Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Medicine 2149

On the Quantitative Evaluation and Enhancement of Prehospital Decisional Capacity

DOUGLAS SPANGLER

ACTA UNIVERSITATIS
UPSALIENSIS
2025

ISSN 1651-6206
ISBN 978-91-513-2469-2
urn:nbn:se:uu:diva-553889

UPPSALA
UNIVERSITET

Dissertation presented at Uppsala University to be publicly examined in H:son Holmdahlsalen, Hus 100, Akademiska sjukhuset, Dag Hammarskjölds 8, Uppsala, Wednesday, 28 May 2025 at 13:00 for the degree of Doctor of Philosophy (Faculty of Medicine). The examination will be conducted in Swedish. Faculty examiner: Professor Markus Skrifvars (University of Helsinki).

Abstract
Spangler, D. 2025. On the Quantitative Evaluation and Enhancement of Prehospital Decisional Capacity. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine* 2149. 73 pp. Uppsala: Acta Universitatis Upsaliensis. ISBN 978-91-513-2469-2.

Prehospital care involves increasingly complex decision-making processes, necessitating commensurate advances in the methods used to assess and improve the quality of patient triage processes. This doctoral project aimed to advance the measurement of patient outcomes in the evaluation of prehospital decision-making and develop interventions to improve those outcomes. In study I, a set of outcome definitions for evaluating referrals to non-emergency care by dispatch nurses was validated, confirming the ability of systematic data abstraction processes to identify patient harms missed by traditional incident reporting systems. In study II, an intervention delivering feedback on process and outcome metrics to dispatch nurses was evaluated, identifying improvements in some process metrics, while impacts on outcomes remained elusive. In study III a machine learning-based approach to estimating composite risk scores was validated internally for use in prehospital contexts. In study IV, similar models for use in Ambulance care were externally validated in a dataset collected from six Swedish regions, finding that model performance remained superior to traditional rule-based risk assessment instruments even when the models were applied in novel settings. Study V is a randomized controlled trial whereby a clinical decision support tool based on these models was found to enhance the ability of dispatchers to identify and prioritize high-risk patients in resource constrained situations. Future directions for study include the incorporation of additional structured and unstructured data in the prediction models, and efforts to evaluate and enhance their fairness and alignment with human assessments of care need. Open-source software packages implementing these tools are available to enhance the transparency of the work and stimulate further development.

Douglas Spangler, Department of Surgical Sciences, Akademiska sjukhuset, Uppsala University, SE-75185 Uppsala, Sweden. Department of Public Health and Caring Sciences, Box 564, Uppsala University, SE-75122 Uppsala, Sweden.

ISSN 1651-6206
ISBN 978-91-513-2469-2
URN urn:nbn:se:uu:diva-553889 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-553889)

For dad

Sometimes science is more art than science
- Rick Sanchez

*Cover image depicts a Principal Component Analysis of ca. 800 000 Swedish ambulance missions, see the Unsupervised Methods section for detail*s

List of papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

I. **Spangler D**, Edmark L, Winblad U, Colldén-Benneck J, Borg H, Blomberg H. Using trigger tools to identify triage errors by ambulance dispatch nurses in Sweden: an observational study. *BMJ Open*. 2020;10(3):e035004.

II. **Spangler D**, Blomberg H. Continuous individual feedback to nurses at emergency medical dispatch centers: a stepped-wedge, interrupted time series analysis. *BMJ Open Qual*. 2025;14(1):e002993.

III. **Spangler D**, Hermansson T, Smekal D, Blomberg H. A validation of machine learning-based risk scores in the prehospital setting. *PLoS One*. 2019;14(12):e0226518.

IV. **Spangler D**, Höglund E, Magnusson C, Wibring K, Palm E, Smekal D, Blomberg H. Estimating generalization loss in prehospital machine learning risk prediction models: A multicenter internal-external validation study. *Submitted manuscript*

V. **Spangler D**, Morelli S, Smekal D, Edmark L, Blomberg H. Machine learning Assisted Differentiation of Low Acuity patients at Dispatch (MADLAD): A Randomized Controlled Trial. *Manuscript*

List of additional papers

Holmström IK, Kaminsky E, Lindberg Y, **Spangler D**, Winblad U. Registered Nurses' experiences of using a clinical decision support system for triage of emergency calls: A qualitative interview study. *Journal of Advanced Nursing*. 2020;76(11):3104–12.

Spangler D, Blomberg H, Smekal D. Prehospital identification of Covid-19: an observational study. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*. 2021;29(1):1–10.

Spangler D, Linder W, Winblad, U. The Impact of the Swedish Care Coordination Act on Hospital Readmission and Length-of-Stay among Multi-Morbid Elderly Patients: A Controlled Interrupted Time Series Analysis. *International Journal of Integrated Care* 23, 17 (2023).

Höglund E, Magnusson C, Lederman J, **Spangler D**, Vloet L, Ebben R. Ambulance quality and outcome measures for general non-conveyed populations (AQUA): A scoping review. *PLoS One*. 2024;19(8): e0306341.

Contents

Abbreviations

API	Application Programming Interface
CDS	Clinical Decision Support
DCE	Discrete Choice Experiment
ED	Emergency Department
EMD	Emergency Medical Dispatch
EMS	Emergency Medical Service
GUI	Graphical User Interface
NEWS	National Early Warning Score
NHS	National Health Service
MAR	Missing At Random
MBS	*Medicinsk Beslutsstöd* (CDS used at SvLc)
MDR	Medical Device Regulation
ML	Machine Learning
PCA	Principal Component Analysis
PIN	Personal Identification Number
RETTS	Rapid Emergency Triage and Treatment System
RCS	Resource Constrained Situation
RCT	Randomized Controlled Trial
SvLc	*Sjukvårdens Larmcentral* (regional EMD center)

Introduction

Prehospital care is in the process of transitioning from a transport-focused service for delivering patients to hospitals, to an advanced healthcare service in and of itself [1,2]. In addition to the traditional task of providing stabilizing treatment to critically ill and injured patients, prehospital care providers are increasingly tasked with identifying patients with conditions not requiring emergency care. These patients are then directed away from the traditional chain of prehospital emergency care to non-emergency care providers [3–6]. This process of stratifying patients based on the acuity of their condition is often referred to as triage, and is particularly difficult when the clinical risks of a condition depend on a large number of factors with complex interactions and a high degree of uncertainty [7,8].

