collaboration with Pedro Faria) ..... 57
Annex B – Script for AVPU Scenario Performance (in collaboration with Pedro Faria) ..... 58
Annex C – Dataset production (in collaboration with Pedro Faria) ..... 59
Annex D - Reporting the Vignette Experiments (using GROVE) ..... 60
Annex E - Inter-rater reliability ..... 61
Annex F - Statistics outputs ..... 62

## Index of Figures

Figure 1.1- NEWS Score - Royal College of Physicians - NHS ..... 5
Figure 1.2 - AVPU scale ..... 6
Figure 1.3 - INEM workflow ..... 8
Figure 1.4 - Use of NEWS in the Portuguese EMS: iTeams® and CODU ..... 8
Figure 2.1 - Mapping response to stimulus in AVPU scale and matching NEWS score ..... 20
Figure 2.2 - Annotation framework ..... 22
Figure 2.3 - Video annotation example ..... 22
Figure 2.4 – AVPU inter-rater Agreement Confusion Matrix ..... 23
Figure 2.5- Feature extraction pipeline for AVPU video-based consciousness classification. ..... 24
Figure 3.1 - Comparison of Vitals only with vitals+ AVPU for predicting NEWS ≥ 5 ..... 30
Figure 4.1 – A modular multimodal architecture integrating physiological monitoring and event-driven video consciousness assessment within a unified workflow ..... 35
Figure 4.2 – iTeams® Dashboard adapted to NEWS score obtained from different pathways ..... 39
Figure 4.3 - iTeams® Dashboard adapted to Video AVPU ..... 40
Figure 4.4 – iTeams® Dashboard example of a flagged quality indicator ..... 40

## Index of Tables

Table 2.1 - Relating BLS, AVPU and the NEWS score ..... 16
Table 2.2 - Mapping face, head and upper limb response to stimulus ..... 20
Table B.1 - Resumed Script per Scenario ..... 58
Table C.1 –Dataset segmented by AVPU binary categories ..... 59
Table D.1- Planned annotation procedure accordingly to GROOVE guidelines. (Hillen et al., 2025) . 60
Table E.1 - AVPU Confusion Matrix (N = 112) ..... 61
Table E.2 – Observed Agreement and Cohen's Kappa ..... 61
Table F.1 - Sample descriptive analysis ..... 62
Table F.2 - Crosstabulation: AVPU × NEWS risk category ..... 62
Table F.3 - Chi-Square Tests ..... 62
Table F.4- Risk Estimates ..... 63
Table F.5 - Multivariable association, standardized physiological parameters, and NEWS ≥ 5 ..... 63
Table F.6 - Reclassification of Risk Across NEWS ≥ 5 after Inclusion of AVPU ..... 63
Table F.7 - Threshold: NEWS ≥ 5 ..... 64
Table F.8 - Threshold: NEWS ≥ 7 ..... 64

**Glossary**

**AI** – Artificial Intelligence.

**ANN** – Artificial Neural Network.

**ATLS®** – Advanced Trauma Life Support.

**AVPU** – Alert / Verbal / Pain / Unresponsive.

**ACVPU** – Alert / Confusion / Verbal / Pain / Unresponsive.

**BLS** – Basic Life Support.

**CNN** – Convolutional Neural Network.

**CODU** – Centro de Orientação de Doentes Urgentes.

**DPIA** – Data Protection Impact Assessment.

**ECG** – Electrocardiography.

**EEG** – Electroencephalography.

**EHR** – Electronic Health Record.

**EMS** – Emergency Medical Services.

**ERC** – European Resuscitation Council.

**fMRI** – Functional Magnetic Resonance Imaging.

**GDPR** – General Data Protection Regulation.

**GCS** – Glasgow Coma Scale.

**HR** – Heart Rate.

**ICU** – Intensive Care Unit.

**INEM** – Instituto Nacional de Emergência Médica.

**ML** – Machine Learning.

**NEWS** – National Early Warning Score.

**NEWS2** – Updated version of the National Early Warning Score

**NHS** – National Health Service (United Kingdom).

**PSPI** – Prkachin and Solomon Pain Intensity score.

**RR** – Respiratory Rate.

**RTS** – Revised Trauma Score.

**SBP** – Systolic Blood Pressure.

**$SpO_2$** – Peripheral Oxygen Saturation.

**TMS-EEG** – Transcranial Magnetic Stimulation combined with Electroencephalography.

## Introduction

Prehospital emergency care is characterized by rapid decision-making under uncertainty, where incomplete information, time pressure, and challenging environmental conditions can compromise the consistency of clinical assessment. (Bhaumik et al., 2022) In this setting, the early identification of patients at risk of deterioration is critical. Under-triage may delay escalation and contribute to adverse outcomes, whereas over-triage may lead to inefficient resource allocation and unnecessary burden on emergency pathways. (Smith et al., 2022)

Consequently, structured risk stratification frameworks have become increasingly important to support frontline decision-making in Emergency Medical Services (EMS). (Burgos-Esteban et al., 2022; Patel et al., 2018) Early warning scores provide a pragmatic mechanism to summarize a patient's physiological condition using routinely collected vital signs. Among these, the National Early Warning Score (NEWS) is widely used across acute care systems and has been adopted as a reference framework for escalation. (Forster et al., 2022;Bourke-Matas et al., 2024, 2025; Maciver, 2021)

NEWS aggregates multiple physiological parameters into a composite score and defines thresholds that can trigger clinical response. (Williams, 2022) However, deterioration is not exclusively represented by numerical physiological values. Neurological compromise, particularly alterations in consciousness, often constitutes a pivotal warning sign that may reflect hypoxia, hypotension, metabolic disturbance, intoxication, central nervous system pathology, or systemic shock. (Zadravecz et al., 2015)

In prehospital practice, the level of consciousness is frequently assessed through the AVPU scale, which classifies patients as Alert, responsive to Verbal stimuli, responsive to Pain stimuli, or Unresponsive. (Janagama et al., 2022) Within the NEWS framework, AVPU is treated as a strongly weighted component as any state other than full alertness produces an immediate score increase that can move a patient across escalation thresholds. (Pimentel et al., 2019) Despite this importance, AVPU assessment remains observational and may be influenced by contextual constraints such as scene safety, noise, lighting, patient positioning, movement during transport, competing tasks, and documentation burden. These constraints motivate the exploration of complementary computational approaches that can support, standardize, or further explain consciousness assessment in the field. (Choi et al., 2023; Pokee et al., 2021; Romero-Brufau et al., 2025)

Recent advances in artificial intelligence (AI) have opened new avenues for decision support in emergency care, particularly through the analysis of multimodal data (Wang et al., 2024). In the prehospital context, two evidence streams are especially relevant. First, physiological measurements

obtained from monitoring devices and wearable sensors provide objective signals may correlate with deterioration risk and neurological status (Breteler et al., 2020; Mayampurath et al., 2020; Romero-Brufau et al., 2025) Second, patient responsiveness can be observed through video, capturing behavioural and interactional cues associated with consciousness. (Al-Tekreeti et al., 2024; Choi et al., 2023) While these modalities have been studied independently in various clinical settings, their use as complementary components of a proof of concept focused on the role of AVPU within a NEWS-based escalation logic remains insufficiently characterized for prehospital workflows.

Accordingly, this dissertation proposes a feasibility and framework study by evaluating a proof of concept to examine how level of consciousness, assessed using AVPU, contributes to deterioration risk stratification in prehospital care, using NEWS as the reference framework. The proof of concept is implemented through two complementary approaches. The first a video-based pipeline grounded in clinically designed scenarios that emulate prehospital interactions and enable the study of responsiveness patterns aligned with AVPU categories. The second is an approach is a sensor-data pipeline that analyses physiological measurements to investigate their association with AVPU and NEWS escalation thresholds. Together, these approaches aim to clarify the operational significance of consciousness assessment and to establish a foundation for future integration into real-time decision-support tools.

**Thesis Organization**

This dissertation is organized as follows.

Chapter 1 introduces the clinical and operational motivation for prehospital deterioration risk stratification, positions the role of level of consciousness within the NEWS framework, and defines the proof-of-concept scope and objectives.

Chapter 2 describes the video-based proof of concept, detailing the clinical design of scenarios, the creation and annotation of the video dataset, and the evaluation protocol used to assess responsiveness recognition aligned with AVPU categories; the chapter also clarifies the collaboration context with the implementation work conducted by Marco Esperança.

Chapter 3 presents the sensor-data approach, including dataset characterization, data quality assessment, analytical methods, and the exploration of relationships between physiological signals, AVPU status, and NEWS-oriented escalation thresholds. *The rational for Tomás Mendes study*

Chapter 4 integrates and evaluates both approaches, comparing their feasibility and complementary value, presenting analytics and dashboard outputs, and discussing the results from a clinical perspective relevance, highlighting operational constraints.

Chapter 5 concludes the dissertation by summarizing the main achievements, articulating limitations, and outlining future work toward a real-world multimodal system for EMS implementation.

## Chapter 1 - Background, motivation, scope and objectives

*Introduces the clinical and operational motivation for prehospital deterioration risk stratification, positions the role of level of consciousness within the NEWS framework, and defines the proof-of-concept scope and objectives.*

### 1.1 Background and Rationale

Identifying deterioration at an early stage has been associated with morbidity, mortality, and healthcare costs. (Abbott et al., 2018), consequently, to support clinical decision, several early warning scores have been developed to synthesize physiological data into structured risk assessments. Among these, the NEWS Score been widely adopted due to its simplicity, reproducibility, and validated prognostic value. (Burgos-Esteban et al., 2022)

NEWS integrates routinely measured vital signs with an assessment of level of consciousness, and is calculated by assigning a weighted score (0–3 points) to six routinely measured physiological parameters: respiratory rate (RR), oxygen saturation (sat.O2), systolic blood pressure (SBP), heart rate (HR), temperature (temp.), and level of consciousness, assessed using the AVPU scale. (Downey et al., 2017)

Each parameter is scored according to the degree of deviation from the normal physiological range, with 0 representing normal values and higher scores reflecting increasing physiological disturbance. The individual parameter scores are then summed to produce an aggregate score. In addition, a supplementary 2 points are added if the patient is receiving supplemental oxygen (binary scoring: yes/no), reflecting the increased clinical risk associated with oxygen requirement.

The obtained NEWS score represents the cumulative burden of physiological abnormality, through aggregate scoring rather than reliance on a single abnormal parameter. Escalation thresholds are linked to the total score: values <5 generally indicate lower risk, a score ≥5 suggests medium risk requiring urgent clinical review, and a score ≥7 indicates high risk and the need for emergency assessment (Figure 1.1).

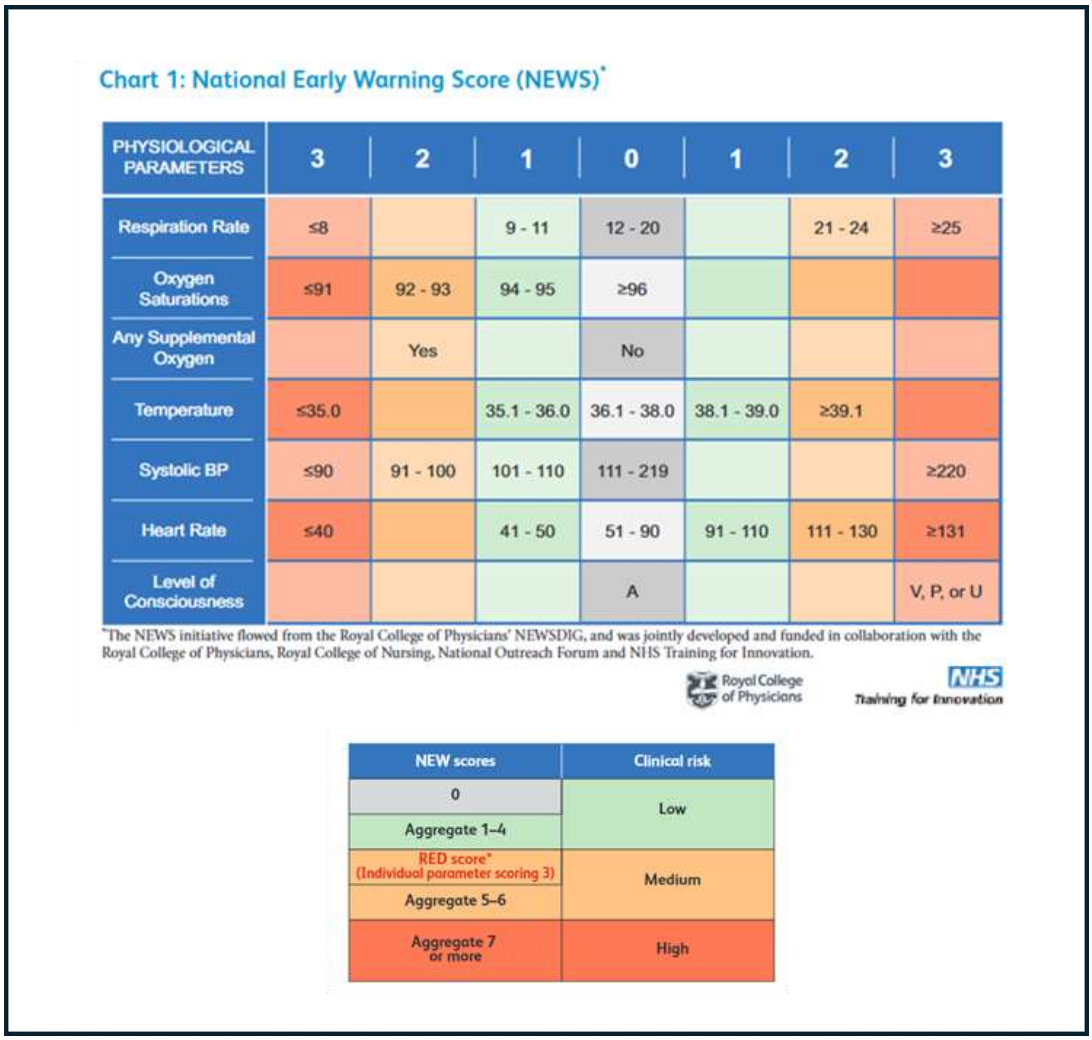


Chart 1: National Early Warning Score (NEWS)*

| PHYSIOLOGICAL PARAMETERS | 3 | 2 | 1 | 0 | 1 | 2 | 3 |
|---|---|---|---|---|---|---|---|
| Respiration Rate | ≤8 | | 9 - 11 | 12 - 20 | | 21 - 24 | ≥25 |
| Oxygen Saturations | ≤91 | 92 - 93 | 94 - 95 | ≥96 | | | |
| Any Supplemental Oxygen | | Yes | | No | | | |
| Temperature | ≤35.0 | | 35.1 - 36.0 | 36.1 - 38.0 | 38.1 - 39.0 | ≥39.1 | |
| Systolic BP | ≤90 | 91 - 100 | 101 - 110 | 111 - 219 | | | ≥220 |
| Heart Rate | ≤40 | | 41 - 50 | 51 - 90 | 91 - 110 | 111 - 130 | ≥131 |
| Level of Consciousness | | | | A | | | V, P, or U |

*The NEWS initiative flowed from the Royal College of Physicians' NEWSDIG, and was jointly developed and funded in collaboration with the Royal College of Physicians, Royal College of Nursing, National Outreach Forum and NHS Training for Innovation.

Royal College of Physicians

NHS Training for Innovation

| NEW scores | Clinical risk |
|---|---|
| 0 | Low |
| Aggregate 1–4 | |
| RED score* (Individual parameter scoring 3) | Medium |
| Aggregate 5–6 | |
| Aggregate 7 or more | High |

Figure 1.1- NEWS Score - Royal College of Physicians - NHS
https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news

In the prehospital context, the use of NEWS has increased alongside EMS digitalization and the integration of portable monitoring devices capable of automatic capture of physiological parameters. (Lindskou et al., 2019) Despite advances in automated data acquisition, the assessment of level of consciousness remains dependent on manual clinical evaluation. (E. et al., 2019; Isabel et al., 2020)This introduces subjectivity, inter-observer variability, and incomplete documentation, particularly in high-pressure and time-critical environments.

The AVPU scale is commonly used in emergency care as a rapid method to assess level of consciousness. Within the NEWS framework, AVPU distinguishes between fully alert patients and those with different degree of impaired consciousness. Despite its apparent simplicity, the AVPU component plays a decisive role in risk stratification and escalation of care (Forster et al., 2022)

### 1.2 Role of level of consciousness within the NEWS framework

AVPU scale was firstly used to access trauma patients and generally taught in ATLS® courses (Parrino et al., 2024),but is now a common component of early-warning scores and rapid response activation criteria. AVPU demonstrated utility for rapid triage decisions while being faster and simpler to apply than consciousness assessment scores like Glasgow Coma Score (GCS) (Janagama et al., 2022) particularly in prehospital settings, where speed and simplicity outweigh the need for detailed neurological granularity. (Figure 1.2)

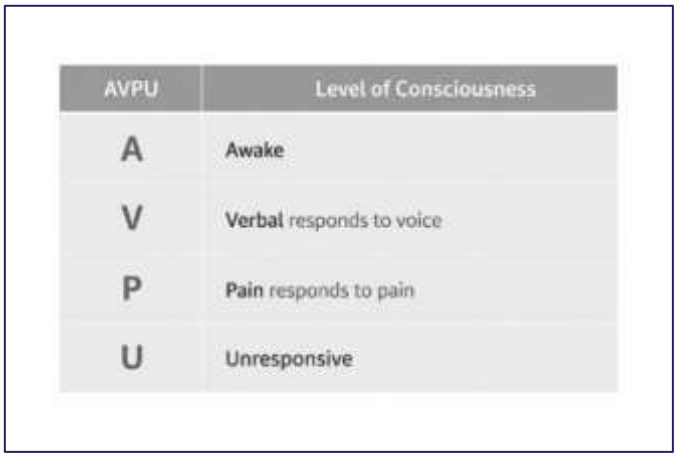



Figure 1.2 - AVPU scale
The ATLS Subcommittee, American College of Surgeons' Committee on Trauma
https://doi.org/10.1097/TA.0b013e31828b82f5

Within the NEWS framework, level of consciousness constitutes a critical physiological domain because alterations in neurological responsiveness precede overt clinical deterioration (Williams, 2019). NEWS was designed as an aggregate score integrating vital signs and consciousness to provide a standardized language of acute illness severity across clinical settings. Literature consistently identifies new-onset confusion or reduced responsiveness as strong independent predictors of adverse outcomes, including unplanned ICU admission and in-hospital mortality (Forster et al., 2022; Williams, 2019). Even subtle deviations from baseline mental status are associated with increased short-term deterioration risk, reinforcing the conceptual basis for including consciousness assessment alongside traditional vital signs in early warning systems.