The judicious practice of triage and the referral of low-acuity patients to non-emergency care holds the potential to alleviate the increasing strain on ambulance services and emergency departments. The practice can however result in serious patient safety hazards if true emergencies are missed. Prehospital care systems incorporating triage to non-emergency care providers must therefore carefully monitor the appropriateness of these triage decisions. They should also provide effective tools for care providers to maximize their capacity to make accurate decisions. Such tools may take retrospective forms such as providing feedback on the results of previous decisions, or prospective forms such as the use of Clinical Decision Support (CDS) systems. Numerous factors including formal education, training, experience, clinical guidelines, personality traits, stress, alertness, availability of supporting tools, etc., impact the ability of a care provider to make appropriate decisions. The final result of the interplay of these factors with regards to a care provider's ability to make appropriate decisions is termed the care providers decisional capacity in this thesis.

The prehospital decision-making process regarding low-acuity patients has received increasing attention in the literature [9–11], but remains under-studied in comparison to decision-making processes in other areas of healthcare. Feedback to care providers about their performance is reported as often lacking in healthcare [12]. Some feedback-based learning will occur naturally in contexts where contacts with patients is prolonged or recurring. In the context of prehospital care however, where contact with patients is brief, care providers are

often not aware of the results of their decisions. CDS systems have traditionally been knowledge- and rule-based; developed based on the clinical domain knowledge of experts. Systems employing predictive statistical models have begun to demonstrate the potential to improve the accuracy of prehospital triage decisions [13–15]. There is a little consensus in the literature regarding how to best operationalize outcomes for low-acuity patients in prehospital care, nor is it well established how to best quantitatively evaluate prehospital decision-making processes [4,16–18]. There is also a lack of studies investigating the ability of machine learning-based risk prediction models to generalize across distinct prehospital care systems. To the authors knowledge, only a single study has investigated whether the theoretical accuracy of risk prediction models in prehospital care can translate into improved decision-making in a Randomized Controlled Trial (RCT) [14].

Aim

The aim of this thesis is to advance the methods used for evaluating and improving the decisional capacity of prehospital care providers, with particular emphasis on the assessment of low-acuity patients. This aim is achieved through five studies related per the diagram below:

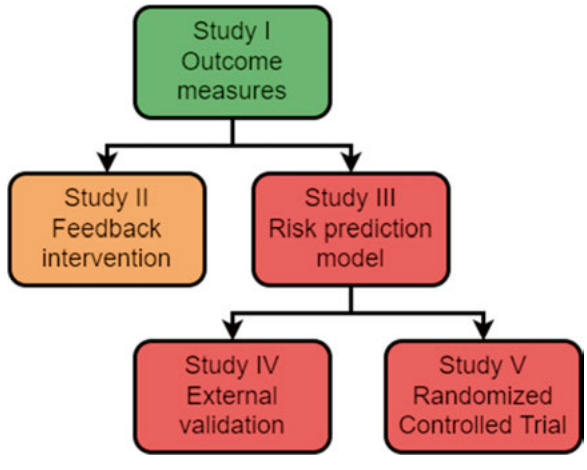

Figure 1 - Flow of included studies

Study I aims to validate and explore the properties of a set of measures used to assess outcomes for patients referred to non-emergency care by dispatch nurses. Study II builds upon these results in evaluating the impact of a feedback intervention seeking to improve outcomes. Finally, studies III – V revolve around the development, validation, and implementation of machine learning models used to predict outcomes for patients in prehospital care.

Background

Prehospital Care Systems

The prehospital care systems of today have their roots in military trauma care systems developed to care for injuries sustained by soldiers in battle. Martial trauma care systems developed over time from the battalion-level surgeons established by Philip II of Macedon in the 4th century BC, to the formalized ambulance services forged in the heat the Napoleonic-, American Civil-, and Crimean wars of the 19th century. Civilian ambulance services bearing the hallmarks of modern systems were developed soon after, and were by the end of that century present in many American and European cities.[19]

The basic process of prehospital care has not changed much since. A sick or injured individual makes contact with a central instance which today is typically an Emergency Medical Dispatching (EMD) center. A resource, typically an ambulance, is then dispatched which makes their way to the individual, provides stabilizing treatment, and transports them to definitive care, typically a hospital Emergency Department (ED). This formula was maintained throughout the 20th century, with the focus on rapid transport and limited on-scene care arguably only strengthened over time. The national standards set by the National Highway Traffic Safety Administration in the US, [20] standardization in the UK around first-aid trained ambulance staff [19] and the emergence of the possibly mythological "Golden hour" in the care of trauma patients all contributed to a focus on rapid stabilization and transport [21,22].

The core mission and focus of the prehospital care system today remains the stabilization and transport of critically ill and injured patients to definitive emergency care. The success of modern public health initiatives has however led to substantial demographic shifts in the populations which these systems serve. The populations in developed countries are rapidly aging, driving substantial increases in the demand for prehospital care. [23–25] Primary- and elder care systems in many countries are furthermore insufficient to fully meet these additional care needs [26–28]. As such, much of the increase in demand has come in the form patients seeking care for low-acuity conditions which do not necessarily require the high-intensity interventions available to care providers staffing ambulances and EDs. This group of patients has been well

characterized in the literature [4,10,29], and the tools and organization of prehospital care systems are beginning to adapt to this new landscape.