However, despite its prognostic strength, the clinical determination of consciousness remains inherently complex. AVPU was designed to assess responsiveness rather than to provide a comprehensive cognitive or delirium evaluation(Williams, 2019). As such, it may fail to capture subtle cognitive changes, potentially underestimating deterioration in patients with acute confusion or delirium. Furthermore, AVPU assessment is rapid and pragmatic but is vulnerable to inter-observer

variability (Haldar et al., 2020), and may be influenced by contextual factors such as environmental conditions, communication barriers, or pre-existing cognitive impairment. (Forster et al., 2022) This reflects a central paradox within early warning systems: one of their most powerful predictors is simultaneously one of the most difficult to standardize and operationalize reliably in routine clinical workflows.

NEWS2 intended to address these challenges by refining the neurological component through the incorporation of “new confusion” (ACVPU), aiming to improve sensitivity for early deterioration. Evidence indicates that this modification increased detection of at-risk patients and strengthened short-term mortality prediction, maintaining high discrimination for death within 24 hours. (Forster et al., 2022).

Nevertheless, greater sensitivity was accompanied by a substantial rise in trigger frequency and escalation events, with some institutions reporting up to a threefold increase in clinical escalations following NEWS2 implementation. While mortality showed modest improvement in certain analyses, specificity decreased and workload increased, raising concerns about alert fatigue and resource utilization. Additionally, large multicentre comparisons have shown that NEWS2 does not consistently outperform NEWS in overall mortality discrimination and may demonstrate reduced performance in specific subgroups, such as patients at risk of type II respiratory failure (Pimentel et al., 2019)

EMS services, as the Portuguese Emergency Medical Service, adopted NEWS over other scores considering its fast and reliable results, reflecting a deliberate choice to preserve prognostic validity while minimizing unnecessary escalation and workload inflation. ((Barros F, 2021)

**1.3 Importance of risk stratification in Prehospital settings**

Clinical deterioration is often preceded by measurable changes in physiological parameters and early detection of these changes allows for interventions that may prevent progression to organ failure, cardiac arrest, or death. In prehospital care, early assessment is particularly critical due to limited diagnostic resources, time constraints, and the need to make rapid decisions regarding transport priority and destination (Smith et al., 2022).

In Portugal the NEWS score was validated after translation and cultural adaptation as a tool for early detection of clinical deterioration, supporting patient safety (Leandro Luís, 2014). Since 2020 the Portuguese National Institute of Medical emergency (INEM) adopted the use of NEWS score to access risk of deterioration. EMS technicians register vital signs and introduce the accessed letter for the AVPU scale when evaluating patients, as a daily routine.

The Portuguese electronic prehospital health record - iTeams® automatically calculates the NEWS score and shares the numeric and colour result, in real time, to the National Emergency Coordinating Centre (CODU) giving both the professionals on scene and the Coordinating Center a proxy of the patient´s deterioration risk (Figure 1.3 and 1.4).

This information is vital for teams on the field as an alert of the patient risk of deterioration is visible on their working screen. Also, the shared NEWS score information allows the CODU to proficiently allocate resources, identify the adequate destination hospital and pre-notify emergency departments of patient´s arrival.

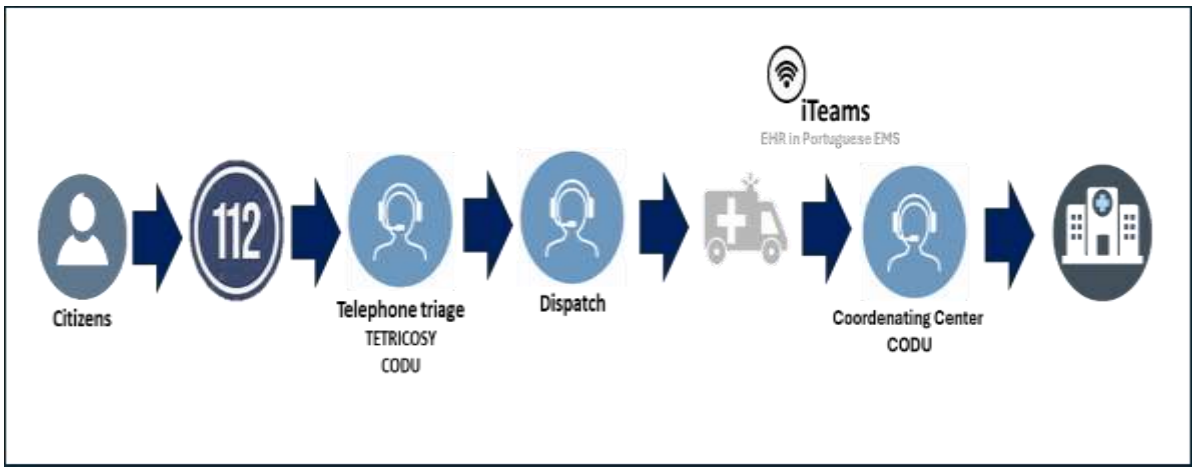



Figure 1.3 - INEM workflow - Author's own work



Figure 1.4 - Use of NEWS in the Portuguese EMS: iTeams® and CODU - Author's own work

### 1.4 Scope and objectives

This work aims to define and structure a methodological framework to support future evaluation and technological development of:

- A video-based analytical structure grounded in clinically designed simulation scenarios that replicate prehospital interactions and enable systematic assessment of behavioural responsiveness aligned with AVPU classification.
- A sensor-data analytical pipeline designed to explore how physiological parameters relate to AVPU categories and NEWS escalation thresholds.

This dissertation seeks to clarify the operational relevance of consciousness assessment within early warning systems and to establish a structured foundation for future validation and integration into real-time decision-support tools in emergency care.

## Chapter 2 -Video-based AVPU classifier

*Describes the video-based proof of concept, detailing the clinical design of scenarios, the creation and annotation of the video dataset, and the evaluation protocol used to assess responsiveness recognition aligned with AVPU categories; the chapter also clarifies the collaboration context with the implementation work conducted by Marco Esperança.*

### 2.1 Chapter Overview

This chapter presents the video-based component developed in this dissertation to evaluate the role of level of consciousness in prehospital deterioration risk stratification. Within the NEWS framework, level of consciousness, recorded using the AVPU scale, is treated as a high-impact clinical sign and can decisively shift escalation decisions. (Patel et al., 2018) Despite its importance, AVPU assessment remains observational, context-dependent, and vulnerable to the variability of EMS settings. (Bhaumik et al., 2022)

The primary purpose of the video-based branch of this dissertation is to establish whether AI video analysis can extract reproducible responsiveness cues coherent with AVPU logic when provided with appropriate data. To enable this evaluation, the major contribution of this chapter is the design and construction, in a controlled environment, of a clinically grounded dataset of videos and derived image sequences suitable for training and validating computer vision models. This controlled dataset is intentionally structured to balance realism with repeatability, ensuring that responsiveness patterns can be observed, annotated, and compared against algorithmic outputs under standardized conditions. The chapter therefore emphasizes the end-to-end dataset creation pipeline, namely the scenario design, recording protocol, quality control, annotation, and ground-truth definition as the critical enabler for subsequent AI experimentation.

### 2.2 Clinical Rationale

#### 2.2.1 Video use for face detection

Video-based face analysis has matured from biometric recognition into a broader toolkit for measuring where the person is looking, whether eyes are open, and how facial features move over time. (Kortli et al., 2020) Modern video analytics provides a practical foundation for assessing clinical state from observable human behaviour, because it can reliably locate faces, track landmarks, and follow motion over time. Face-recognition literature describes robust pipelines built around face detection, feature extraction and classification, with explicit attention to operating on both images and video in (near) real time and under variability such as head orientation, lighting, and facial expression. (L. Li et al., 2020)

These capabilities are not limited to identifying individuals; the same landmarking and tracking steps can quantify *behavioural signals* that matter clinically (e.g., eye-opening state, head/eye direction, and purposeful movement). In practice, this shift from identity recognition to behaviour interpretation is what enables video to support physiological or neurological screening.

Modern reviews emphasize deep-learning (e.g., CNN-based) representations and the importance of robustness in real conditions (pose/illumination/low resolution), which is directly relevant to clinical environments where cameras cannot be perfectly controlled. (L. Li et al., 2020)

### 2.2.2 Drowsiness

A mature and widely deployed example of video-based system to assess human alertness in safety-critical environments is found in the automotive domain, where driver drowsiness detection systems are integrated into commercial vehicles. These systems aim to identify early signs of reduced alertness to prevent accidents, using non-intrusive sensing technologies that operate continuously and in real time (Essahraui et al., 2025)

Within this field, research has explored three principal dimensions of drowsiness detection, namely physiological, behavioural, and vehicle-based approaches (Ramzan et al., 2019). Physiological methods rely on bio signals such as electroencephalography, electrocardiography, or electrooculography, which provide direct indicators of fatigue but require intrusive sensors. Vehicle-based methods infer drowsiness indirectly from driving patterns such as lane deviation, steering behaviour, or pedal inputs, but are highly dependent on road conditions and driving context. Behavioural methods, by contrast, analyse observable cues from the driver, including eye closure, blink rate, yawning, head pose, and gaze direction, using machine learning (Ramzan et al., 2024) and artificial neural network models (Inkeaw et al., 2022).

Hybrid systems (Ramzan et al., 2024) combining two or more of these dimensions have demonstrated robustness by compensating for the limitations of individual modalities. However, such systems introduce additional complexity, sensor dependencies, and integration challenges that limit large-scale deployment. Consequently, contemporary research and industrial implementations have largely converged on camera-based behavioural techniques approaches (Ramzan et al., 2019). Contemporary systems analyse observable behavioural cues such as eye closure duration, blink frequency, yawning, head pose, and gaze direction. Vision-based systems do not attempt to infer internal neurological states directly; instead, they classify externally observable behaviours that correlate with reduced alertness, making them well suited for real-world, real-time safety applications.

### 2.2.3 Consciousness Assessment in Healthcare

Advances in neuroimaging have transformed the conceptual and clinical understanding of consciousness. Functional modalities such as Functional Magnetic Resonance Imaging (fMRI) (Cainer et al., 2024).Electroencephalography (EEG), Positron Emission Tomography (PET), and Transcranial Magnetic Stimulation combined with Electroencephalography (TMS-EEG) have demonstrated the ability to detect residual and covert awareness (Kazazian et al., 2025; Sanz et al., 2021).

Recent reviews advocate for multimodal and probabilistic frameworks that integrate neuroimaging with behavioural and electrophysiological measures to improve diagnostic precision and prognostic confidence (Demertzi et al., 2020,Sanz et al., 2021). Rather than establishing a single gold standard, these approaches emphasize complementary tools to reduce misdiagnosis and better guide clinical and ethical decision-making (Kazazian et al., 2025). However, despite their diagnostic potential, advanced neuroimaging methods remain limited by cost, infrastructure requirements, and the need for specialized expertise, restricting their use to highly specialized centers. Consequently, most emergency and prehospital settings continue to rely primarily on rapid behavioural observation.

In response to this gap, video-based technologies and computer vision have emerged as scalable tools capable of objectifying behavioural responsiveness. Automated systems using visible and infrared imaging have demonstrated accuracy in detecting motion and responsiveness in simulated disaster environments, including night-time conditions (Pokee et al., 2021). These contactless approaches enable remote triage and provide quantifiable metrics of motor activity, supporting more objective and reproducible assessments in environments where direct neurological examination may be delayed or unsafe. By converting observable behaviours into computational data, such systems may reduce inter-observer variability and enhance consistency in high-pressure scenarios.

Beyond gross motor detection, qualitative video analyses of clinician–patient interactions reveal that subtle behaviours, such as gaze direction, micro-movements, and interactional timing may indicate preserved communicative competence even in patients diagnosed with disorders of consciousness (Papadimitriou et al., 2023). These nuanced indicators are often overlooked by standardized scoring systems but can potentially be captured through computational modelling. The convergence of neuroimaging insights, video analytics, and multimodal assessment frameworks points toward a future in which consciousness evaluation becomes objective, scalable, and sensitive. Within this broader landscape, simplified bedside tools such as AVPU remain essential, serving as the practical interface between complex neuroscientific understanding and real-time clinical decision-making.

The AVPU scale is a widely used simplified tool for rapid assessment of consciousness in emergency and ward settings. Designed for speed and ease of application, AVPU categorizes responsiveness into four hierarchical levels that can be assessed within seconds. Studies comparing AVPU with more detailed instruments such as the Glasgow Coma Scale (GCS) demonstrate strong correlation between categories, particularly in identifying severe impairment (Janagama et al., 2022). This comparability supports AVPU's utility in time-critical prehospital environments, where rapid triage decisions often take precedence over detailed neurological granularity.

In emergency medical services (EMS), the practicality of neurological scales is essential. An expert panel evaluating multiple neurologic assessment tools for older adults with traumatic brain injury found that simplified approaches as AVPU and motor-focused scales were highly rated for usability and feasibility, though none fully satisfied all clinical needs (Wasserman et al., 2015).

A dichotomous strategy centered on alertness and motor response, reinforces the conceptual simplicity underlying AVPU, narrowing the gap between comprehensiveness and feasibility in prehospital triage algorithms.

Also, when examined in inpatient settings, AVPU demonstrates reasonable predictive capability for mortality and clinical deterioration, although more granular scales such as GCS may show stronger prognostic discrimination (Zadravecz et al., 2015, McNarry & Goldhill, 2004).

This trade-off reflects a broader principle in clinical risk assessment: tools optimized for speed and routine use may sacrifice some predictive precision. Nonetheless, AVPU's simplicity enables frequent reassessment and broad adoption, making it an effective screening instrument even if it is not the most detailed prognostic tool available.

Bedside evaluation, including airway, breathing, circulation, and neurological responsiveness remain the cornerstone of safe patient management (Hill et al., 2023). Within this framework, AVPU provides an accessible and reproducible method for rapidly categorizing mental status as part of initial stabilization and ongoing monitoring. As emerging technologies such as computer vision and sensor-based analytics evolve, AVPU may serve as both a reference framework and a target variable for automated systems, bridging traditional clinical examination with next-generation decision-support tools

### 2.2.4 Video to access consciousness

Within healthcare-oriented computer vision, the conceptual pathway from observable behaviour to clinically meaningful inference has already been established in adjacent domains such as automated

pain detection. In the Pain Detection/Classification Framework, pain intensity is operationalized through the Facial Action Coding System (FACS) and the Prkachin and Solomon Pain Intensity (PSPI) metric, which aggregates specific facial action units including orbital tightening, levator contraction, lip raising, and eye closure (Elgendy et al., 2021).
These systems extract features followed by feature selection and machine-learning classification. The result is a structured pipeline in which subtle facial movements are translated into quantifiable clinical indicators, demonstrating that externally observable facial dynamics can reliably encode internal physiological states.

Building on this logic, the AI-Based Visual Early Warning System extends the paradigm from discrete symptom detection to continuous risk monitoring. Rather than relying exclusively on intermittent measurements, visual signals as facial expression changes, head posture, and eye activity can be modelled as continuous temporal data streams that correlate with clinical deterioration (Al-Tekreeti et al., 2024).

This visual early warning approach reframes the patient's face and body as dynamic sensors, capable of contributing to escalation logic alongside traditional physiological parameters. In this context, AVPU becomes a natural target variable for computer vision systems, as it already constitutes the central nervous system component within early warning scores and reflects graded responsiveness observable through behaviour.

Importantly, the AVPU categories themselves are grounded in observable stimulus–response patterns. The AVPU-AVSR study demonstrates that these behavioural definitions can be computationally reconstructed from synchronized whole-body video, facial video, and speech recognition inputs (Choi et al., 2023). Using pose estimation to detect arm movement and pain stimulus application, facial landmark tracking to determine eye-opening state, and speech recognition to identify verbal commands, the system algorithmically reproduced the clinical examination workflow. In simulated emergency scenarios, the AVPU-AVSR system achieved a multi-class accuracy of 0.95 and a sensitivity of 1.00 for detecting impaired consciousness, without misclassifying impaired patients as alert. (Choi et al., 2023)

Taken together, these studies delineate a coherent trajectory from facial and motion analytics to structured clinical decision variables. Pain detection research establishes that facial action units and temporal dynamics can encode clinically relevant internal states. Visual early warning frameworks demonstrate that such signals can be integrated into continuous monitoring architectures. Finally, the

AVPU-AVSR system confirms that video-derived cues can be mapped directly onto standardized consciousness categories.

In an early warning workflow, AVPU may therefore function as a machine-readable behavioural endpoint, while preserving a transparent and clinically interpretable link between what is observed on video and what is scored in practice.

**2.3 Proposing an AVPU video-observable classifier**

Rational for binary AVPU approach

Assessment of consciousness is the first step in emergency care because it determines the entire course of action. In Basic Life Support (BLS), responsiveness is simplified into a rapid, binary clinical judgment: the responder determines whether the patient is responsive or unresponsive by applying verbal and, if necessary, physical stimuli. This assessment does not aim to establish a neurological diagnosis; rather, it functions as a trigger for immediate downstream actions such as airway management, breathing assessment, activation of emergency response systems, and initiation of cardiopulmonary resuscitation. Both the European Resuscitation Council (Soar et al., 2021) and the American Heart Association (Donoghue et al., 2025) emphasize that uncertainty should default toward treating the patient as unresponsive, prioritizing speed and safety over diagnostic precision.

Within this framework, consciousness assessment becomes the gateway to the entire resuscitation pathway. The rescuer first ensures scene safety, then applies verbal and physical stimuli to provoke a reaction. Any meaningful response indicates responsiveness; absence of response defines unresponsiveness. Once unresponsiveness is confirmed, the BLS algorithm proceeds immediately. The logic is intentionally binary and action-oriented and reflects a safety-first philosophy in which rapid recognition of potential life-threatening states outweighs the need for detailed neurological stratification.

The AVPU scale can be understood as a structured formalization of this BLS responsiveness check within professional and prehospital environments. Although AVPU is not explicitly taught to lay rescuers, its categories map directly onto BLS logic. Spontaneous or meaningful interaction corresponds to Alert; response to verbal command aligns with Verbal; reaction only to physical stimulation matches Pain; and absence of response defines Unresponsive. Conceptually, AVPU encodes the presence or absence of observable responses to escalating stimuli. From an operational perspective, this structure makes AVPU highly adaptable to prehospital conditions where altered consciousness is clinically relevant and where rapid, reproducible categorization is essential.

Framing AVPU within a binary “impaired versus non-impaired” paradigm is therefore clinically coherent. In practical terms, Alert may represent a non-impaired state, whereas Verbal, Pain, and Unresponsive can be grouped as varying degrees of impaired responsiveness. Such dichotomization mirrors BLS principles, preserves clinical interpretability, and aligns with escalation-based workflows. A video-based AVPU classifier grounded in this logic does not attempt to replace detailed neurological examination; rather, it digitizes the first judgment in emergency care, namely whether a response is present and what stimulus elicits it. By embedding this binary safety-first reasoning into automated systems, the approach remains clinically defensible, operationally efficient, and consistent with established emergency algorithms.

A binary approach to AVPU is conceptually aligned with the way consciousness is operationalized within the National Early Warning Score. In NEWS, the neurological component does not require detailed stratification of mental status but instead assigns a score based on whether the patient is Alert or exhibits any deviation from alertness—traditionally captured as response to Voice, Pain, or Unresponsive. In practical terms, this functions as a dichotomous variable: “Alert” scores 0, while any impaired level of responsiveness triggers additional 3 points and contributes directly to escalation thresholds. Thus, NEWS already embeds a binary safety logic in which the critical distinction is not the depth of impairment but its presence. Framing AVPU as “non-impaired (Alert)” versus “impaired (V/P/U)” therefore mirrors the NEWS structure, preserves clinical interpretability, and supports automated integration into early warning workflows where the detection of deterioration, rather than fine neurological grading, is the primary objective. (Table 2.1)

Table 2.1 - Relating BLS, AVPU and the NEWS score

| BLS concept | AVPU equivalent | NEWS |
|---|---|---|
| Responds spontaneously | Alert (A) | 0 |
| Responds to voice | Verbal (V) | 3 |
| Responds only to physical stimulus | Pain (P) | |
| No response | Unresponsive (U) | |

### 2.4 Dataset Design Principles

AVPU is determined through interaction, observation, and response to stimuli rather than through a single measurable signal. Consequently, modelling AVPU-related states from video requires a dataset that captures temporal interaction dynamics, including stimulus delivery and patient reaction, rather than static appearance. The dataset was planned (annex 4) in alignment with the GROVE guidelines (Hillen et al., 2025) and designed according to three principles:

- First, clinical plausibility was prioritized, and each scenario was grounded in common prehospital assessment behaviours used to establish responsiveness, namely the BLS sequence explained previously.

- Second, reproducibility was enforced through controlled conditions, namely the same interaction patterns were repeated across recordings using consistent camera placement and scene layout, reducing unwanted variance to enable reliable algorithmic comparison.

- Third, computational tractability guided data structuring was maintained as recordings were produced to preserve visibility of body regions relevant to responsiveness, enabling downstream extraction of image frames and motion features commonly used in AI video pipelines.

This controlled dataset strategy supports the proof-of-concept objective. Rather than attempting to claim clinical grade automation of AVPU, the objective is to demonstrate that responsiveness cues aligned with AVPU logic can be represented in data and subsequently learned and evaluated by AI models. Establishing such a dataset is a prerequisite for credible feasibility testing and provides a foundation for future expansion toward less constrained real-world settings.

### 2.5 Controlled Environment Dataset Construction

The development of a dataset suitable for AI-based video analysis requires careful control of acquisition conditions to minimize sources of variability. In prehospital environments, responsiveness cues may be obscured by factors such as camera motion, variable illumination, occlusions, background activity, and inconsistent viewpoints. For a proof-of-concept study, these uncontrolled variables can compromise the interpretation of model performance, making it difficult to determine whether prediction errors arise from algorithmic limitations or from poor-quality visual input. Consequently, a controlled recording environment was adopted to isolate the phenomenon of interest: patient responsiveness during AVPU assessment.

The recording setup was designed to standardize key acquisition parameters while preserving clinically plausible interactions between the patient and the emergency responder. Participants were seated facing a tripod-mounted camera positioned approximately 1.5 m away at eye level under controlled studio lighting, ensuring consistent framing and minimizing variations in illumination. The interviewer was positioned slightly off-axis to administer the predefined AVPU stimulus sequence without obstructing the participant's face or upper body, allowing unobstructed capture of facial expressions, head movements, and upper-limb responses relevant to responsiveness assessment.

Video acquisition was performed using a Canon EOS RP camera equipped with a Canon EF 16–35 mm f/4L IS USM lens configured at a 35 mm focal length. Recordings were captured at a resolution of 1920 × 1080 pixels and a frame rate of 25 frames per second, generating MPEG-4 video clips with durations ranging from approximately 6 to 12 seconds. Each participant and simulated consciousness-level condition constituted an individual dataset entry, enabling consistent representation of the different AVPU response scenarios. (annex A - scenarios)

The patient role was primarily performed by healthcare-related professionals, while the emergency medical services (EMS) responder role was assumed by the study organizers to ensure standardized execution of the assessment protocol across all recordings.

Following acquisition, individual video clips were processed to generate image sequences for computer vision applications. Frame-level extraction enabled the development of image-based classification models, while temporal segmentation supported sequence-based approaches. Associated timestamps and metadata—including scenario type, simulated AVPU condition, and annotation references—were preserved throughout the processing pipeline to ensure traceability and reproducibility. Consequently, the resulting dataset represents not only a collection of standardized video recordings but also a curated experimental resource specifically designed to support the development and evaluation of AI models for automated responsiveness assessment.

**2.6 Scenario Specification and Scripted Interaction Design**

A critical step in dataset creation is the definition of scenarios that encode clinically meaningful responsiveness patterns. Scenarios were specified to reflect the operational logic of AVPU assessment, where the responder escalates from observation to verbal prompting and to the application of a pain stimulus, while monitoring for reaction.

The scenario scripts were designed to capture interaction stages that are interpretable by clinicians and labelable for AI training.

Scenarios representing an alert state emphasize spontaneous engagement, orientation cues, and prompt behavioural response to verbal interaction. Scenarios representing response to Verbal stimuli focus on delayed or reduced engagement that becomes evident only after prompting. Scenarios representing response to Pain stimuli encode minimal responsiveness until a discomfort stimulus is applied, after which a motor or behavioural reaction becomes visible. Unresponsive scenarios encode the absence of observable reaction across escalation steps. The dataset design therefore mirrors the operational pathway by which AVPU is applied, ensuring that the recorded data supports both clinical plausibility and computational learning of stimulus–response relationships.

While scripts were structured, performances were not intended to be theatrical; rather, they were executed to preserve realistic motion patterns and interaction timing. This approach supports AI training by providing consistent event structure while avoiding artificial cues that would not generalize beyond the controlled environment. (annex B - Script)

**2.7 Ground Truth Definition**

In this study, the prediction target was explicitly defined at two complementary levels: the clip level and the global scenario level. At the clip level, each video segment was annotated according to the type of stimulus applied (verbal stimulus, verbal plus tactile stimulus, or pain stimulus) and the observable response.

Response modalities included eye behaviour, facial expression changes, head movement, and upper limb movement. (Table2.2)

At the global level, these events were mapped on to a consciousness proxy aligned with AVPU logic and consistent with Basic Life Support principles (Figure 2.1):

- “Alert (A)” was operationally defined as spontaneous or clearly purposeful response to interaction, particularly to verbal stimulus, without the need for escalation.
- “Non-Alert (V/P/U)” was defined as absence of spontaneous responsiveness and requirement of escalating stimuli to elicit any reaction, or complete absence of observable response even after pain stimulus.

In practical terms, “responds to voice” required consistent eye or facial engagement following verbal/touch prompting; “responds to pain” required absence of response to voice/touch but clear motor or facial movement only after painful stimulus; “unresponsive” required no observable reaction across all escalation steps.

These codes were written and collaboratively shared to standardize labelling decisions, reduce inter-rater ambiguity and ensuring that the video-derived ground truth mirrors the operational semantics of AVPU and the binary “Alert versus impaired” distinction embedded in NEWS, while remaining observable, reproducible, and computationally tractable.

Table 2.2 - Mapping face, head and upper limb response to stimulus

| Structure in focus | Identification of “A” | Identification of "Non-A" (VPU)** |
|---|---|---|
| Eyes | Blink and eye movement | No movement |
| | opened eyelids | Closed eyelids |
| Lips | Any mouth or lips movement | No movement of mouth or lips |
| Face | Frown | No facial expressions |
| | Jaw movement | |
| | Any face muscle contraction | |
| Head | Any movement | No movement |
| Upper limbs | Any movement | No movement |
| Sound | NA | NA |

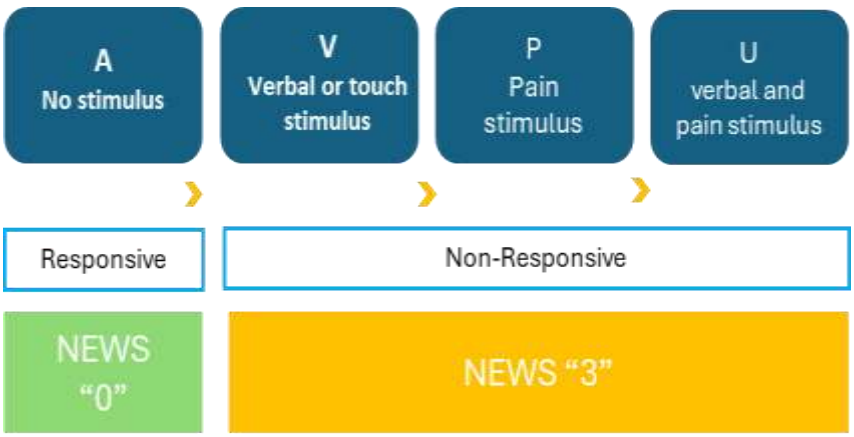



Figure 2.1 - Mapping response to stimulus in AVPU scale and matching NEWS score. (Author´s own work)

**2.8 Annotation Strategy**

The annotation strategy aimed for multi-layer labels that could be usable for machine learning and feasible for the annotator. Each clip recorded global fields, namely the take number, the intended scenario namely “A”, “V” “P” and “U”, and the stimulus sequence with the approximate timing of each stimulus onset. Then each clip was divided into segments relative to stimuli, such as baseline (alert), voice and touch window, pain window, and unresponsive window.

For each segment, annotators record eye state, facial expression state and movement. A subset of frames was annotated, with presence or absence of key action units such as eyebrow raise, tight eye closure, or lip tightening used both for model calibration and for checking annotation bias.

For this proof of concept, there were two independent annotators per clip. One a clinician familiar with AVPU and similar prehospital assessments, and a trained non-clinician using the same codebook. An interface showing the video timeline with markers for each stimulus and simple controls to enter segment-level labels was used. Before starting full annotation, a small training set of ten practice clips was used with consensus labels and brief explanations, so annotators could calibrate their interpretations and clarify ambiguities.

To consolidate ground truth from annotators majority rules and clear decision procedures were defined. All variables were defined as categorical, namely the presence or absence of a response or scenario classification, if the annotators agree, that category become the ground truth. When there was disagreement or both rated “unsure,” the clip or segment was marked as ambiguous and excluded from primary analyses.

For the scenario label (responds to voice versus responds to pain only), the scripted condition was used as an intended label, but annotators independently classified what they saw. Clips where annotators disagreed with the intended script, such as detecting a clear response to voice in a “pain-only” scenario, were flagged and reviewed to check whether the acting or filming deviated from protocol.

To anchor the labelling system, a subset of clips (12) across participants and conditions were selected for expert consensus. Two experienced emergency clinicians reviewed these clips together and agreed on final labels. This subset of clips was defined as the gold-standard reference.

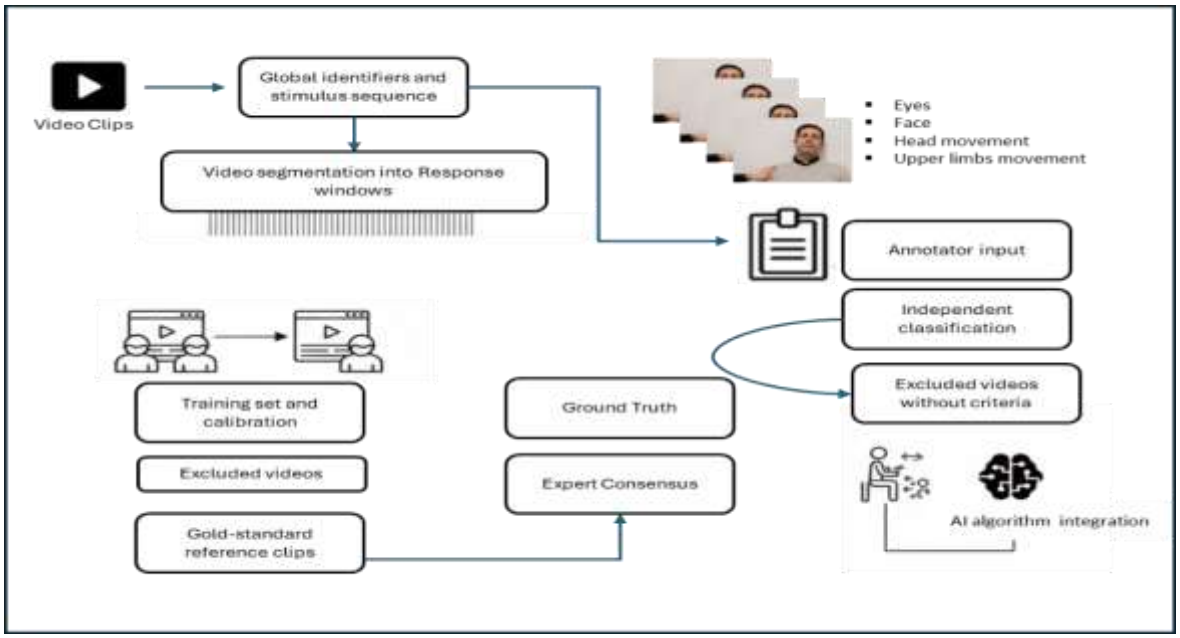



Figure 2.2 - Annotation framework (Author´s own work)




Figure 2.3 - Video annotation example (Author´s own work)

### 2.9 Validation framework

Inter-rater reliability was assessed on 112 video vignettes using Cohen's kappa (Annex5). For the four-category AVPU classification (Alert, Verbal, Pain, Unresponsive), observed agreement was 79.5%, with κ = 0.72, indicating substantial agreement. When AVPU categories were collapsed into a binary reaction presence label (Reaction vs. No Reaction), observed agreement increased to 93.8%, with κ =

0.79, reflecting substantial to near-excellent agreement. These findings support the reliability of the annotation protocol prior to model training.

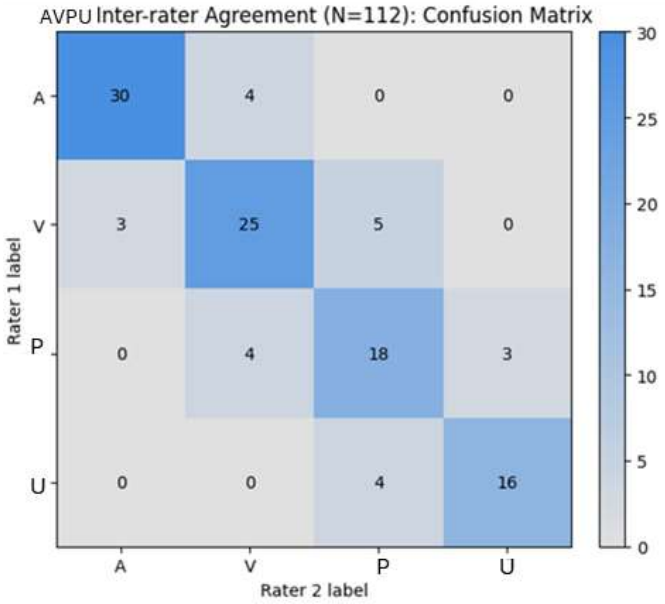



Figure 2.4 – AVPU inter-rater Agreement Confusion Matrix. ChatGPT, Version 4.0, Open AI, 2 February 2026, openai.com

The same two experienced emergency clinicians watched samples of each scenario and rated how realistic each clip was for them at the corresponding level of responsiveness, and whether the facial responses looked purposeful. Clips judged as unrealistic or ambiguous by any of the experts were removed.

Once the models are trained validation will be obtained by comparing model predictions to the consolidated ground truth and especially to the expert gold subset, and by examining performance separately on clips with high versus low annotator agreement. This analysis will tell us whether the model has learned robust patterns tied to the intended construct.

### 2.10 AI Video Pipeline and Collaboration Context

The controlled dataset created in this chapter served as the input resource for the AI video application used to extract and classify responsiveness cues. The technical implementation of the computer vision pipeline, including model development and training, was conducted in collaboration with Marco Esperança. The contribution of this dissertation within the collaboration is centred on the conception

and construction of the dataset and its clinical grounding, including scenario specification, acquisition protocol, data organization, annotation strategy, and evaluation framing.