In the United States, the use of Community Paramedics as a kind of ambulatory primary care service has seen adoption, with 78% of respondents to a 2023 survey reporting some form of Mobile Integrated Health/Community Paramedicine project [30]. The practice of care provider initiated non-conveyance is however rare given that ambulance services are reimbursed only for transports to EDs, and the legal risks involved. There have been efforts to expand the scope of services for which Emergency Medical Service (EMS) providers are reimbursed to include on-scene care without transport to an ED, although the project evaluating this change at the national level was terminated early due to a lack of participation [31]. Nonetheless, given the lack of universal health insurance coverage in the US, EMS continues to serve as a safety-net service, particularly for socioeconomically disadvantaged populations. Efforts to effectively and safely modulate the level of care provided to low-acuity patients thus continue by necessity, driven especially by the pressure placed on EMS systems by the Covid-19 pandemic [32–34].

In Europe, financing is more commonly value-, or budget-based and primary care systems are generally functional. As such, non-conveyance, conveyance to alternate destinations, and referral by dispatchers to non-emergency care is more common. The UK National Health Service (NHS) for instance has undertaken systematic efforts to safely reduce avoidable conveyances [35], and was early to fund research on appropriate care pathways for low-acuity patients [5,36,37]. In the NHS England ambulance trust, these sustained efforts have resulted in 17% of callers to the EMD center being referred to non-emergency care, and an ambulance non-conveyance rate of 29.7% in 2024 [38]. This high rate of non-conveyance is facilitated by not only a high level of competency through the use of paramedic practitioners [39], but also the use of organizational adaptations. These include changes in legislation enabling these practitioners to directly prescribe medications [40] and effective CDS tools such as the National Early Warning Score (NEWS) which is implemented across all NHS England ambulance trusts [41]. In the Netherlands, the non-conveyance rate in 2015 was 20.4% in a system based on primarily nurse-staffed ambulances [42]. Published information in the English language regarding non-conveyance rates in other major European countries is scant [4].

In Sweden, the organization of prehospital care reflects the decentralized nature of the healthcare system at large, with each of the 21 regions organizing its own prehospital care system. As in the Netherlands, ambulances are staffed by at least one nurse, and care practitioners thus have a relatively broad base of clinical education to draw upon. Non-conveyance practices are varied, with non-conveyance rates across the six regions included in the dataset for Study

IV ranging between 10 and 22%. Regions employ a variety of resources to appropriately meet patient care needs, including single-responder units for patient assessment and non-nurse staffed patient transport vehicles [43]. The national emergency number (112) is in most regions managed by a government owned corporation (SOS Alarm) which employs dispatchers without formal medical education in the primary call taking role [44]. Several regions however have accommodations with SOS Alarm entailing the use of nurses in a secondary triage role [45]. Four regions as of 2025 (Uppsala, Västmanland, Sörmland, and Värmland) handle EMD as an integral part of the ambulance service under the moniker *Sjukvårdens Larmcentral* (SvLc), including the regions where the studies in this thesis primarily take place. These employ exclusively nurses in the primary call-taking role.

Prehospital Care Pathways

To provide a referential framework for the patient flows assumed throughout this thesis, an idealized prehospital care system including alternate care pathways for low-acuity patients is described below and in Figure 2. In a typical prehospital care system, most callers to the emergency phone number will receive an ambulance response with either a high or low priority and in turn be transported by the ambulance to an ED. This is the primary prehospital "chain of care", corresponding to the bolded patient flow in Figure 2. In many modern systems this is not the only care pathway available, and a share of patients are referred to alternate forms of care after assessment by the dispatch center or ambulance. These are denoted by nodes with dotted outlines in Figure 2. These alternate types of care can be divided into Alternate prehospital *resources* which can be dispatched to the patient (orange), and alternate forms of definitive *care* (red). These alternate forms of resources and definitive care can be nothing and none, respectively; a determination that no further care is required for the time being. Upon being provided some form of definitive care, patients will experience an outcome (purple). The most important of these include whether and how care was subsequently provided (admission to in-patient care being the most commonly evaluated), and survival for some period of time [18]. Having survived their care episode, patients may re-contact the prehospital- or emergency care system, constituting another commonly measured outcome. "Frequent fliers" who do so often are in fact a popular topic of research in and of themselves [46–48].

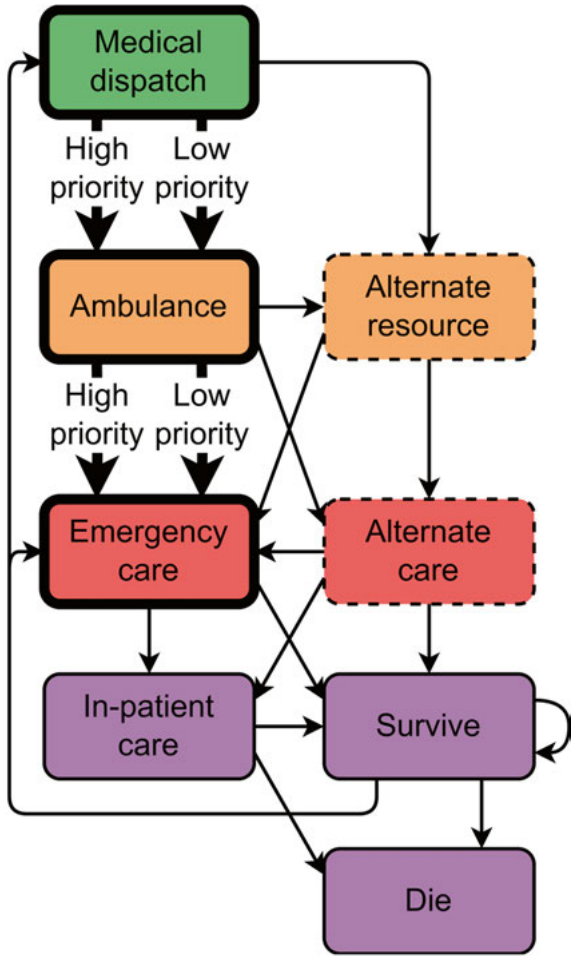

Figure 2 - Diagram of Prehospital care pathways. Green, yellow, and red, nodes represent the actors a patient may be in contact with throughout a prehospital episode of care. Purple nodes represent some outcomes relevant to evaluating prehospital care.

In this model, the first task of a care provider at the dispatch center is to determine whether a patient has high-acuity condition requiring a high priority ambulance response with lights and sirens, termed a "Priority 1" response in Sweden. Once the presence of such a condition is ruled out, the provider may continue to assess this lower-acuity patient to determine which form of care is most appropriate to meet their needs in the most effective manner. The provider may determine that the patient requires an ambulance, but that driving with lights and sirens is not necessary (a "Priority 2" response), that some alternate resource (e.g., a mobile physician- or psychiatric team, single-staffed transport unit, or even a taxi) will suffice, or that no resource needs to be sent. Care providers both at the dispatch center and on the ambulance make similar decisions regarding the appropriate level of definitive care. In addition to

traditional definitive care at an emergency department, prehospital care providers may refer patients to a number of non-emergency definitive care options, including for instance primary- and urgent care centers, direct admission to hospital units, and nursing advice lines.

Complicating this picture is the fact that these resources and forms of definitive care may be combined in many ways; it may be that a patient requires care at a hospital ED but that no prehospital care is necessary to stabilize the patient, and that a taxi is sufficient to facilitate transport. Conversely, a patient may require an immediate prehospital intervention, but following treatment in the field, no further care is necessary. Thus, multiple dimensions must be considered by prehospital care providers in assessing the risks and needs of a given patient. These low-acuity patients for whom high priority transport via ambulance to an ED is not indicated present with a wide array of signs and symptoms, often in the context of multi-morbidity and age-related frailty [8,49]. These patients furthermore often present with non-specific complaints which are difficult to differentiate using existing methods [43,49,50]. It is the complexity within this patient cohort, both in terms of potential treatment pathways and clinical presentation, which motivates this doctoral project.

The hand-over of a patient to a definitive care provider or the determination that self-care is sufficient to meet the patient's needs constitutes the end of the prehospital care episode. Information about what treatment a patient receives following the prehospital care episode is however critical in evaluating and improving the prehospital care process. By observing post-handover care processes and outcomes, conclusions may be drawn regarding the accuracy of decisions made by prehospital care providers. Commonly observed outcomes of this type in the literature include the level of care and interventions provided by the receiving facility, final diagnoses, subsequent recontact with the prehospital/emergency care system, and patient mortality [13,17,29,43,51].

Clinical Decision Support tools

Clinical Decision Support (CDS) systems are today widely used in prehospital care. Almost exclusively, these are rule-based *expert systems* whereby clinical domain knowledge is encoded into a set of pre-specified decision rules. These systems include both tools for assessing risks within specific clinical cohorts, and tools for triaging patients across a broad spectrum of clinical presentations. Examples of the former include qSOFA and SIRS for screening patients for sepsis [52], the Cincinnati (a.k.a. FAST) and NIH stroke scales [53], and sets of trauma activation criteria such as those originally promulgated by the American College of Surgeons [54]. These condition-specific risk assessment tools are of secondary interest in this thesis.

CDS systems appropriate for use across the heterogeneous patient populations cared for by prehospital care systems are employed at both dispatch centers and on ambulances. In the United States and United Kingdom, the Medical Priority Dispatching System is the *de facto* standard for triage at dispatch [55,56]. In the Scandinavian context, criteria-based systems including the Norwegian medical index and its derivatives are most common [57,58]. At SvLc an internally developed CDS system (*Medicinskt Beslutsstöd*, or MBS) is used. This expert system is based on the Advanced Medical Life Support assessment structure which is commonly used in prehospital care internationally [59]. The system generates a five-level prioritization decision based on an initial patient complaint which is modified by answers to follow-up questions specific to the complaint category. Priorities 1A and 1B differentiate between higher- and lower-priority lights and sirens responses. Priorities 2A and 2B similarly differentiate between ambulance responses travelling according to normal traffic rules. If no findings indicating a need for prehospital care are identified using the system, a referral is recommended. Given the relatively high level of education of the users, wide latitude is given to deviate from the priority recommendation based on their clinical judgement of the situation [60].

The most commonly used triage system among Swedish ambulance services is the Rapid Emergency Triage and Treatment System (RETTS), which combines a clinical symptom categorization with a vital sign scoring system, resulting in a five-level prioritization scale [61]. Perhaps the most internationally popular general-purpose risk assessment tool is NEWS, which was originally developed in the United Kingdom to predict clinical deterioration in the hospital setting. It has since been validated in a variety of settings including prehospital care [62–64]. Unlike RETTS, NEWS is entirely based on vital signs which are routinely collected by ambulance crews, and thus can be calculated based on retrospectively collected data. This, along with its ubiquity in the literature, makes it an ideal candidate for comparative evaluations of new approaches to risk assessment.