The collaboration was structured to ensure that the AI pipeline could be developed against a dataset that was purpose-built for responsiveness modelling. This separation of roles is important for scientific clarity: the dataset contribution establishes the experimental validity of the proof of concept, while the implementation contribution enables computational testing. The dissertation therefore emphasizes that the reliability of conclusions drawn from the AI pipeline depends critically on the quality, structure, and clinical coherence of the controlled dataset.

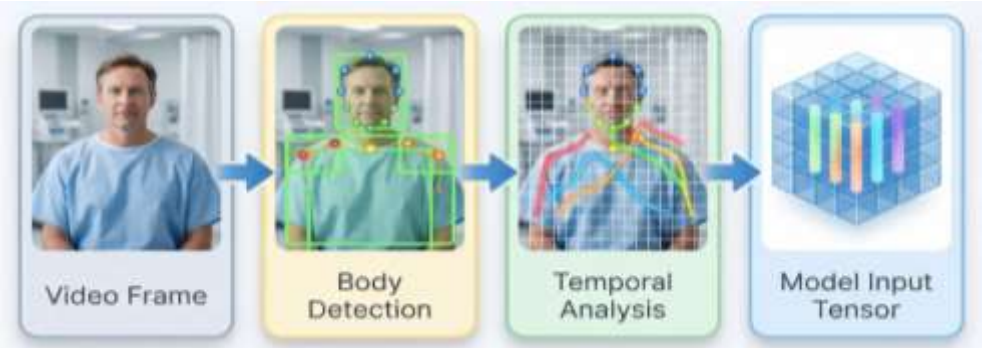



Figure 2.5- “Feature extraction pipeline for AVPU video-based consciousness classification”. ChatGPT, Version 4.0, Open AI, 2 February 2026, openai.com

**2.11 Summary and Link to the Sensor Data-Based Approach**

This chapter described the video-based branch of the proof of concept, with emphasis the contribution of creating, in a controlled environment, a clinically grounded dataset of videos and derived image sequences suitable for AI video applications. The dataset was engineered to represent AVPU-related responsiveness through structured interaction scenarios, consistent acquisition conditions, and temporally defined annotations, thereby enabling reproducible feasibility testing of computer vision models.

While video captures behavioural responsiveness and interaction dynamics, it does not directly represent internal physiological mechanisms of deterioration. For this reason, the dissertation complements the video-based approach with a sensor-data approach, presented in the next chapter, which analyses physiological measurements and their relationship with AVPU status and NEWS-based escalation thresholds. Together, these two modalities provide complementary perspectives on consciousness and deterioration in prehospital risk stratification.

## Chapter 3 - Vital signs data models

*Presents the sensor-data approach, including dataset characterization, data quality assessment, analytical methods, and the exploration of relationships between physiological signals, AVPU status, and NEWS-oriented escalation thresholds. The rational for Tomás Mendes study*

### 3.1 Overview

This chapter presents the sensor-data approach that underpins the analytical framework of this dissertation, integrating physiological time-series modelling with deterioration risk stratification. Vital signs constitute the immediate, continuously available, and physiologically grounded indicators of patient status across both prehospital and in-hospital environments (Leenen et al., 2020). Traditional early warning systems, including aggregate-weighted scores such as NEWS, rely on threshold-based interpretations of heart rate, respiratory rate, systolic blood pressure, temperature, oxygen saturation, and level of consciousness. Although clinically intuitive and operationally simple, these rule-based systems are inherently static: they apply additive points to single observations and do not explicitly capture temporal evolution, multivariate interactions, or individualized physiological baselines.

Recent advances in wearable wireless monitoring devices, continuous bedside sensors, and large-scale electronic health record (EHR) infrastructures have enabled the acquisition of high-frequency, longitudinal physiological data streams. These developments raise critical questions for contemporary risk stratification: Can continuously acquired vital-sign data meaningfully support or enhance deterioration scores such as NEWS? Can altered consciousness, traditionally assessed through subjective categorical scales such as AVPU, be inferred or anticipated from physiological patterns? Can risk assessment be improved through higher temporal resolution, trend detection, and automated multivariate interpretation?

Contemporary research reflects a paradigm shift from static, threshold-driven scoring toward data-driven approaches capable of modelling longitudinal trajectories, inter-variable dependencies, and dynamic, individualized risk estimation. The evolution of predictive modelling in healthcare has paralleled advances in machine learning and deep learning applied to structured EHR data, including transformer-based and hierarchical architectures capable of encoding long clinical sequences and complex temporal dependencies (Renc et al., 2025, Y. Li et al., 2023). Within this framework, vital signs are no longer treated as isolated measurements but as components of temporally structured risk representations.

The rationale for this line of work, developed in collaboration with Tomás Mendes, emerges from a central methodological asymmetry as in NEWS, consciousness contributes to the total score yet

remains a categorical, clinician-dependent, and potentially variable input (E. et al., 2019), whereas the remaining physiological parameters are objectively measured and increasingly captured through automated systems. Exploring whether multivariate physiological signals can be structured, characterized, and modelled to examine their relationship with AVPU status and NEWS-oriented escalation thresholds therefore represents a foundational step toward objective and temporally sensitive risk assessment.

This exploration is relevant in the prehospital setting, where early identification of the critically ill patient is essential and diagnostic resources are limited (Smith et al., 2022). In such environments, vital signs are often the primary data source guiding triage and escalation decisions. Using exploratory data from INEM, this chapter validates the necessity of relating physiological patterns to AVPU classifications. By examining associations between sensor-derived vital signs, consciousness status, and NEWS escalation thresholds, the chapter establishes the conceptual and methodological foundations for future automated decision-support systems capable of integrating objective physiological dynamics with established clinical risk frameworks.

### 3.2 Current Evidence for Vital Signs Monitoring by wearable devices

Continuous monitoring using wearable wireless devices has demonstrated technical feasibility and patient acceptability in inpatient settings. Literature review reports acceptable accuracy for heart rate, respiratory rate, and oxygen saturation, though evidence for improved clinical outcomes remains limited (Leenen et al., 2020,Leenen et al., 2020). Most investigations focus on validation and feasibility rather than escalation performance or mortality reduction.

Nevertheless, in high-risk surgical populations, continuous monitoring identified sustained abnormal patterns, such as progressive tachycardia, rising respiratory rate, and declining oxygen saturation, hours before clinical diagnosis of adverse events (Breteler et al., 2020). Transient changes related to mobilization were distinguishable from persistent pathological trends, emphasizing the importance of temporal modelling rather than isolated threshold breaches.

Prehospital wireless monitoring systems have similarly demonstrated feasibility for real-time acquisition and transmission of physiological data, supporting early situational awareness and interprofessional coordination (Koceska et al., 2020). Although continuous monitoring in prehospital care is not yet universal, these findings establish a technical foundation for extending trend-based analytics beyond hospital walls.

Despite technological maturity, systematic reviews highlight a persistent gap between signal acquisition and outcome-driven validation, underscoring the need for robust prospective evaluation before widespread implementation (Leenen et al., 2020; Muralitharan et al., 2021).

### 3.3 Sensor-Derived Physiological Data and Dataset Characterization

Sensor-derived datasets consist of multivariate time-series data characterized by irregular sampling, missingness, noise, and heterogeneity across care settings. Predictive modelling of biomedical temporal data must address challenges including variable measurement frequency, signal artifact, and short- versus long-term temporal dependencies(Patharkar et al., 2024). Physiological streams typically include heart rate, respiratory rate, systolic blood pressure, oxygen saturation, and temperature. Beyond absolute values, derived features such as variability indices, slopes, minimum and maximum values, and frequency of directional changes capture instability patterns. In sepsis prediction research, variability-based features extracted from four vital signs improved early detection compared with static summaries (Bloch et al., 2019).

Similarly, modelling strategies that incorporate dual-window feature extraction, capturing both recent deviations and baseline trends, have demonstrated predictive discrimination for inpatient deterioration using vital signs alone (Romero-Brufau et al., 2025). These findings indicate that longitudinal structure and contextualization of measurements are central to extracting prognostic information.

From a structural perspective, patient trajectories may be conceptualized as temporally ordered event sequences. Advanced architectures, including convolutional neural networks combined with dynamical systems approaches (Lehman et al., 2015), enable simultaneous modelling of short-term fluctuations and long-range dependencies, reinforcing the importance of preserving temporal resolution in physiological data.

Robust preprocessing is essential. Continuous monitoring data are prone to motion artifact, signal dropout, and underestimation biases under certain physiological conditions (e.g., arrhythmias) (Breteler et al., 2020). Data quality assessment therefore includes physiologically plausible range filtering, artifact detection, alignment with NEWS observation intervals, and transparent handling of missing data to prevent distortion of risk modelling outputs.

### 3.4 Vital Signs Monitoring and Clinical Risk Prediction

Vital signs are consistently identified as primary predictors of acute deterioration, ICU transfer, and mortality. Machine-learning–based early warning systems that rely exclusively on vital signs frequently outperform aggregate-weighted scores such as MEWS or NEWS (Muralitharan et al., 2021). Deep-

learning models using four routinely measured vital signs (heart rate, systolic blood pressure, respiratory rate, and temperature) demonstrated superior discrimination for in-hospital cardiac arrest compared with traditional track-and-trigger systems, identifying more than 50% of arrests up to 14 hours before the event (Kwon et al., 2018). False-alarm rates were reduced through recurrent neural network architectures capable of modelling sequential dependencies.

In large multicentre cohorts, vital-sign–only models have achieved high predictive performance for, ward deterioration and ICU transfer while maintaining manageable alert burdens (Romero-Brufau et al., 2025). Similarly, pediatric deterioration models based exclusively on objective physiological parameters outperformed modified bedside early warning systems in both discrimination and efficiency (Mayampurath et al., 2020).

Sepsis detection studies further demonstrate that physiological instability precedes overt diagnosis, with models identifying deterioration several hours before antibiotic initiation [(Abd El-Aziz & Rayan, 2025; Bloch et al., 2019; Sadeghi et al., 2018). Across settings, trend-based features, particularly slopes, variability, and minimum values, consistently outperform single-observation metrics (Churpek et al., 2016).

Collectively, these findings challenge the sufficiency of static threshold-based scoring and support embedding longitudinal physiological modelling within risk prediction frameworks aligned with NEWS escalation logic.

**3.5 Consciousness assessment in risk scores**

Consciousness assessment, as discussed in chapter 2, is a prognostically influential component of aggregate early warning systems such as NEWS. Any deviation from “Alert” in the AVPU contributes three points to the total score, reflecting strong evidence that altered mentation is independently associated with unplanned ICU admission, cardiac arrest, and short-term mortality. (Karunarathna et al., 2025)

Despite its predictive strength, consciousness assessment differs fundamentally from other parameters because it remains a categorical, clinician-dependent input, susceptible to inter-observer variability and documentation inconsistency (Haldar et al., 2020). Unlike continuously measured physiological variables, AVPU is not automatically captured, limiting its integration into high-frequency monitoring systems.

### 3.6 AVPU in the NEWS Score

A real-world prehospital sample validation

In Portugal patient assessment is performed by EMS according to standardized clinical protocols. INEM uses the NEWS score. Physiological parameters are recorded using portable monitoring devices, results are registered in iTeams® and the NEWS score is automatically calculated using the recorded vital signs and the on-scene professional assessment of consciousness (AVPU).

An iTeams Data sample, anonymized and fully de-identified, was collected in a retrospective observational design from the first 6 month of 2024 of adult, non-pregnant, patients attended by EMS teams with recorded NEWS score.

A quantitative analytical approach was used to snap-shot the relationship between physiological parameters (respiratory rate, heart rate, oxygen saturation, use of supplemental oxygen, systolic blood pressure, body temperature), level of consciousness (AVPU), and deterioration risk classification measured by NEWS. The AVPU scale was operationalized as a binary variable (A vs VPU) and the moderate or high risk of clinical deterioration defined as NEWS ≥ 5.

The analysis showed a total of 111,202 EMS events. Most patients were classified as Alert (AVPU = A) at initial assessment (90.6%), while 10,454 (9.4%) presented with impaired level of consciousness (Verbal, Pain, or Unresponsive; V/P/U). The overall distribution of NEWS values showed a median score of 2 (IQR 1–4). Most patients were classified as low risk (NEWS ≤ 5) (82.7%), while 19,248 (17.3%) met criteria for moderate or high deterioration risk (NEWS ≥ 5). (Table F.1)

A marked difference in deterioration risk classification was observed according to level of consciousness. Among alert patients, the majority were classified as low risk, with approximately 86% presenting NEWS < 5. In contrast, among patients with impaired consciousness (V/P/U), nearly four out of five patients (≈79%) were classified as NEWS ≥ 5. This difference was highly statistically significant ($p < 0.001$) and demonstrated a strong crude association between impaired consciousness and elevated NEWS. Non-alert patients exhibited a substantially higher probability of being classified as moderate or high risk compared with alert patients. (Table F.2, F.3, F.4)

Despite the data set having properties that require careful handling in regression models namely the fact that AVPU is part of the NEWS score (predictor is a component of the outcome score) and as the scope of this analysis is just a snap shot of the AVPU-NEWS relation, a multivariable logistic regression model with NEWS ≥ 5 as the outcome and AVPU (V/P/U vs Alert) as the primary exposure was applied, adjusting for all physiological parameters, with prior standardization.

In the presented model the Impaired level of consciousness was strongly and independently associated with elevated NEWS classification. The adjusted odds of NEWS ≥ 5 were markedly higher among V/P/U patients compared with alert patients (adjusted odds ratio ≈ 48; $p < 0.001$), even after accounting for all other physiological parameters. The average marginal effect of impaired consciousness corresponded to an absolute increase of approximately 20 percentage points in the probability of being classified as NEWS ≥ 5, holding other variables constant. (Table F.5)

Predictive performance was evaluated using ROC curve analysis with elevated NEWS defined as NEWS ≥ 5. A model including physiological parameters alone demonstrated good discrimination (AUC = 0.864). When AVPU (binary A vs VPU) was added to the physiological model, discrimination improved further (AUC = 0.904), yielding an absolute increase of ΔAUC = 0.039. These results indicate that level of consciousness contributes additional clinically meaningful information to NEWS-based deterioration risk classification beyond vital signs alone.

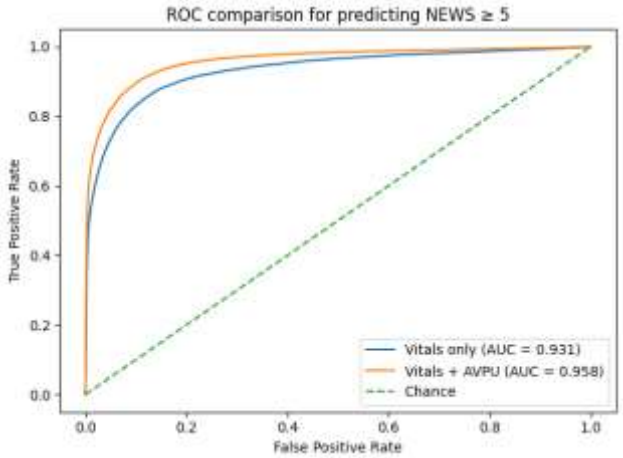



**Figure 3.1** - Comparison of Vitals only with vitals+ AVPU for predicting NEWS ≥ 5

The reclassification analysis demonstrated that inclusion of AVPU resulted in 3,072 patients (2.8% of the cohort) being up classified across the clinically relevant escalation threshold (NEWS < 5 to NEWS ≥ 5). No patients were down classified. These findings indicate that assessment of consciousness functions as a unidirectional modifier of risk classification, identifying additional patients who meet escalation criteria without reducing risk status in any case. From an operational perspective, incorporation of AVPU enhances detection of patients at moderate or high risk of clinical deterioration, particularly among those with borderline physiological abnormalities whose total NEWS would otherwise remain below the escalation threshold. (Table F.6)

Sensitivity analysis using alternative escalation thresholds confirmed the robustness of the AVPU contribution to deterioration risk stratification. At the threshold of NEWS ≥ 5, inclusion of AVPU resulted in upward reclassification of 3,072 patients (2.8% of the cohort), with no downward reclassification observed. When applying the higher threshold of NEWS ≥ 7, 3,123 patients (2.8%) were similarly up classified, representing a substantial proportion of those ultimately categorized as high acuity. These findings demonstrate that the impact of impaired level of consciousness on risk stratification remains consistent across clinically relevant escalation thresholds and predominantly influences patients positioned near decision boundaries. (Table F.7 and F.8)

**3.7 Vital signs to access Consciousness**

Are there reasonable patterns?

Consciousness level, operationalized in NEWS through AVPU, is traditionally considered a clinical neurological assessment rather than a sensor-derived parameter. However, physiological abnormalities frequently accompany altered awareness.

A study examining associations between vital signs and impaired consciousness demonstrated that hypotension, hypoxia, fever, and abnormal respiratory rate were strongly linked to reduced levels of awareness in emergency contexts (Ikeda et al., 2002). These findings support the hypothesis that constellations of vital-sign abnormalities may serve as indirect indicators of neurological compromise.

Physiological patterns vary by etiologic Intracranial pathology may produce hypertensive responses with bradycardia, whereas metabolic or toxic etiologies often present with hypotension or hypoxia (Ikeda et al., 2002). Such interactions illustrate that consciousness states reflect systemic physiological processes rather than isolated neurological events.

Trend-based modelling studies indicate that incorporating slopes and minimum systolic blood pressure values improves deterioration detection (Churpek et al., 2016). Continuous monitoring thus enables identification of evolving patterns—such as progressive hypoxia combined with rising respiratory rate—that may precede transition from “Alert” to “Voice” or “Pain” responsiveness.

Although AVPU remains a subjective categorical input, exploring whether reproducible physiological trajectories correlate with impaired consciousness represents a foundational step toward objective augmentation of neurological assessment within risk stratification systems.

### 3.8 From Threshold-Based Scores to Data-Driven Escalation Models

Traditional early warning scores operate through additive point systems with predefined thresholds. While transparent and clinically interpretable, they assume independence between parameters and ignore nonlinear interactions and individualized baselines.

Machine-learning–based early warning systems address these limitations by learning complex multivariate patterns directly from longitudinal data (Muralitharan et al., 2021). Gradient boosting approaches enhance discrimination while controlling alert burden (Romero-Brufau et al., 2025), and recurrent neural networks reduce false positives through sequence-aware modelling.

Hybrid deep-learning frameworks combining convolutional neural networks with dynamical systems modelling further demonstrate improved outcome prediction by explicitly encoding temporal structure (Lehman et al., 2015). These approaches signal a conceptual shift from episodic scoring toward continuous risk estimation, rather than replacing NEWS or AVPU, data-driven escalation models can be conceptualized as augmentation layers. Continuous physiological streams provide temporal context, while NEWS thresholds offer interpretable anchors for clinical action. Integration of these paradigms preserves transparency while enhancing sensitivity and timeliness.

### 3.9 Data models and Collaboration Context

Within this dissertation, the sensor-data modelling framework was developed in collaboration with Tomás Mendes. The analytical exploration of relationships between physiological parameters, AVPU status, and NEWS escalation thresholds provided the conceptual basis for future computational modelling. The present work contributes the clinical grounding of the dataset structure, definition of relevant physiological features, alignment with NEWS scoring intervals, and framing of deterioration thresholds. This collaboration was designed to ensure that dataset conceptualization and clinical interpretation reinforce methodological consistency, while computational development provides a structured pathway for potential future predictive applications.

### 3.10 Implications for Clinical Risk Stratification and Decision Support

The convergence of wearable sensors, electronic records, and advanced analytics enables reframing risk assessment as a continuous process rather than a sequence of discrete evaluations. Longitudinal modelling enhances differentiation between transient physiological fluctuations and sustained deterioration patterns (Breteler et al., 2020; Churpek et al., 2016).

High positive predictive value is essential to minimize alarm fatigue and maintain clinical trust (Romero-Brufau et al., 2025). Intelligent trend interpretation, rather than simple threshold crossing, is central to achieving this balance.

In prehospital environments, wireless monitoring systems may extend these capabilities beyond hospital settings, supporting earlier identification of high-risk patients and improved coordination between emergency teams and receiving institutions (Koceska et al., 2020). However, implementation must prioritize interpretability, transparency, and integration with existing escalation protocols. Data driven systems should augment, not obscure clinical judgment.

### 3.11 Research Gaps

Despite promising retrospective performance, machine-learning–based early warning systems lack prospective validation (Muralitharan et al., 2021). Methodological heterogeneity in prediction windows, outcome definitions, and evaluation metrics limits comparability across studies.

Irregular sampling, missing data, and signal artifact remain significant modelling challenges in biomedical time-series analysis (Patharkar et al., 2024). Additionally, integration of objective physiological modelling with neurological assessment frameworks such as AVPU remains underexplored.

Generalizability across care settings, particularly from inpatient to prehospital populations, requires dedicated validation studies (Romero-Brufau et al., 2025). Furthermore, advanced architectures such as deep-learning models must incorporate explainability mechanisms to ensure clinical acceptance and regulatory compliance. Future research should therefore focus on prospective validation, standardized evaluation protocols, and integration of continuous physiological modelling with established escalation systems such as NEWS, ensuring that data-driven innovations enhance safety, efficiency, and clinical relevance.

## Chapter 4 -NEWS automation in EMS

*Integrates and evaluates both approaches, discussing the clinical perspective relevance, proposing a* multimodal architecture model for EMS deployment and *highlighting privacy, ethical, governance and operational constraints.*

### 4.1 Integration of both approaches

This dissertation developed two complementary evidence streams to support the same operational objective: improving the reliability and interpretability of consciousness assessment within a NEWS-oriented escalation logic in the prehospital setting. The video-based branch focuses on observable responsiveness during structured EMS-like interactions. The sensor-data branch focuses on physiological state captured from monitoring devices and recorded in routine EMS workflows.

These modalities are not redundant. They capture different aspects of the clinical construct underlying AVPU. Video captures the behavioural manifestation of responsiveness to escalating stimuli, whereas physiological parameters capture systemic instability that may co-occur with altered consciousness and deterioration risk.

The integration proposed in this chapter does not assume that video can replace clinical examination, nor that physiology can fully explain consciousness state. Instead, the aim is to define an architecture in which physiological measurements continue to support NEWS computation and escalation thresholds, combined in an operationally feasible way for EMS and auditable for clinical governance.

Figure 4.1 illustrates the proposed multimodal architecture as a modular pipeline that integrates physiological monitoring and event-driven video assessment within a unified workflow. The system begins with data acquisition from standard EMS monitoring devices—capturing HR, RR, $SpO_2$, SBP, temperature, and oxygen supplementation—and short video clips recorded during structured responsiveness checks. These inputs undergo preprocessing and quality control to ensure plausibility and to flag low quality or incomplete segments before relevant physiological trends and video-derived responsiveness cues are extracted.

Two inference modules, a vital signs model and a video responsiveness model, generate NEWS-aligned risk outputs and a structured consciousness-support indicator with rationale flags, which are then aggregated and presented in a clinically interpretable format. The final outputs are delivered directly to EMS operational platforms, namely iTeams® in the field and CODU dashboards at the coordinating centre within the Portuguese EMS service (INEM).

A transversal governance layer spans the entire pipeline, incorporating data privacy and consent management, data minimisation, audit logging, and model oversight to ensure controlled, transparent, and accountable deployment.

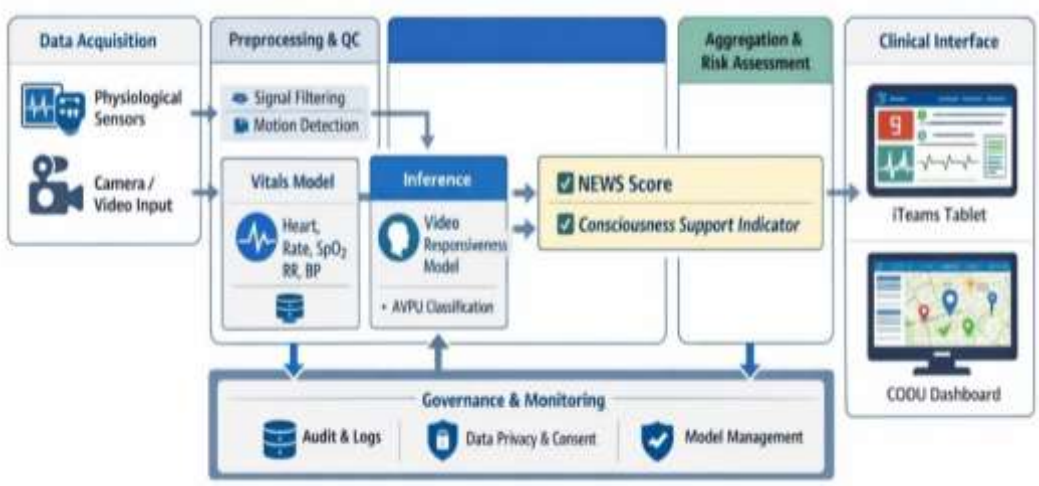



Figure 4.1 – A modular multimodal architecture integrating physiological monitoring and event-driven video consciousness assessment within a unified workflow. ChatGPT, Version 4.0, Open AI, 2 February 2026

### 4.2 Relevance from a clinical perspective

From a clinical perspective, the relevance of this work must be interpreted within the broader transformation of emergency medicine driven by digitalisation and artificial intelligence. The COVID-19 pandemic acted as a structural accelerator of digital adoption, shifting teleconsultation, virtual pathways, and data-driven surveillance from optional innovation to core healthcare infrastructure [(Baudier et al., 2023; Clipper, 2020; Getachew et al., 2023; Wosik et al., 2020). In this context, AI-enabled assessment is no longer framed as experimental enhancement but increasingly as a necessary component of resilient emergency systems. However, the literature consistently emphasises that technological expansion alone does not guarantee clinical benefit; sustained uptake depends on trust, perceived usefulness, workflow alignment, and organisational readiness (Olakotan et al., 2025)

Within emergency and prehospital care, the digitalisation of clinical records constitutes a foundational enabler of predictive analytics and automated decision support [(Sarwar et al., 2023; Zhang et al., 2020). Electronic health records (EHRs) aggregate structured and unstructured data at scale, enabling outcome modelling, though with well-documented challenges including missingness, coding variability, temporality bias, and governance constraints. Simultaneously, AI applications in EMS—ranging from dispatch optimisation to risk prediction—are expanding, yet remain constrained by retrospective design, limited external validation, privacy barriers, and integration difficulties

[(Elfahim et al., 2025) These limitations reinforce a central conclusion of the reviewed evidence: AI deployment in emergency contexts must be clinically grounded, transparent, and operationally integrated rather than technologically isolated (Topol, 2019).

Against this background, the multimodal architecture proposed in this dissertation represents a focused but meaningful advance within EMS digital transformation. It does not attempt to exploit the full analytical potential of EHR ecosystems, nor does it claim comprehensive automation of patient assessment. Instead, it addresses a high-impact but operationally variable clinical determinant - consciousness level within NEWS logic, by combining physiological monitoring data with structured video-derived responsiveness cues.

The sensor-data findings confirm that transitions from Alert to impaired responsiveness materially influence escalation classification, supporting existing evidence that altered mental status is a strong predictor of adverse outcomes under NEWS frameworks. The video-based modelling demonstrates that AVPU, often considered inherently subjective, can be decomposed into observable stimulus–response patterns that are annotatable and computationally interpretable under controlled conditions.

Clinically, this integration is coherent with the augmentation paradigm described in high-performance medicine literature, where AI systems are designed to support rather than replace clinician judgement in high-uncertainty environments. In prehospital, where decisions are made under time pressure and incomplete information, even incremental gains in reliability and interpretability of escalation triggers can influence downstream triage accuracy and resource allocation (Elfahim et al., 2025). Therefore, from a clinical relevance standpoint, this represents a concrete step toward clinically interpretable, governance aware, and workflow compatible decision support in emergency medical services**.**

### 4.3 System architecture and deployment in EMS

#### 4.3.1 Acquisition and operational triggers

The deployment of a multimodal architecture in EMS must be grounded in operational realism. In the prehospital environment, any additional technological layer that disrupts workflow, increases cognitive load, or prolongs on-scene time risks undermining its own clinical value. For this reason, acquisition within the proposed system is designed to build upon existing data streams rather than create new documentation tasks. Physiological parameters (heart rate, respiratory rate, oxygen saturation, systolic blood pressure, temperature, and supplemental oxygen status) are already routinely captured through standard EMS monitoring devices. These measurements are time-stamped

and transmitted through established electronic record infrastructures (e.g., iTeams®), forming the primary input to the vital signs' component of the architecture.

Video acquisition, by contrast, requires stricter constraints due to its sensitivity and governance implications. Rather than continuous recording, the system relies on short, event-driven clips triggered specifically during the structured responsiveness assessment. This targeted capture reflects the clinical moment in which AVPU is determined and avoids unnecessary data collection. By limiting recording to brief assessment windows and enforcing minimal retention policies with controlled access, the architecture adheres to a privacy-by-design principle (Van Rest et al., 2014) while maintaining clinical relevance.

Once acquired, both data streams enter a unified processing pipeline. Physiological signals undergo automated plausibility filtering, detection of missingness, and artefact identification to ensure that downstream computation is not distorted by sensor noise or transient instability. In parallel, video segments are evaluated for basic quality metrics, including lighting adequacy, motion blur, and occlusion that could compromise behavioural interpretation.

Only data meeting predefined quality thresholds proceed to inference; otherwise, the system explicitly flags insufficient quality and defaults to vital signals-only operation if acquisition constrains refer to video. If constrains occur on both data streams the determination of a NEWS score reverses to manual input.

Feature extraction and inference modules operate separately but in coordination, together they derive a NEWS-aligned risk features and compute an automated score consistent with established escalation logic. The physiological branch predicts an AVPU score based on vital signs automatically proposing a NEWS score. The video branch extracts structured stimulus–response indicators proposing a binary AVPU result: "A-alert" or "VPU-non alert", that incorporates respectively as a "zero" or "three" in the NEWS score. Importantly, the output of this video module is not a black-box categorical assignment but an interpretable consciousness-support indicator, preserving transparency and clinical interpretability.

These parallel outputs converge in an aggregation layer that maps results, into NEWS-oriented escalation classes. The system presents the automated NEWS score alongside the consciousness-support indicator and associated rationale flags within the EMS interface, supporting rather than supplanting clinician judgement.

An overarching audit layer records model versioning, input quality metrics, generated outputs, and user interactions, ensuring traceability and facilitating governance oversight.

In this configuration (Figure 4.1), the multimodal architecture remains tightly aligned with the existing NEWS workflow used in EMS, it does not introduce a competing assessment but augments an established escalation pathway through structured, quality-controlled data flow.

### 4.3.2 Preprocessing and quality control

Preprocessing and quality control are central to the safety and credibility of the proposed system. In emergency settings, sensor artefacts, environmental constraints, and time pressure are common, therefore, automated decision-support must explicitly recognise uncertainty rather than conceal it.

For physiological data, quality control mechanisms include range plausibility checks, identification of abrupt spikes inconsistent with physiological patterns, and explicit marking of missing or incomplete data. These steps ensure that NEWS computation reflects genuine patient state rather than measurement error.

For video-derived inputs, quality control is equally essential. Occlusion, insufficient illumination, excessive motion, or inability to reliably track relevant anatomical regions may invalidate behavioural interpretation. When such limitations are detected, the system does not attempt forced classification but instead signals insufficient quality and reverts to vitals-based assessment alone. This degraded-mode capability is not a limitation but a safety feature: it ensures that multimodal augmentation enhances clinical reliability without creating dependency on data streams that may be unavailable or unreliable in real-world EMS conditions.

### 4.3.3 Feature extraction and inference modules

The architecture separates the video and vital signs streams into two inference modules. The vital signs model produces risk features that summarize physiological instability by supporting NEWS computation and trend interpretation. This model infers the consciousness level (AVPU) from the physiological parameters and can be implemented as rule-based NEWS calculation alone, particularly as an augmentation layer that learns associations between multivariate physiological patterns and escalation thresholds (chapter 3).

Video responsiveness model produces a structured “responsiveness cue” output aligned with AVPU logic, classifying patients, as presently done by EMS technicians, as “alert” versus nonresponsive. This binary proposed AVPU result integrate NEWS to compute a score. (Chapter 2).

The video module should output the class label but also a concise rationale that maps to the AVPU examination logic, namely “A – alert, “V” - response after verbal stimulus,” “P” - response only after pain stimulus,” or “U” - no response detected within clip”. This aligns the computational output with the clinician’s mental model and supports trust calibration.

#### 4.3.4 Aggregation and explanation (NEWS + consciousness support)

Aggregation occurs at the operational level used by EMS namely in the assessment of the NEWS score with its escalation thresholds.

The vital signs module proposes a NEWS result with a probable AVPU derived from the physiologic trend. The video module proposes an AVPU result through feature detection to stimulus, shown separately. The integration module combines both results to a final NEWS result. Clinical validation of NEWS output result determines acceptance.

If any or both modules present uncertainty, the manual input of NEWS parameters remains the definitive reference, ensuring that clinician judgement overrides automated outputs and that escalation decisions continue to rely on validated clinical assessment rather than solely on algorithmic inference.

The integrated system should present:

- NEWS score in the same manner used today, to preserve workflow familiarity, referring to the way it was obtained in a label (V & V - Vitals and video, VS - vital signs only, M - manual input) (Figure 11)

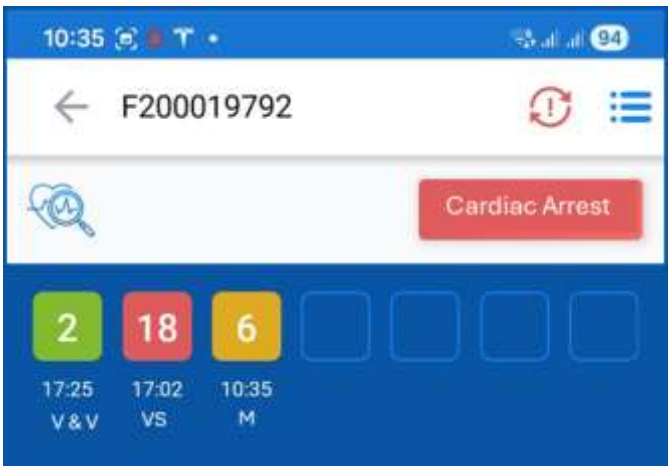



Figure 4.2 – iTeams® Dashboard adapted to NEWS score obtained from different pathways

- Consciousness-support indicator, namely the AVPU obtained through video, showing the observed responsiveness “A”, “V”, “P” and “U” (Figure 12)

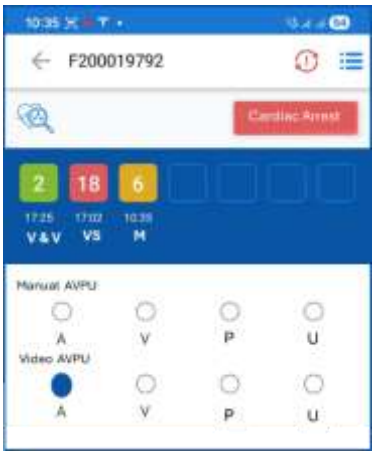



Figure 4.3- iTeams® Dashboard adapted to Video AVPU

- Confidence and quality flags to avoid over-interpretation (e.g., “video quality low,” “occlusion present,” “insufficient duration”). (Figure 13)




Figure 4.4 – iTeams® Dashboard example of a flagged quality indicator

### 4.3.5 Edge versus cloud deployment

In EMS environments, inference location affects connectivity requirements, privacy exposure, and operational reliability. Edge inference (on-device or in-vehicle) reduces dependency on network connectivity and allows video processing without transmitting identifiable raw footage. Cloud

inference supports heavier models and centralized oversight but introduces latency and transmission risks.