Recently, risk assessment tools based on statistical methods including Machine Learning (ML) have become popular. The application of ML to prehospital triage typically entails predicting whether some outcome will occur for a given patient and using the predicted probability of that outcome as the basis for decision rules. Only a few cases of ML-based triage are known to the author to be in active use in prehospital or emergency care. In 2018, Levin et al. published a retrospective validation of a ML-based CDS system for use at EDs [13,65], and a non-randomized evaluation of this tool in clinical practice was recently published [66]. In Denmark, a ML-based system to identify cardiac arrest was validated as outperforming human dispatchers in retrospective data, but a randomized trial failed to identify a significant improvement in identification time [14,67]. Nonetheless it is reported to be in clinical use at dispatch

centers [68]. There is also a relatively large number of recent retrospective validation studies reporting positive theoretical results [15,69–73]. More broadly, ML-based CDS have begun to see some clinical use, particularly in the field of image processing and the analysis of electronic medical records. There is however growing concern that many commercially available ML-based tools have not been validated sufficiently, and the extent to which such products are actually used in clinical practice is often unclear [74,75].

Quality in Prehospital Care

Healthcare quality is a complex and multifaceted concept and may also be defined and measured in a variety of ways. Perhaps the most ubiquitous model for understanding healthcare quality was established by Avedis Donabedian in the 1980s. This model proposes three categories of quality measures: *Structures*, *Processes*, and *Outcomes*. Measures of *structural* quality assess the settings and preconditions for providing care; in the context of prehospital care this might be the number of ambulances available per capita, or the training of care providers. Measures of *processes* quality evaluate the care processes that are thought to contribute to beneficial patient outcomes, for instance ambulance response times and adherence to protocols. Measures of *outcome* quality then represent the ultimate goals of healthcare: Rates of subsequent patient mortality or satisfaction for instance. These three categories of quality are considered causally linked, with the structure of the care system impacting care processes, which in turn impact patient outcomes. [76]

Assessment of quality in prehospital care has traditionally been focused on measuring process quality. Ambulance response times are the predominant metric of prehospital care system performance outside of academia, and are often highlighted in the media as a measure of quality [77–79]. Response times are, as suggested by Donabedian's theory, strongly impacted by structural factors such as per capita ambulance density, geography, and clinical practices. The association of this metric with patient outcomes such as mortality is however surprisingly tenuous [80–84], with important exceptions including responses to cardiac arrest [85–87] and severe/penetrating trauma [88,89]. Other common process measures include protocol adherence, documentation completeness, appropriate intervention delivery, and agreement between care providers at different stages of the chain of care [90].

Mortality is often considered the "gold standard" of patient outcomes given its objectivity, but it can be difficult to apply as a measure of quality for a number of reasons. As noted above, mortality can be difficult to impact using the tools available to prehospital care providers outside of a narrow set of patient conditions. Mortality is furthermore substantially impacted by clinical

and demographic characteristics, and is thus subject to substantial confounding and selection bias outside the setting of a randomised trial. The use of survival as a measure of quality has thus traditionally been limited to specific high-acuity conditions where death is common such as cardiac arrest [80,87,91].

As the practice of ambulance non-conveyance becomes more widespread, outcome measures suited to the assessment of low-acuity patients have been increasingly investigated. In a review of outcome measures used in this context, mortality was found to be the most frequently investigated outcome, despite the above issues. Beyond mortality, the majority of outcomes in the literature were found to be based on subsequent contacts by the patient with the healthcare system, for instance hospital admission or re-contact with the ambulance service [18]. The method by which care providers are thought to impact these outcomes is not the effectiveness of their interventions *per se*. Instead, it is expected that care providers will identify patients at high risk for e.g. mortality or admission and convey them to a hospital. As such, these measures generally seek to evaluate the ability of care providers to accurately determine the level of care required by a given patient. The focus of outcome measures in this context is furthermore generally on the direct patient safety hazards associated with underestimating the severity of the patient's condition, or "under-triage". The consequences of the inverse error of "over-triage" is felt primarily at the system level on downstream care providers in the form of overburdened EMS systems and ED overcrowding [32,92,93]. It is thus the ability of care providers to determine the appropriate level of care which has come to dominate the evaluation of care quality with regards to low-acuity patients.

Decisional Capacity

In this thesis, the term "decisional capacity" is used to denote the ability of a care provider to accurately assess patient care needs in a given context. This capacity can be influenced by a wide variety of factors. The extent of a care providers education, training, and experience is for instance thought to affect decisional capacity, although there is little high-quality evidence that this is the case [94,95]. It is also likely that idiosyncratic factors such as a care provider's own degree of risk tolerance impacts decision-making: The nurses investigated in study II for instance ranged from 3 – 41% in the proportion of patients referred to non-emergency care. Beyond the individual characteristics of care providers, organizational and contextual factors such as clinical guidelines, available assessment tools, and the fundamental setting (i.e. remote, on-scene, or in-hospital) in which the care is provided have a substantial impact on decisional capacity.

While not generally articulated as such, one assumption regarding the decisional capacity of care providers is foundational to outcome measurement in prehospital care. This is the assumption that as a patient progresses along the prehospital chain of care, the decisional capacity of care providers increases. That is to say, the decisions of care providers earlier in the chain of care can be judged using the decisions, assessments, and interventions of care providers further along the chain. It is for instance common to consider a patient assessed by an ambulance care provider as requiring high-priority transport to a hospital as having been under-triaged by dispatch if the ambulance was sent with a low priority [96,97]. Similarly, decisions made by paramedics on scene are often judged by the determinations made by physicians at the hospital [98,99]. Given a steadily increasing level of education and available assessment tools, e.g. sight, touch, and vital signs on the ambulance, blood tests and radiological imaging at the hospital, this assumption would seem to hold. However, as the levels of competency and assessment tools available in early phases of the prehospital chain of care increase, the question is muddied: Does a physician over the phone have a greater or lesser capacity to determine a patient's level of care need than a nurse at the patient's bedside? While assuming an increasing decisional capacity would seem to hold by and large given the current state of prehospital systems, this assumption must be considered in selecting outcome measures related to care provider decisions.

A quantitative expression of this assumption may be found in the generally decreasing rates of high-priority transport as the patient moves along the chain of care. In the dataset used in study IV for instance, 45% of patients received an ambulance with a lights and sirens response, while only 16% were transported to an ED with lights and sirens. Some patients may have received stabilizing treatments resulting in a reduced level of urgency, but the bulk of this difference is due to over-triage at dispatch stemming from the lower degree of decisional capacity inherent to the triage of patients over the telephone. The goal of enhancing decisional capacity in this setting can be expressed in terms of shrinking the margin of safety rendered necessary by the uncertainty about the patient's true condition at each point along the prehospital chain of care.

Evaluation

In evaluating the decisional capacity of a care provider or decision-making system, a number of aspects must be considered. For the purpose of this thesis, we define four general categories of concern. The *Alignment* of a system concerns the degree to which the goals of the system as operationalized coincide with the intended goals of the system. The *Performance* of a system concerns the precision with which the system achieves these operationalized goals. The *Generality* of a system (analogous to robustness or transferability) concerns the extent to which a system maintains its performance in new settings.

Finally, the *Fairness* of a system concerns its ability to make decisions which are unbiased across groups of patients. While these properties have some application to the measurement of decision-making processes in general, the following sections will be primarily concerned with the application of these concepts to CDS systems and the development of risk prediction models.

Alignment

Alignment is a term used to describe the extent to which the goals of a system match those which the designers of the system intend it to have [100]. The term is drawn from the emerging field of AI safety research, which tends to focus the potential consequences of misalignment in super-intelligent autonomous systems [101]. Problems relating to alignment are however already apparent in existing CDS systems. The heterogeneity of clinical presentations within prehospital care entails a particular difficulty in defining the need for prehospital care in terms of measurable quantities. This contributes to the wide range of outcomes used in the literature [17,18,102–104].

As discussed previously, outcome measures are often based on the assumption that assessments made by care providers further along the chain of prehospital care can be considered a standard against which earlier decisions can be evaluated [97,105]. It is however difficult to conceive of such a healthcare utilization measure which perfectly corresponds with a prehospital decision. In study I for instance, hospital admission within 72 hours was considered as a reasonable measure by which the decision to refer a patient to non-emergency care could be measured. While it is reasonable to assume that there is some correlation between hospital admission and emergency care need, it is not reasonable to assume that all patients admitted to the hospital required an ambulance response and vice versa. Thus, while hospital admission may be an appropriate *proxy* measure for ambulance need, it is not perfectly aligned with the need for an ambulance. This gap between our true and operationalized goals will lead some degree of error if, for instance, a prediction model for hospital admission is used to prioritize ambulance responses. Problems relating to alignment in this context thus occur primarily due to the difficulty of operationalizing our true latent and often poorly defined goals in terms of measurable quantities. These issues are exacerbated by the subjective nature of many healthcare decisions: Different physicians may for example come to different conclusions about whether a patient needs to be admitted to in-patient care. This results in measurement errors with respect to the patients true underlying need for hospital care which are problematic in terms of alignment if they are not randomly distributed.

Mortality is commonly used to validate prehospital risk scores, and is rightly held forth as an objective measure of patient outcome [52,62]. Its use as an outcome in training predictive models may however also lead to particularly troublesome issues relating to alignment. Consider that the goal of the prehospital care system is to prioritize providing interventions which prevent death. To the extent these interventions are effective, the apparent mortality rate among patients receiving an ambulance will be lower than it would be in the counterfactual case of not receiving an ambulance. It is in fact precisely those patients whose deaths can be prevented that we would like to prioritize the most. We are however typically not able to directly observe these counterfactual deaths in retrospective data since the patients we have data on received an ambulance. These relationships are illustrated by Figure 3 below:

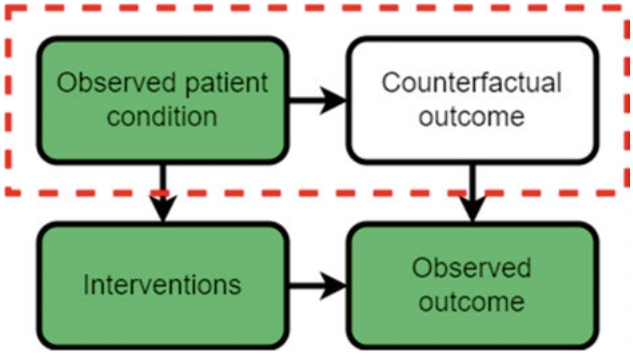

Figure 3- Graph of causal relationships. The relationship between the nodes denoted by a red outline is the relationship we are interested in measuring. Green backgrounds indicate observed outcomes, white backgrounds indicate unobserved outcomes.

This figure illustrates how observed outcomes are a product of (at least) two distinct causal pathways, one relating the underlying risk of the outcome occurring without intervention, and another relating to the effect of those interventions. If a model were naively developed to predict observed mortality and used to prioritize ambulance responses, the model estimate would be biased by the life-saving interventions provided by the ambulance. The model would thus tend to prioritize patients whose death is unavoidable over patients who can be saved by a prompt ambulance response. This issue arises as our true goal entails asking related to counterfactual mortality, i.e.: “What is the likelihood of dying given if no interventions are performed?”, whereas our measure of mortality as operationalized include the effect of ambulance interventions. These issues are often side-stepped by focusing on predictive- rather than causal modelling [106], but it cannot be ignored if clinical risk prediction models are to see widespread deployment. Addressing such causal questions is an unsolved problem and the topic of active research [107,108]. The outcome compositing approach described later constitutes one possible approach to mitigating sources bias arising from the fundamental difficulties in answering such causal questions.