A pragmatic hybrid approach is recommended: compute video-derived cues on the edge and transmit summary outputs to CODU (Haider et al., 2026; Madhavi Latha Vadlamudi, 2025). Namely NEWS outputs, as currently done, adding the information whether it was inferred form Vital Signs, Vital Signs + Video or Vital Signs + manual insertion.

#### 4.3.6 Latency and degraded modes

Decision support in prehospital triage must operate within clinically meaningful time windows. The proposed system therefore prioritizes near-real-time inference, with target latency bounded by the assessment cycle (seconds rather than minutes). Latency contributors include sensor sampling frequency, preprocessing time (artifact detection), model runtime, and network uplink delays when transmitting to CODU.(Fukaguchi et al., 2022)

The architecture explicitly supports degraded modes. If video is unavailable or low quality, the system falls back to vitals-only logic, if connectivity is lost, local computation continues and synchronizes logs when network becomes available.

### 4.4 Ethical, privacy and governance considerations

The proposed multimodal architecture, combining physiological monitoring with event-driven video assessment of responsiveness, operates within a particularly sensitive ethical and legal context. In emergency medical services (EMS), decision-support technologies are deployed in situations marked by urgency, patient vulnerability, and frequently reduced autonomy. Ethical robustness is therefore not optional but foundational to legitimacy. The governance framework must align with European standards on patients' rights(Duwell & Mieth, 2000), quality and safety of care, transparency, and non-discrimination, while remaining operationally feasible in prehospital environments. Importantly, the system augments NEWS-based escalation without displacing clinician judgement, thereby preserving professional accountability and avoiding algorithmic substitution of clinical responsibility.

From a data protection perspective, the architecture processes special category data, including health data and potentially identifiable video data, requiring strict compliance with the General Data Protection Regulation (GDPR)(European Union, 2016). Core GDPR principles as lawfulness, purpose limitation, data minimisation, integrity, and accountability are translated into design decisions. Therefore, video capture is event-triggered rather than continuous, retention is limited to the minimum necessary period, inference may occur at the edge to reduce transmission of identifiable

data, and comprehensive audit logging ensures traceability. Data minimisation is thus embedded structurally in the acquisition logic rather than treated as a purely declarative policy.

Transparency, explainability, and inclusivity further structure the governance approach. In line with European ethical principles for digital health(Seroussi & Zablit, 2024), individuals should be informed about how digital systems function, how data is processed, and when artificial intelligence contributes to assessment. The consciousness-support module therefore produces structured stimulus–response rationales aligned with AVPU logic, accompanied by quality and confidence indicators to prevent over-reliance on uncertain outputs. Inclusivity requires ongoing monitoring for bias and performance disparities across demographic groups, ensuring that visual or physiological variability does not disadvantage specific populations.

Accountability and oversight are equally central. Although positioned as decision-support, automated NEWS computation and video-derived AVPU input may materially influence escalation and triage decisions. Clear delineation of responsibility is therefore essential. Clinicians remain final decision-makers, algorithmic outputs are advisory, and system logs record model versions, input quality metrics, and user interactions. Proportionality and purpose limitation must guide deployment, restricting use strictly to clinical assessment unless additional legal bases and safeguards are established for research or model optimisation.

Finally, operational adoption would require a formal Data Protection Impact Assessment (DPIA), given the high-risk nature of processing health and video data under GDPR. A distinction must be made between routine vital signs monitoring that are justified under the provision of healthcare and protection of vital interests and video capture, which introduces heightened sensitivity. In cases where patients are unconscious and consent cannot be obtained, processing may be ethically and legally justified only if it is necessary for immediate care, minimally intrusive, and governed by strict retention, access control, and audit mechanisms. Where capacity is regained, retrospective transparency should be ensured. Under these conditions, multimodal decision-support in compromised consciousness can be ethically defensible, provided governance remains intrinsic to system design rather than an external compliance add-on.

### 4.6 Operational constraints and implementation Challenges

Prehospital emergency medical services operate in environments defined by unpredictability, urgency, and limited environmental control. Patient assessment frequently occurs in homes, public spaces, roadside incidents, or moving ambulances, where lighting conditions, ambient noise, confined spaces, and continuous motion can compromise both physiological signal stability and video quality. Motion artefacts, transient sensor displacement, and occlusions during patient handling are common.

Additionally, rapid transitions between scene assessment and transport introduce discontinuities in acquisition. Any digital architecture intended for EMS must therefore be inherently robust to incomplete, noisy, or intermittent data streams and explicitly designed to tolerate degraded input conditions rather than assuming hospital-level stability.

Operational workload represents an equally significant constraint. EMS staff performs multiple tasks, from patient assessment to emergency intervention (e.g., CPR), monitoring, documentation, communication with coordinating centres and transport, all within compressed time windows. Introducing additional recording steps or complex interfaces risks increasing cognitive load and documentation burden. (Smith et al., 2022) Evidence from digital health implementation studies indicates that systems perceived as intrusive or misaligned with workflow may contribute to fatigue and resistance rather than safety gains. (Stachwitz & Debatin, 2023) For this reason, the proposed multimodal model is deliberately aligned with existing NEWS-based escalation logic, avoiding parallel assessment pathways and ensuring that outputs are integrated into familiar decision structures rather than adding procedural layers.

Technical infrastructure variability further complicates implementation. Prehospital connectivity may fluctuate depending on geography, network coverage, and environmental interference, making cloud-dependent designs vulnerable to latency or temporary inaccessibility. A pragmatic architecture must therefore privilege edge-based processing for time-sensitive components, while transmitting only summarised outputs when connectivity allows. Conservative design principles are essential, and alerting mechanisms must be calibrated to avoid excessive or low-specificity triggers that could generate alert fatigue and erode EMS staff trust.

Beyond operational constraints, successful implementation depends on organisational readiness and user acceptance. Clinical environments are often resistant to changes that introduce perceived complexity or uncertainty. Adoption will depend on demonstrable usefulness, ease of use, adequate training, leadership endorsement, and iterative pilot testing with structured feedback loops.

Financial considerations, from hardware acquisition, software integration, to training and ongoing support, must ultimately be justified by measurable benefits in patient safety, workflow efficiency or risk assessment improvement. Therefore, deployment should follow a staged pathway with initial feasibility and data quality validation under real EMS conditions, followed by prospective observational evaluation, and only subsequently broader impact assessment. Operational realism, rather than technological ambition, remains the decisive determinant of responsible integration into prehospital care.

## Chapter 5 - Conclusions and future work

*Concludes the dissertation by summarizing the main achievements, articulating limitations, and outlining future work toward a real-world multimodal system for EMS implementation.*

### 5.1 Summary of the main achievements

This dissertation investigated the operational significance of level of consciousness within prehospital deterioration risk stratification, adopting the National Early Warning Score (NEWS) as the reference escalation framework and the AVPU scale (Alert, Verbal, Pain, Unresponsive) as the clinical method for consciousness classification. The work was motivated by a central tension in emergency medical services (EMS): consciousness is one of the most prognostically influential components of early warning scores, yet its assessment remains observational, context-dependent, and particularly vulnerable to variability under field conditions. The thesis therefore sought to move from this conceptual and operational challenge toward structured methodological clarification and system-level integration.

The first achievement addressed the problem at its behavioural root: how AVPU assessment can be operationalised in a reproducible and computationally analysable manner. A controlled, clinically grounded video dataset was constructed using scripted responsiveness scenarios aligned with the AVPU examination pathway. Rather than simulating clinical deterioration, the dataset functioned as a methodological feasibility platform, translating stimulus–response logic into standardised acquisition, annotation, and inter-rater agreement procedures. Importantly, AVPU categories were operationally collapsed into the distinction most relevant within NEWS logic—Alert versus impaired responsiveness—thereby directly linking behavioural observation to escalation frameworks.

Building on this behavioural formalisation, the second achievement clarified the role of consciousness within NEWS determination from a physiological perspective. While NEWS is conventionally derived from vital signs, the dissertation highlighted that consciousness is not an ancillary variable but a decisive escalation determinant. Analysis of Portuguese EMS data reinforced how strongly AVPU status influences NEWS classification, empirically demonstrating its impact on risk stratification. This opened a structured investigation pathway: exploring the relationship between physiological trends and AVPU categories and examining how vital signs analysis can inform or support the consciousness component within NEWS computation.

The third achievement connected these two strands, behavioural modelling and physiological analysis, into a coherent system-level integration pathway. Rather than proposing automation that replaces clinical judgement, the dissertation defined a decision-support architecture aligned with

existing EMS workflows. In this model, physiological monitoring continues to drive standard NEWS computation, while structured vital signs analysis and video-derived responsiveness cues provide transparent support for the consciousness domain. The resulting framework emphasises modularity, interpretability, degraded modes, and governance safeguards, establishing not a deployable product, but a clinically grounded foundation for future validation and responsible integration of multimodal decision-support in prehospital emergency care.

**5.2 Contributions to clinical practice and digital innovation**

Consciousness assessment represents a pivotal determinant of deterioration escalation in prehospital care. Although its importance is well established conceptually within emergency medicine, this dissertation advances the discussion by clarifying and structurally framing how impaired AVPU status influences escalation classification under NEWS logic. By making explicit the weight attributed to consciousness within the score, the work highlights why variability in AVPU assessment can materially affect risk stratification. This perspective redirects attention toward the contexts in which observational inconsistency may arise and where structured support or standardisation could improve reliability. Furthermore, the dissertation reframes AVPU not as a static categorical label but as a structured stimulus–response process, thereby strengthening interpretability by aligning computational modelling with the clinical examination logic used in practice.

From a digital innovation standpoint, the contribution lies in proposing a feasible and conservative pathway toward multimodal decision support in prehospital environments. The development of a clinically grounded video dataset, a structured annotation framework, and inter-rater reliability procedures establishes methodological foundations that can be extended in future work. The proposed architecture demonstrates how physiological analytics and video-derived responsiveness cues can be integrated without displacing clinician judgement, ensuring that outputs remain interpretable and consistent with established NEWS escalation language. This approach reflects safety-oriented digital health principles, prioritising consistency, clarity, and workflow alignment over premature automation of complex clinical reasoning.

A practical implication concerns documentation practices. Manual AVPU recording, particularly in high-pressure scenarios, contributes to cognitive load and may be incomplete or inconsistently captured. A structured decision-support system that automates the detection and provisional recording of consciousness level—while preserving clinician validation—has the potential to reduce administrative burden and improve data completeness. Standardised capture mechanisms may enhance consistency across providers and care transitions, supporting reliable NEWS computation and clearer communication between teams.

By strengthening the reliability and continuity of consciousness assessment, such an approach may also improve early recognition of clinical change. Consistent and structured documentation of responsiveness facilitates timely detection of deterioration and supports appropriate escalation decisions. Rather than replacing clinical expertise, the system aims to reinforce situational awareness and reduce ambiguity in high-pressure environments, thereby contributing to safer and coherent prehospital risk stratification.

**5.3 Study limitations**

Limitations concerns may rise from the conceptual and structural boundaries of NEWS and AVPU as the last is part of the former.

In emergency settings the assessment of responsiveness / consciousness has already proven its independent prognostic value, but it also remains the only qualitative bio signal used transversely in emergency, from early warning scores to specific scores trauma scores as the revised trauma score (RTS). A word to say is that it is not possible to dissociate AVPU scale and emergency scores.

The video-based component presents methodological limitations inherent to controlled camera-based data acquisition. The dataset was constructed using scripted scenarios and simulated responsiveness performed by healthcare-related participants under standardized lighting, positioning, and recording conditions. Although ensuring reproducibility and reliable annotation of stimulus–response patterns, it does not replicate the variability of real prehospital environments. In EMS settings, responsiveness assessment may be influenced by patient comorbidities, intoxication, agitation, severe pain, communication barriers, or movement during transportation. Also, environmental constraints such as poor lighting, occlusion, camera instability, and dynamic scene conditions may compromise video quality. Therefore, the dataset supports feasibility testing of behavioural modelling but cannot be interpreted as evidence of field-ready automation of AVPU assessment.

Limitations also arise within the physiological data component as despite physiological parameters being objectively measured, prehospital datasets are subject to irregular sampling, transient artefacts, missingness, and operational variability. While exploring the association between physiological patterns and impaired consciousness, it does not address demonstration, as such, the physiological modelling should be interpreted only as exploratory and foundational of automated consciousness prediction.

The proposed modular multimodal architecture integrating video-derived responsiveness and physiological analytics remains conceptual and has not undergone prospective validation. Although

the system is designed with degraded modes, quality flags, and clinician override mechanisms, its real-world performance under operational stress remains untested. Critical properties such as latency, robustness to incomplete inputs, false-alert burden, and interaction between modules can became constrains in live EMS workflows.

Also, implementation constraints must be acknowledged as prehospital emergency care operates in unpredictable, high-pressure environments with limited environmental control, fluctuating connectivity, and substantial cognitive workload. The introduction of additional technological layers can increase documentation burden or alert fatigue if not carefully aligned with workflow. Moreover, issues of clinician trust, organisational readiness, training requirements, financial cost, and governance compliance represent substantial barriers to adoption. Therefore, any real-world deployment would require staged feasibility testing, prospective evaluation, and careful operational integration before consideration of routine use.

**5.4 Recommendations for future research**

The work presented in this dissertation establishes a structured methodological and architectural foundation, however, its real-world impact depends on rigorous future validation and progressive refinement. Several research trajectories emerge directly from the limitations and achievements identified throughout the thesis, including the continuation of work currently being developed by Marco Esperança and Tomás Mendes within their respective technical domains.

The video-based modelling work led by Marco Esperança represents a critical development pathway. Building on the controlled dataset and annotation framework defined in this dissertation, ongoing work focuses on refining computer vision architectures capable of detecting structured responsiveness cues aligned with AVPU logic. This includes experimentation with temporal sequence models, landmark-based feature encoding, and optimization of inference pipelines suitable for edge deployment in EMS environments. Future research should expand training data toward semi-controlled or pilot real-world recordings, evaluate robustness under environmental variability (lighting, motion, occlusion), and assess model explainability to ensure alignment with clinician reasoning. External validation across diverse demographic and operational contexts will be essential to strengthen generalisability and mitigate bias.

The physiological modelling framework under the work of Tomás Mendes constitutes another key research direction. The exploratory analyses presented in this thesis clarified the structural impact of AVPU within NEWS and reviewed the association between physiological parameters and impaired responsiveness. Ongoing computational work aims to model multivariate vital-sign trajectories using longitudinal and machine-learning approaches to better characterise deterioration dynamics. Future

research should investigate whether trend-aware models can anticipate transitions in consciousness state before overt behavioural impairment occurs and whether such models improve early detection performance relative to static NEWS computation alone. Advanced architectures capable of modelling temporal dependencies and inter-variable interactions may enable sensitive and individualized risk estimation.

The proposed multimodal architecture should be evaluated under operational conditions to assess latency, robustness to degraded inputs, workflow compatibility, false-alert burden, and clinician acceptance. Such studies should measure not only technical performance (e.g., agreement with clinician-assigned AVPU and NEWS) but also operational indicators, including scene time, documentation completeness, escalation appropriateness, and perceived cognitive load. Outcome-oriented studies should further explore associations with downstream clinical endpoints such as ICU admission, hospital mortality, and time-to-intervention.

An additional avenue for future research concerns the exploration of broader multimodal fusion strategies beyond the current integration of video and physiological parameters. In particular, the inclusion of audio-derived features such as speech presence, response latency, vocal coherence, tone variation, or detection of verbal engagement that may provide complementary indicators of responsiveness that are directly aligned with AVPU examination logic. Furthermore, structured integration with EHR data, including baseline cognitive status, prior neurological conditions, medication history, recent hospitalisations, or comorbidity profiles, could contextualise real-time assessments and reduce misclassification risk in patients with chronic impairment or pre-existing altered mental status. Future multimodal fusion models should therefore investigate hierarchical or attention-based architectures capable of combining physiological trends, video-derived cues, audio signals, and structured EHR data while preserving interpretability and maintaining alignment with NEWS-based escalation logic.

Also, real-world EMS pilot studies are needed to capture natural variability, including diverse age groups, comorbidities, intoxication states, agitation, language barriers, and environmental complexity during transport. Multi-centre collaborations would enhance representativeness and support fairness auditing across demographic subgroups.

Initial feasibility pilots should prioritise data quality monitoring and workflow integration before escalation to large-scale outcome evaluation. Economic assessment, including cost-effectiveness and resource allocation impact, should be incorporated in later phases to determine long-term sustainability.

## 5.5 Conclusion

This dissertation examines the role of consciousness assessment in the early recognition of clinical deterioration within prehospital emergency medical services. By positioning the AVPU scale within the framework of the National Early Warning Score (NEWS), the work clarifies how changes in alertness function as decisive escalation determinants in risk stratification. Although this relationship is well recognised in emergency medicine, AVPU assessment in the field remains observational, context-sensitive, and vulnerable to variability under operational pressure. Addressing this tension formed the conceptual core of the thesis.

To respond to this challenge, the dissertation proposed a structured, modular multimodal architecture that integrates video-derived responsiveness cues with physiological analytics while remaining aligned with established EMS workflows.

Rather than introducing a parallel assessment system, the proposed model preserves the primacy of NEWS-based escalation logic and clinician judgement. It reframes AVPU not as a static categorical label but as a structured stimulus–response process, enabling computational exploration without detaching the model from clinical reasoning.

The dissertation does not claim readiness for full automation of consciousness assessment nor immediate clinical deployment. Its contribution is foundational defining methodological, analytical and architectural principles that make future validation feasible.

The thesis advocates for a measured and responsible evolution in emergency medicine, going from episodic, manually documented observations toward structured, data-informed, and workflow-aligned decision support. The framework presented here offers a coherent pathway for future prospective validation and careful implementation, supporting the safe and ethically grounded integration of digital innovation into prehospital deterioration risk stratification.