Performance

Performance is here considered the ability of a system to predict a given outcome within the context it was developed. The outcomes investigated in the literature on prehospital and emergency care as described above are often binary. Studies on prehospital triage often seek to evaluate decisions with respect to over-triage and/or under-triage, i.e. overestimating/underestimating the true acuity of the patient's condition [13,109]. The costs associated with these errors are generally recognized to be unequal – It is for instance generally considered worse to miss a case of true ambulance need than to send an ambulance unnecessarily [110]. The precise relative costs of over- and under-triage are however unclear and may vary depending on circumstances.

While the ultimate decisions made in healthcare are often binary, the degree of confidence with which they are made is not. A binary decision can be thought of as the level of confidence in a decision crossing some threshold in the mind of the decision maker, including some margin of safety as discussed previously. This can be considered in the context of a call to the EMD center: At the beginning of the call, the care provider has no information about the caller's condition, and proceeds to interview the patient. At some point, the provider concludes that they have sufficient information and makes a decision. The confidence with which a decision was made is however not typically available in medical records and is often only captured in the context of scientific studies [111].

Many decision support tools make these confidence levels explicit by generating continuous- or ordinal-scaled risk estimates: NEWS for instance results in a risk score between 0 and 21 [112]. Most statistical methods for predicting binary outcomes similarly result in a continuous estimate of the probability of the outcome occurring, which is then converted to a binary class prediction by comparing the estimate to some threshold value. This threshold is selected to correspond to some desired trade-off between e.g. sensitivity and specificity. The overall ability to discern between higher- and lower-risk patients without reference to specific threshold values is referred to as its *discriminatory power* or *discrimination*. Given our interest in evaluating the overall performance of ML models rather than their ability to make specific decisions, comparing the discrimination of various methods was the preferred approach performance in studies III-IV. This is however not the only method available for evaluating the value of continuous risk scores: Decision curve analysis for instance allows a researcher to determine appropriate threshold values based on the relative costs of false negative and false positive cases [113]. In addition to a high degree of discriminatory power, the predictions of a well-performing system should correspond to the true probability of the outcome occurring. The property is denoted at the system's *calibration*.

While the methods described above apply to any decision-making process, evaluating the performance of complex ML models such as those employed in this project involves some additional issues. The simple models typically used to answer research questions (e.g. regression models with a small number of parameters in relation to the number of observations) tend to have a low risk of "over-fitting" the data they were trained on. This phenomenon entails a model fitting to statistical noise or irrelevant patterns which happen to be present in a sample, rather than the true data generating process present in the full population. Prediction errors due to variance arise from this which are typically estimated and accounted for when reporting results through the use of hold-out samples, bootstrapping, or cross-validation [114]. These approaches establish the validity of a given model with regards to the population represented by the sample it was trained on, but not necessarily the ability to generalize to new populations.

Generality

Predictive models are invariably developed based on retrospective data in a specific setting but can only be applied prospectively. This entails a conundrum as no model can be guaranteed to maintain its predictive value over time, nor can it be guaranteed to function in new settings. Furthermore, as a model grows in terms of the number of included predictors and parameters, the risk also grows that it will capture idiosyncratic patterns which vary across time and between settings. Confidence in the ability of a model to maintain its predictive value over time and be resistant to "concept drift" as well as to system shocks such as pandemics must therefore be built empirically [115]. If one has the aim of using models in settings where data to train prediction models is not available, then confidence must also be built that the model maintains its performance when extended to new settings of a similar type.

Demonstrating the *generality* of a risk prediction model in terms of these features is thus crucial to ensuring the safety of decision support systems, especially those based on complex predictive models, when they are applied continually over time in a given setting, and extended to new settings. The former can be evaluated for instance by testing model performance in prospectively gathered data as done in study III, while evaluating the latter required testing the performance of the models in new contexts, which was the aim of study IV.

Fairness

It must further be ensured that decisions are made fairly across various subgroups of patients, particularly with regards to vulnerable groups in the population in terms of gender, age, ethnicity, and socioeconomic status. As predictive models use data regarding historical patient assessments and outcomes, any biases present in past decisions will thus be "baked in" to prospectively

applied risk prediction models. While it is generally agreed that models ought to be fair, there is substantial debate in the literature regarding how to achieve or even measure fairness [116,117]. Considering fairness purely in terms of the distribution of risk predictions and outcomes, it may be evaluated in terms of the equality of treatment or the equality of outcomes. In other words, should the predictions have a similar interpretation (i.e. be well-calibrated) across groups, or should the predictions result in similar outcomes (i.e. have levels of sensitivity and specificity) across groups? These two goals are mutually exclusive, and any *post hoc* adjustments of model predictions to improve one will result in a degradation of the other [118]. This situation is further complicated by the fact that the patient characteristics we believe may be subject to undesirable bias are often also associated with the health status of the patient. We may for instance worry that men are taken more seriously than women calling the emergency number (i.e. that there is some bias on the part of dispatchers we wish to remove). Male sex is however also likely an independent predictor of mortality given shorter life-spans, greater risk-taking behaviors [119,120], and selection effects stemming from differing propensities to contact emergency hotlines. [121] Separating appropriate from inappropriate sources of bias is likely requires delving beyond the analysis of predicted outcomes to identify the underlying causal models resulting in those predictions. This is a difficult problem to be sure, and is the topic of substantial ongoing research [107,122,123].