**Bibliographic References and Sources**


Abbott, T. E. F., Cron, N., Vaid, N., Ip, D., Torrance, H. D. T., & Emmanuel, J. (2018). Pre-hospital National Early Warning Score ( NEWS ) is associated with in- hospital mortality and critical care unit admission : A cohort study. *Annals of Medicine and Surgery*, *27*(December 2017), 17–21. https://doi.org/10.1016/j.amsu.2018.01.006

Abd El-Aziz, R. M., & Rayan, A. (2025). Early detection of sepsis using machine learning algorithms. *Alexandria Engineering Journal*, *111*. https://doi.org/10.1016/j.aej.2024.10.005

Al-Tekreeti, Z., Moreno-Cuesta, J., Madrigal Garcia, M. I., & Rodrigues, M. A. (2024). AI-Based Visual Early Warning System. *Informatics*, *11*(3). https://doi.org/10.3390/informatics11030059

Barros F, Lourenço J. Nunes J. (2021). Novo Verbete Nacional de Socorro. *LIFE SAVING*. https://issuu.com/lifesaving/docs/revista_lifesaving_n19_final_low/122

Baudier, P., Kondrateva, G., Ammi, C., Chang, V., & Schiavone, F. (2023). Digital transformation of healthcare during the COVID-19 pandemic: Patients' teleconsultation acceptance and trusting beliefs. *Technovation*, *120*. https://doi.org/10.1016/j.technovation.2022.102547

Bhaumik, S., Hannun, M., Dymond, C., DeSanto, K., Barrett, W., Wallis, L. A., & Mould-Millman, N. K. (2022). Prehospital triage tools across the world: a scoping review of the published literature. In *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* (Vol. 30, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s13049-022-01019-z

Bloch, E., Rotem, T., Cohen, J., Singer, P., & Aperstein, Y. (2019). Machine Learning Models for Analysis of Vital Signs Dynamics: A Case for Sepsis Onset Prediction. *Journal of Healthcare Engineering*, *2019*. https://doi.org/10.1155/2019/5930379

Bourke-Matas, E., Doan, T., Bowles, K. A., & Bosley, E. (2024). A prediction model for prehospital clinical deterioration: The use of early warning scores. *Academic Emergency Medicine*. https://doi.org/10.1111/acem.14963

Bourke-Matas, E., Doan, T., Bowles, K. A., & Bosley, E. (2025). Prevalence of clinical deterioration in the pre-hospital setting. *EMA - Emergency Medicine Australasia*, *37*(1). https://doi.org/10.1111/1742-6723.14511

Breteler, M. J. M., KleinJan, E., Numan, L., Ruurda, J. P., Van Hillegersberg, R., Leenen, L. P. H., Hermans, M., Kalkman, C. J., & Blokhuis, T. J. (2020). Are current wireless monitoring systems capable of detecting adverse events in high-risk surgical patients? A descriptive study. *Injury*, *51*. https://doi.org/10.1016/j.injury.2019.11.018

Burgos-Esteban, A., Gea-Caballero, V., Marín-Maicas, P., Santillán-García, A., Cordón-Hurtado, M. de V., Marqués-Sule, E., Giménez-Luzuriaga, M., Juárez-Vela, R., Sanchez-Gonzalez, J. L., García-Criado, J., & Santolalla-Arnedo, I. (2022). Effectiveness of Early Warning Scores for Early Severity Assessment in Outpatient Emergency Care: A Systematic Review. In *Frontiers in Public Health* (Vol. 10). https://doi.org/10.3389/fpubh.2022.894906

Cainer, T. L., Bagshaw, A. P., & Cavanna, A. E. (2024). Mental imagery paradigms for the assessment of awareness in patients with disorders of consciousness: a systematic literature review. In *Journal of Psychopathology* (Vol. 30, Number 3). https://doi.org/10.36148/2284-0249-N565

Choi, D. H., Hong, K. J., Do Shin, S., Kim, S., Chung, M., Kim, K. H., Song, K. J., Cho, M., Yoon, D., & Lee, J. (2023). Measurement of level of consciousness by AVPU scale assessment system based on automated video and speech recognition technology. *American Journal of Emergency Medicine*, *74*. https://doi.org/10.1016/j.ajem.2023.09.044

Churpek, M. M., Adhikari, R., & Edelson, D. P. (2016). The value of vital sign trends for detecting clinical deterioration on the wards. *Resuscitation*, *102*. https://doi.org/10.1016/j.resuscitation.2016.02.005

Clipper, B. (2020). The Influence of the COVID-19 Pandemic on Technology: Adoption in Health Care. *Nurse Leader*, *18*(5). https://doi.org/10.1016/j.mnl.2020.06.008

Donoghue, A. J., Auerbach, M., Banerjee, A., Blewer, A. L., Cheng, A., Kadlec, K. D., Lin, Y., Diederich, E., Sawyer, T., Stallings, D. T., Toft, L. E. B., Torman, D., Wright, J. I., Schexnayder, S. M., & Dainty, K. N. (2025). Part 12: Resuscitation Education Science: 2025 American Heart Association Guidelines for Cardiopulmonary Resuscitation and Emergency Cardiovascular Care. *Circulation*, *152*(16 2). https://doi.org/10.1161/CIR.0000000000001374

Downey, C. L., Tahir, W., Randell, R., Brown, J. M., & Jayne, D. G. (2017). Strengths and limitations of early warning scores: A systematic review and narrative synthesis. In *International Journal of Nursing Studies* (Vol. 76). https://doi.org/10.1016/j.ijnurstu.2017.09.003

Duwell, M., & Mieth, D. (2000). The Charter of Fundamental Rights of the European Union. In *Biomedical Ethics* (Vol. 5, Number 2). https://doi.org/10.54648/337857

E., M., A., D. B., M., van de S., M., D., L., A., R., B., & R., A. B. (2019). Implementation of an automated early warning scoring system in a surgical ward: Practical use and effects on patient outcomes. *PLoS ONE*, *14*(5).

Elfahim, O., Edjinedja, K. L., Cossus, J., Youssfi, M., Barakat, O., & Desmettre, T. (2025). A Systematic Literature Review of Artificial Intelligence in Prehospital Emergency Care. In *Big Data and Cognitive Computing* (Vol. 9, Number 9). https://doi.org/10.3390/bdcc9090219

Elgendy, F., Alshewimy, M., & Sarhan, A. (2021). Pain detection/classification framework including face recognition based on the analysis of facial expressions for e-health systems. *International Arab Journal of Information Technology*, *18*(1). https://doi.org/10.34028/iajit/18/1/14

Essahraui, S., Lamaakal, I., El Hamly, I., Maleh, Y., Ouahbi, I., El Makkaoui, K., Filali Bouami, M., Pławiak, P., Alfarraj, O., & Abd El-Latif, A. A. (2025). Real-Time Driver Drowsiness Detection Using Facial Analysis and Machine Learning Techniques. *Sensors*, *25*(3). https://doi.org/10.3390/s25030812

European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation – GDPR). *Official Journal of the European Union*, *EN*(679).

Forster, S., McKeever, T. M., & Shaw, D. (2022). Effect of implementing the NEWS2 escalation protocol in a large acute NHS trust: a retrospective cohort analysis of mortality, workload and ability of early warning score to predict death within 24 hours. *BMJ Open*, *12*(11). https://doi.org/10.1136/bmjopen-2022-064579

Fukaguchi, K., Goto, T., Yamamoto, T., & Yamagami, H. (2022). Experimental Implementation of NSER Mobile App for Efficient Real-Time Sharing of Prehospital Patient Information With Emergency Departments: Interrupted Time-Series Analysis. *JMIR Formative Research*, *6*(7). https://doi.org/10.2196/37301

Getachew, E., Adebeta, T., Muzazu, S. G. Y., Charlie, L., Said, B., Tesfahunei, H. A., Wanjiru, C. L., Acam, J., Kajogoo, V. D., Solomon, S., Atim, M. G., & Manyazewal, T. (2023). Digital health in the era of COVID-19: Reshaping the next generation of healthcare. *Frontiers in Public Health*, *11*. https://doi.org/10.3389/fpubh.2023.942703

Haider, E., Mujahid, M. U. F., & Hidig, S. M. (2026). Harnessing hybrid edge-fog computing framework for emergency medical response management. *Annals of Medicine & Surgery*, *88*(1). https://doi.org/10.1097/ms9.0000000000004495

Haldar, M., Verma, A., Jaiswal, S., & Sheikh, W. R. (2020). Interrater reliability of four neurological scales for patients presenting to the emergency department. *Indian Journal of Critical Care Medicine*, *24*(12). https://doi.org/10.5005/jp-journals-10071-23603

Hill, M., Moreda, M., Navarro, J., & Mulkey, M. (2023). Assessing Patients With Altered Level of Consciousness. *Critical Care Nurse*, *43*(4). https://doi.org/10.4037/ccn2023449

Hillen, M. A., Visser, L. N. C., Labrie, N. H. M., van Vliet, L. M., Saha, S., Blanch-Hartigan, D., & Smets, E. M. A. (2025). Development of GROVE: A Guideline for RepOrting Vignette Experiments conducted in a healthcare context. *Patient Education and Counseling*, *136*. https://doi.org/10.1016/j.pec.2025.108750

Ikeda, M., Matsunaga, T., Irabu, N., & Yoshida, S. (2002). Using vital signs to diagnose impaired consciousness: Cross sectional observational study. *British Medical Journal*, *325*(7368). https://doi.org/10.1136/bmj.325.7368.800

Inkeaw, P., Srikummoon, P., Chaijaruwanich, J., Traisathit, P., Awiphan, S., Inchai, J., Worasuthaneewan, R., & Theerakittikul, T. (2022). Automatic Driver Drowsiness Detection Using Artificial Neural Network Based on Visual Facial Descriptors: Pilot Study. *Nature and Science of Sleep*, *14*. https://doi.org/10.2147/NSS.S376755

Isabel, A., Figueira, R., & Pereira, M. (2020). *Projetar Enfermagem-Revista Científica de Enfermagem Avaliação da pessoa em situação crítica: Aplicação do National Early Warning Score Critically ill patient assessment: Application of the National Early Warning Score*.

Janagama, S. R., Newberry, J. A., Kohn, M. A., Rao, G. V. R., Strehlow, M. C., & Mahadevan, S. V. (2022). Is AVPU comparable to GCS in critical prehospital decisions? – A cross-sectional study. *American Journal of Emergency Medicine*, *59*. https://doi.org/10.1016/j.ajem.2022.06.042

Karunarathna, I., Aluthge, P., Jayawardana, A., Fernando, C., Vidanagama, U., Ekanayake, U., Hapuarachchi, T., Senanayaka, S., Gunasena, P., Gunawardana, K., Rajapaksha, S., Warnakulasooriya, A. T., Athulgama, P., Abeyek, M., Dius, S., Ranwala, R., Godage, S., Senadeera, S., Wijepala, J., … Png, R. (2025). The AVPU Scale: A Rapid and Practical Tool for Assessing Consciousness in Clinical and Emergency Settings. *Uva Clinical Research Lab*.

Kazazian, K., Monti, M. M., & Owen, A. M. (2025). Functional neuroimaging in disorders of consciousness: towards clinical implementation. In *Brain* (Vol. 148, Number 7). https://doi.org/10.1093/brain/awaf075

Koceska, N., Komadina, R., Simjanoska, M., Koteska, B., Strahovnik, A., Jošt, A., Maček, R., Madevska-Bogdanova, A., Trajkovik, V., Tasič, J. F., & Trontelj, J. (2020). Mobile wireless monitoring system for prehospital emergency care. *European Journal of Trauma and Emergency Surgery*, *46*(6). https://doi.org/10.1007/s00068-019-01130-4

Kortli, Y., Jridi, M., Al Falou, A., & Atri, M. (2020). Face recognition systems: A survey. In *Sensors (Switzerland)* (Vol. 20, Number 2). https://doi.org/10.3390/s20020342

Kwon, J. M., Lee, Y., Lee, Y., Lee, S., & Park, J. (2018). An algorithm based on deep learning for predicting in-hospital cardiac arrest. *Journal of the American Heart Association*, *7*(13). https://doi.org/10.1161/JAHA.118.008678

Leandro Luis. (2014). *Tradução, validação e aplicação dos sistemas de pontuação de alerta precoce "VIEWS" e "NEWS" em Portugal*.

Leenen, J. P. L., Leerentveld, C., van Dijk, J. D., van Westreenen, H. L., Schoonhoven, L., & Patijn, G. A. (2020). Current evidence for continuous vital signs monitoring by wearable wireless devices in hospitalized adults: Systematic review. In *Journal of Medical Internet Research* (Vol. 22, Number 6). https://doi.org/10.2196/18636

Lehman, L. W., Ghassemi, M., Snoek, J., & Nemati, S. (2015). Patient prognosis from vital sign time series: Combining convolutional neural networks with a dynamical systems approach. *Computing in Cardiology*, *42*. https://doi.org/10.1109/CIC.2015.7411099

Li, L., Mu, X., Li, S., & Peng, H. (2020). A Review of Face Recognition Technology. *IEEE Access*, *8*, 139110–139120. https://doi.org/10.1109/ACCESS.2020.3011028

Li, Y., Mamouei, M., Salimi-Khorshidi, G., Rao, S., Hassaine, A., Canoy, D., Lukasiewicz, T., & Rahimi, K. (2023). Hi-BEHRT: Hierarchical Transformer-Based Model for Accurate Prediction of Clinical Events Using Multimodal Longitudinal Electronic Health Records. *IEEE Journal of Biomedical and Health Informatics*, *27*(2). https://doi.org/10.1109/JBHI.2022.3224727

Lindskou, T. A., Mikkelsen, S., Christensen, E. F., Hansen, P. A., Jørgensen, G., Hendriksen, O. M., Kirkegaard, H., Berlac, P. A., & Søvsø, M. B. (2019). The Danish prehospital emergency healthcare system and research possibilities. In *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* (Vol. 27, Number 1). BioMed Central Ltd. https://doi.org/10.1186/s13049-019-0676-5

Maciver, M. (2021). Pre-hospital use of early warning scores to improve detection and outcomes of sepsis. *British Journal of Community Nursing*, *26*(3). https://doi.org/10.12968/bjcn.2021.26.3.122

Madhavi Latha Vadlamudi. (2025). The rise of edge computing in healthcare data processing. *World Journal of Advanced Research and Reviews*, *26*(1). https://doi.org/10.30574/wjarr.2025.26.1.1092

Mayampurath, A., Jani, P., Dai, Y., Gibbons, R., Edelson, D., & Churpek, M. M. (2020). A Vital Sign-Based Model to Predict Clinical Deterioration in Hospitalized Children. *Pediatric Critical Care Medicine*, *21*(9). https://doi.org/10.1097/PCC.0000000000002414

McNarry, A. F., & Goldhill, D. R. (2004). Simple bedside assessment of level of consiousness: Comparison of two simple assessment scales with the Glasgow Coma Scale. *Anaesthesia*, *59*(1). https://doi.org/10.1111/j.1365-2044.2004.03526.x

Muralitharan, S., Nelson, W., Di, S., McGillion, M., Devereaux, P. J., Barr, N. G., & Petch, J. (2021). Machine learning–Based early warning systems for clinical deterioration: Systematic scoping review. *Journal of Medical Internet Research*, *23*(2). https://doi.org/10.2196/25187

Olakotan, O., Samuriwo, R., Ismaila, H., & Atiku, S. (2025). Usability Challenges in Electronic Health Records: Impact on Documentation Burden and Clinical Workflow: A Scoping Review. In *Journal of Evaluation in Clinical Practice* (Vol. 31, Number 4). https://doi.org/10.1111/jep.70189

Papadimitriou, C., Lindemann, L., & Meehan, A. J. (2023). Making the visible seen: The interactional competence of a person in a disordered state of consciousness. *Social Science and Medicine*, *336*. https://doi.org/10.1016/j.socscimed.2023.116261

Parrino, C. R., Fransman, R. B., Varone, A. J., & Galvagno, S. M. (2024). Advanced Trauma Life Support. In *Recent Strategies in High Risk Surgery*. https://doi.org/10.1007/978-3-031-56270-9_11

Patel, R., Nugawela, M. D., Edwards, H. B., Richards, A., Le Roux, H., Pullyblank, A., & Whiting, P. (2018). Can early warning scores identify deteriorating patients in pre-hospital settings? A systematic review. In *Resuscitation* (Vol. 132). https://doi.org/10.1016/j.resuscitation.2018.08.028

Patharkar, A., Cai, F., Al-Hindawi, F., & Wu, T. (2024). Predictive modeling of biomedical temporal data in healthcare applications: review and future directions. In *Frontiers in Physiology* (Vol. 15). https://doi.org/10.3389/fphys.2024.1386760

Pimentel, M. A. F., Redfern, O. C., Gerry, S., Collins, G. S., Malycha, J., Prytherch, D., Schmidt, P. E., Smith, G. B., & Watkinson, P. J. (2019). A comparison of the ability of the National Early Warning Score and the National Early Warning Score 2 to identify patients at risk of in-hospital mortality: A multi-centre database study. *Resuscitation*, *134*. https://doi.org/10.1016/j.resuscitation.2018.09.026

Pokee, D. Q., Pereira, C. B., Mösch, L., Follmann, A., & Czaplik, M. (2021). Consciousness detection on injured simulated patients using manual and automatic classification via visible and infrared imaging. *Sensors*, *21*(24). https://doi.org/10.3390/s21248455

Ramzan, M., Abid, A., Fayyaz, M., Alahmadi, T. J., Nobanee, H., & Rehman, A. (2024). A Novel Hybrid Approach for Driver Drowsiness Detection Using a Custom Deep Learning Model. *IEEE Access*, *12*. https://doi.org/10.1109/ACCESS.2024.3438617

Ramzan, M., Khan, H. U., Awan, S. M., Ismail, A., Ilyas, M., & Mahmood, A. (2019). A Survey on State-of-the-Art Drowsiness Detection Techniques. In *IEEE Access* (Vol. 7, pp. 61904–61919). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2019.2914373

Renc, P., Grzeszczyk, M. K., Oufattole, N., Goode, D., Jia, Y., Bieganski, S., Mcdermott, M. B. A., Was, J., Samir, A. E., Cunningham, J. W., Bates, D. W., & Sitek, A. (2025). Foundation model of electronic medical records for adaptive risk estimation. *GigaScience*, *14*. https://doi.org/10.1093/gigascience/giaf107

Romero-Brufau, S., Smunyahirun, R., Filhol, T., Niederhauser, L., Trakoolwilaiwan, T., & Singh, G. (2025). Vital Signs–Only Machine Learning Model for Acute Inpatient Deterioration: A Retrospective Multicenter Study. *Mayo Clinic Proceedings: Innovations, Quality and Outcomes*, *9*(5). https://doi.org/10.1016/j.mayocpiqo.2025.100663

Sadeghi, R., Banerjee, T., & Romine, W. (2018). Early hospital mortality prediction using vital signals. *Smart Health*, *9–10*. https://doi.org/10.1016/j.smhl.2018.07.001

Sanz, L. R. D., Thibaut, A., Edlow, B. L., Laureys, S., & Gosseries, O. (2021). Update on neuroimaging in disorders of consciousness. In *Current Opinion in Neurology* (Vol. 34, Number 4). https://doi.org/10.1097/WCO.0000000000000951

Sarwar, T., Seifollahi, S., Chan, J., Zhang, X., Aksakalli, V., Hudson, I., Verspoor, K., & Cavedon, L. (2023). The Secondary Use of Electronic Health Records for Data Mining: Data Characteristics and Challenges. In *ACM Computing Surveys* (Vol. 55, Number 2). https://doi.org/10.1145/3490234

Seroussi, B., & Zablit, I. (2024). Implementation of Digital Health Ethics: A First Step with the Adoption of 16 European Ethical Principles for Digital Health. *Studies in Health Technology and Informatics*, *310*. https://doi.org/10.3233/SHTI231331

Smith, J., Andrew, E., & Smith, K. (2022). Pre-hospital early warning scores are associated with requirement for medical retrieval services. *Australasian Journal of Paramedicine*, *19*. https://doi.org/10.33151/AJP.19.956

Soar, J., Böttiger, B. W., Carli, P., Couper, K., Deakin, C. D., Djärv, T., Lott, C., Olasveengen, T., Paal, P., Pellis, T., Perkins, G. D., Sandroni, C., & Nolan, J. P. (2021). European Resuscitation Council Guidelines 2021: Adult advanced life support. In *Resuscitation* (Vol. 161, pp. 115–151). https://doi.org/10.1016/j.resuscitation.2021.02.010

Stachwitz, P., & Debatin, J. F. (2023). Digitalization in healthcare: today and in the future. In *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* (Vol. 66, Number 2). https://doi.org/10.1007/s00103-022-03642-8

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. In *Nature Medicine* (Vol. 25, Number 1). https://doi.org/10.1038/s41591-018-0300-7

Van Rest, J., Boonstra, D., Everts, M., Van Rijn, M., & Van Paassen, R. (2014). Designing privacy-by-design. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *8319*. https://doi.org/10.1007/978-3-642-54069-1_4

Wang, Y., Yin, C., & Zhang, P. (2024). Multimodal risk prediction with physiological signals, medical images and clinical notes. *Heliyon*, *10*(5). https://doi.org/10.1016/j.heliyon.2024.e26772

Wasserman, E. B., Shah, M. N., Jones, C. M. C., Cushman, J. T., Caterino, J. M., Bazarian, J. J., Gillespie, S. M., Cheng, J. D., & Dozier, A. (2015). Identification of a neurologic scale that optimizes EMS detection of older adult traumatic brain injury patients who require transport to a trauma center. *Prehospital Emergency Care*, *19*(2). https://doi.org/10.3109/10903127.2014.959225

Williams, B. (2019). The national early warning score and the acutely confused patient. In *Clinical Medicine, Journal of the Royal College of Physicians of London* (Vol. 19, Number 2). https://doi.org/10.7861/CLINMEDICINE.19-2-190

Williams, B. (2022). The National Early Warning Score: From concept to NHS implementation. *Clinical Medicine, Journal of the Royal College of Physicians of London*, *22*(6), 499–505. https://doi.org/10.7861/clinmed.2022-news-concept

Wosik, J., Fudim, M., Cameron, B., Gellad, Z. F., Cho, A., Phinney, D., Curtis, S., Roman, M., Poon, E. G., Ferranti, J., Katz, J. N., & Tcheng, J. (2020). Telehealth transformation: COVID-19 and the rise of virtual care. In *Journal of the American Medical Informatics Association* (Vol. 27, Number 6). https://doi.org/10.1093/jamia/ocaa067

Zadravecz, F. J., Tien, L., Robertson-Dick, B. J., Yuen, T. C., Twu, N. M., Churpek, M. M., & Edelson, D. P. (2015). Comparison of mental-status scales for predicting mortality on the general wards. *Journal of Hospital Medicine*, *10*(10). https://doi.org/10.1002/jhm.2415

Zhang, D., Yin, C., Zeng, J., Yuan, X., & Zhang, P. (2020). Combining structured and unstructured data for predictive models: a deep learning approach. *BMC Medical Informatics and Decision Making*, *20*(1). https://doi.org/10.1186/s12911-020-01297-6

## Annexes

### ANNEX A - AVPU scenario video recording (in collaboration with Pedro Faria)

To construct a dedicated dataset for the video-based component of this study, a series of recordings was conducted over a three-month period in 2025. The recording setup and technical production were carried out in collaboration with Pedro Faria, using his professional filming equipment. The objective was to ensure standardized acquisition conditions while preserving clinically plausible stimulus–response interactions aligned with AVPU assessment logic.

#### Scenario setup

- The recording device was secured on a stable tripod and aligned horizontally with the participant's eye level, positioned at an approximate distance of 1.5 meters to ensure consistent framing.
- Uniform illumination was provided by a studio LED light source (Viltrox RB08 @3200K) attached to the camera mount, delivering neutral colour temperature and minimizing shadow variability across sessions.
- Participants were seated on a chair or bench, oriented directly toward the camera. Head orientation and gaze direction were adjusted in accordance with the intended AVPU scenario and the predefined stimulus progression.
- The interviewer stood slightly off-axis, either to the left or right of the participant, allowing unobstructed visual capture of facial and upper-body responses while performing the structured stimulus sequence.

#### Recording Setup and Output

- Recordings were captured using a Canon EOS RP camera equipped with a Canon EF 16–35mm f/4L IS USM lens, configured at a 35mm focal length and f/4 aperture.
- Video files were produced in MPEG-4 format at a resolution of 1920×1080 pixels and a frame rate of 25 frames per second, with audio encoded using Linear PCM and H.264 standards, stored in stereo MP4 AAC format at 48 kHz.
- Each participant and corresponding simulated consciousness-level (CL) condition constituted a distinct entry within the dataset.
- Individual clips ranged between approximately 6 and 12 seconds in duration, capturing the structured interaction between interviewer and participant according to the specific stimulus phase of the AVPU sequence.

**ANNEX B – Script for AVPU Scenario Performance (in collaboration with Pedro Faria)**

The scripted scenarios were designed to replicate the operational logic of AVPU assessment as performed in emergency and prehospital care. Each interaction follows a structured escalation pathway: observation → verbal stimulus → tactile/verbal repetition → pain stimulus. The actor's performance encodes clinically meaningful responsiveness patterns that are both medically plausible and computationally labelable.

The objective was not theatrical realism, but reproducible stimulus–response dynamics. Each scenario preserves natural motion timing and avoids exaggerated gestures to ensure that AI models learn physiologically credible patterns rather than artificial cues.

At the global level, scenarios are mapped to AVPU categories and to a binary classification framework:

- **Alert (A)** → non-impaired
- **Verbal (V), Pain (P), Unresponsive (U)** → impaired

This structure mirrors Basic Life Support logic and supports both clinical interpretability and machine-learning annotation.

| Scenario | Baseline State | Response to Observation | Response to Voice | Response to Pain | AVPU Label | Binary Label |
|---|---|---|---|---|---|---|
| Alert | Eyes open | Spontaneous engagement | Immediate, purposeful | Not required | A | Non-Impaired |
| Verbal | Eyes closed | None | Delayed/minimal response | Not required | V | Impaired |
| Pain | Eyes closed | None | None | Reflexive withdrawal/grimace | P | Impaired |
| Unresponsive | Eyes closed | None | None | None | U | Impaired |

Table B.1 - Resumed Script per Scenario

**ANNEX C – Dataset production (in collaboration with Pedro Faria)**

The participant's recruitment was mainly related to healthcare and done in research, medical or EMS settings. Recordings that did not meet real-life conditions and subject natural responsiveness were discarded.

The filming, according to AVPU standards, as well as the labelling of the dataset entries were supported by my colleague in ManagiDiTH program, Pedro Faria.

The produced dataset was then used for training and validation of an experimental algorithm being developed by our colleague Marco Esperança.

**Dataset:**

• Number of videos: 136.

• Number of subjects: 58.

• Gender proportion: 36 female, 22 male.

• Race diversity: 49 Caucasian, 4 brown, 3 black, 2 Indian.

• Age range: Between 25 and 55 years old.

• Professional background: MD (6), nurse (30), first responder (16), healthcare consultant (6).

• Simulations: Response to verbal and touch stimulus, pain stimulus and unresponsive.

• GRPD and privacy concerns: Context brief and informed consent and were provided to all the participants prior to video recording, accordingly to ISCTE University ethics commission guidelines (CEISTA/ 2025.08).

| Alert | | | Unresponsive (U) | | |
|---|---|---|---|---|---|
| **27** | | | **109** | | |
| V | P | N.A.* | V | P | N.A.* |
| 12 | 14 | 1 | 56 | 52 | 1 |

NA* out of scope reactions (e.g. convulsion)

Table C.1 –Dataset segmented by AVPU binary categories

## ANNEX D - Reporting the Vignette Experiments (using GROVE)

| GROVE Criteria | Description (per GROVE) | How This Study Addresses the Criteria |
|---|---|---|
| **1. Rationale** | Explain why a vignette-based design was chosen and why the study could not be conducted in a non-simulated setting. | A video vignette design is used to simulate prehospital emergency assessments of responsiveness, as systematic manipulation of verbal, tactile, and painful stimuli in real patients would be unethical and impractical. Vignettes allow controlled, reproducible evaluation of observable responses relevant to AV classification. |
| **2. Vignette Content** | Describe how vignette content was developed and refined, including key methodological choices. | Video vignettes are scripted to represent prehospital emergency scenarios, following a standardized AVPU stimulus sequence (observation, verbal/light tactile stimulation, pain stimulation, no response). All non-manipulated elements are kept constant across vignettes. |
| **2.1 Healthcare Scenario** | Describe the healthcare context, setting, characters, and sources informing vignette content. | Scenarios represent emergency medical service encounters in a prehospital context. The portrayed subject is assessed by a simulated responder applying AVPU-consistent stimuli. Content is informed by clinical guidelines and emergency assessment workflows. |
| **2.2 Manipulation & Standardization** | Describe experimental manipulations and how other elements were standardized. | The experimental manipulation is the type and intensity of stimulus applied (none, verbal/light tactile, pain). Camera placement, lighting, subject positioning, and scenario duration are standardized across all vignettes. |
| **2.3 Mode of Delivery** | Describe vignette modality and delivery details needed for replication. | Vignettes are delivered as short video recordings captured using fixed cameras focused on the face and upper limbs. Videos are viewed by annotators in a standardized digital environment. |
| **2.4 Expert Involvement** | Describe expert involvement in vignette development. | Emergency care clinicians contribute to scenario design, stimulus definition, and development of annotation guidelines to ensure clinical realism and relevance. |
| **2.5 Pilot Testing** | Describe if and how pilot testing was conducted. | Pilot testing of vignette clarity, realism, and annotation feasibility is planned prior to full annotation. Feedback will be used to refine stimulus timing and annotation criteria. |
| **3. Outcomes** | Explain selected outcomes and how they relate to real-world outcomes. | Outcomes are expert-assigned AVPU labels (Alert, Verbal, Pain, Unresponsive) and a derived binary reaction indicator. These outcomes serve as proxies for real-world clinical assessments used in NEWS scoring. |
| **4. Vignette Validity & Realism** | Describe how manipulation success, realism, and validity are evaluated. | Vignette validity will be evaluated through expert review and inter-rater agreement analysis. Agreement between annotators will be quantified to assess reliability of the operationalized AVPU outcomes. |
| **5. Participants** | Provide rationale for participant selection. | Annotators will be clinicians or trained healthcare professionals with experience in emergency assessment or AVPU scoring, ensuring alignment with the real-world population performing such evaluations. |
| **6. Availability** | Describe availability and reuse of vignettes and data. | Availability of vignettes and annotations for research, training, or further development will be described, subject to ethical, institutional, and data protection constraints. |

Table D.1- Planned annotation procedure accordingly to GROOVE guidelines. (Hillen et al., 2025)

**ANNEX E - Inter-rater reliability**

| R1 \ R2 | A | V | P | U | Row Total |
|---|---|---|---|---|---|
| **A** | 30 | 4 | 0 | 0 | 34 |
| **V** | 3 | 25 | 5 | 0 | 33 |
| **P** | 0 | 4 | 18 | 3 | 25 |
| **U** | 0 | 0 | 4 | 16 | 20 |
| **Column Total** | 33 | 33 | 27 | 19 | **112** |

Table E.1 - AVPU Confusion Matrix (N = 112)

| Outcome | % Agreement | κ |
|---|---|---|
| AVDS (4-level) | 79.5% | 0.72 |
| Binary Reaction | 93.8% | 0.79 |

Table E.2 – Observed Agreement and Cohen's Kappa

**ANNEX F - Statistics outputs**

**Study population characteristics (N = 111,202)**

| Variable | Overall |
|---|---|
| **Total observations** | 111,202 |
| **Alert (AVPU = A)** | 100,748 (90.6%) |
| **Non-alert (V/P/U)** | 10,454 (9.4%) |
| **NEWS (median, IQR)** | 2 (1–4) |
| **NEWS ≤ 5 (low risk)** | 91,954 (82.7%) |
| **NEWS > 5 (higher risk)** | 19,248 (17.3%) |

Table F.1 - Sample descriptive analysis

**Association between AVPU category and elevated NEWS score**

| AVPU category | NEWS < 5 | NEWS ≥ 5 | Total |
|---|---|---|---|
| Alert (A) | 86,497 (85.9%) | 14,251 (14.1%) | 100,748 |
| V/P/U | 2,240 (21.4%) | 8,214 (78.6%) | 10,454 |
| **Total** | 88,737 | 22,465 | 111,202 |

Table F.2 - Crosstabulation: AVPU × NEWS risk category

| Test | Value | df | p-value |
|---|---|---|---|
| Pearson Chi-Square | >10,000 | 1 | <0.001 |

Table F.3 - Chi-Square Tests

| Measure | Value | 95% CI |
|---|---|---|
| Risk Ratio (VPU vs Alert) | **5.57** | ~5.45 – 5.69 |
| Odds Ratio (VPU vs Alert) | **≈ 22.3** | ~21.6 – 23.0 |

Table 3- **Risk Estimates**

**Multivariable regression model**

| Predictor | Coding / Scale | Effect Estimate | 95% CI | p-value |
|---|---|---|---|---|
| **Impaired consciousness (V/P/U)** | V/P/U vs Alert | aOR ≈ 48 | ~45–52 | <0.001 |
| **Respiratory rate (standardized)** | per +1 SD | Directional ↑ | — | <0.001 |
| **Oxygen saturation (standardized)** | per +1 SD | Directional ↓ | — | <0.001 |
| **Supplemental oxygen use** | Yes vs No | Directional ↑ | — | <0.001 |
| **Systolic blood pressure (standardized)** | per +1 SD | Directional ↓ | — | <0.001 |
| **Heart rate (standardized)** | per +1 SD | Directional ↑ | — | <0.001 |
| **Temperature (standardized)** | per +1 SD | Directional ↑ | — | <0.001 |

Table 4 - Multivariable association, standardized physiological parameters, and NEWS ≥ 5

**Average marginal effect of V/P/U: +20.4 percentage points increase in probability of NEWS ≥ 5.**

**Reclassification Matrix**

| NEWS without AVPU (Base) | NEWS < 5 (Full) | NEWS ≥ 5 (Full) | Total |
|---|---|---|---|
| **NEWS < 5** | 81,654 | 3,072 (2.8%) | 84,726 |
| **NEWS ≥ 5** | 0 | 26,476 | 26,476 |
| **Total** | 81,654 | 29,548 | 111,202 |

Table 5 - Reclassification of Risk Across NEWS ≥ 5 after Inclusion of AVPU

NEWS base: The score calculated excluding the AVPU component (−3 points for V/P/U).

NEWS full: The recorded NEWS including AVPU

Up-classification indicates patients crossing the escalation threshold (NEWS < 5 to NEWS ≥ 5) after inclusion of AVPU

**Sensitivity analysis**

| NEWS without AVPU (Base) | NEWS < 5 (Full) | NEWS ≥ 5 (Full) | Total |
|---|---|---|---|
| **NEWS < 5** | 81,654 | 3,072 (2.8%) | 84,726 |
| **NEWS ≥ 5** | 0 | 26,476 | 26,476 |
| **Total** | 81,654 | 29,548 | 111,202 |

Table F.7 - Threshold: NEWS ≥ 5

| NEWS without AVPU (Base) | NEWS < 7 (Full) | NEWS ≥ 7 (Full) | Total |
|---|---|---|---|
| **NEWS < 7** | 94,642 | **3,123** | 97,765 |
| **NEWS ≥ 7** | 0 | 13,437 | 13,437 |
| **Total** | 94,642 | 16,56 | 111,202 |

Table 6 - Threshold: NEWS ≥ 7