Enhancement

In this thesis, two data-driven approaches to enhancing decisional capacity are considered. The first is the delivery of personalized feedback reports to care providers, and the second is the use of risk prediction models as clinical decision support tools. This section will provide a general background on the theory and context of these interventions, while their implementation is described later in the thesis.

Feedback

A lack of feedback to care providers regarding outcomes for their patients is often cited as a barrier to achieving optimal decision-making in the prehospital settings [95,124]. In many prehospital care systems, ambulance services and the hospitals to which patients are handed over are not integrated, and information regarding patient outcomes must often be sought out by care providers on an *ad hoc* basis. Care providers are thus rarely able to gauge whether the assumptions made regarding the condition of specific patients were correct. Care providers also have little ability to compare their behavior to that of their colleagues in a quantitative manner. Behavior change theory suggests that both of these are key components of learning [125].

To address this deficit, feedback on the performance of care providers can be provided. In prehospital care, this can take the form of qualitative discussions (e.g. post-event debriefings), performance data delivered to care providers, or some combination of thereof. Often, feedback interventions are directed towards improving care within specific clinical categories, e.g. cardiac arrest, stroke, or trauma. The effects of such interventions have generally been found to be moderate, with the strongest impacts being found on measures of process quality such as documentation and protocol adherence. [124,126]

The literature suggests a number of considerations which contribute to achieving effective clinical feedback interventions [124,127,128]. The most comprehensive framework is Clinical Performance Feedback Intervention Theory which synthesizes a set of high-confidence findings into a feedback cycle consisting of a set of mechanisms necessary for successful feedback delivery [128]. The literature suggests that feedback should ideally be *timely*; delivered in close proximity to the event, and *credible*; transparent, statistically accurate and delivered by trusted sources. The information presented in a feedback report should be displayed both visually and in numerical form, with a minimum of complexity, and be accompanied by an appropriate reference measure. The intent of providing the feedback should be clear, and specific recommendations for action should be provided [127]. Many of these considerations were taken into account in designing the feedback intervention described in study II.

Clinical Decision Support

Studies III – V in this thesis are directed towards developing tools to provide prospective decision support to care providers. A CDS will generally consist of some underlying engine for generating recommendations, and an interface for presenting the recommendations to the user [129,130]. The studies involved in this thesis are primarily concerned with developing and evaluating the engines used in CDS systems. Nonetheless, particularly in study V consideration had to be given to the mechanisms by which CDS influences care providers. In qualitative research, CDS was considered particularly useful in rare situations, and in situations involving patients with unclear or vague symptoms. Dispatchers are also under substantial time pressure to make decisions, which must be considered in designing appropriate CDS systems [60].

Whereas the basis for recommendations made by traditional knowledge-based CDS systems are relatively simple to communicate, communicating the factors contributing to recommendations made by more complex machine learning-based systems is a challenge. As such, the *explainability* of risk prediction models has come to the forefront as a topic of research and has been suggested as a key to ensuring the reliability and usefulness of CDS. Explainability as a concept may be relevant at several levels, from estimating measures of variable importance for model development and validation, to

generating explanations for end users regarding individual predictions [131]. These factors must also be considered in the context of the already high cognitive load placed on care providers, and an excess of information perceived to be irrelevant may result in "alert fatigue", reducing the impact of CDS and potentially the overall decisional capacity of care providers. Taken together, these factors and discussions with users pointed towards developing a rather minimalistic interface for the CDS tool under study, minimizing the number of interactions necessary to include patients, and limiting the amount of information provided to users regarding model assessments.

In providing this type of decision support we seek to offload some of the cognitive burden placed on care providers. Care should be taken in considering exactly what burdens we are lifting. Whereas medicine often involves relatively intricate diagnostic decision-making, prehospital triage rather involves ruling out emergent conditions with a sufficient degree of confidence. In this sense, risk prediction models are a good fit. Nonetheless, it is important to recognize that statistical models are, to borrow a term, golems – single minded creatures created through wizardry, dedicated entirely and unflinchingly to a task set out before them [132]. Our model may thus accurately predict some outcome, but it remains the task of the human care provider to interpret the prediction in the context of an individual patient. In developing interventions using prediction models, we should strive to identify the strengths and weaknesses of human and algorithmic decision-making processes. Prediction models are highly capable in terms of rapidly reducing subtle patterns in large amounts of information to simple quantitative measures. It is however arguably the hallmark of human-level intelligence to operate flexibly across multiple levels of reasoning, jumping out of a given system of logic as deemed appropriate [133]. A model cannot itself reason about factors outside of the model, and as such a human must remain as our final defense against the inevitable misalignment of such tools with our underlying goals. As machine learning systems become increasingly substantial and able to embed ever more layers of isomorphism to reality within their parameters, this differential aptitude for "out-of-system" thinking may perhaps erode. Whether such developments entail an increased or decreased need for human oversight only time will tell. Perhaps we will know the answer only after our course in this regard is irreversibly set.

Methods

Data Sources and Collection

Obtaining the data necessary to characterize the prehospital care decision process is often a major barrier to performing prehospital research [134]. There are two general approaches to gathering retrospective quantitative outcome data. These are manual abstraction through review of medical records and automated extraction from electronic medical record systems. The former is often employed when the desired variables are difficult to define, require some human judgement to assess, or when variables must be drawn from unstructured data such as text or audio recordings. Manual abstraction may also be necessary when a technical data linkage between various sources simply is not available, as was the case when the data collection for study I was performed. While manual review has some comparative advantages, it is time-intensive and not a feasible approach to gathering the large amounts of data necessary to develop machine learning models such as those investigated in this project. The learnings from study I were therefore used to define the technical specifications for the automated extraction process upon which the dataset used in the remainder of the studies are based.

For the Uppsala and Västmanland regional prehospital care systems, data from all parts of the prehospital chain of care described previously in Figure 2 is available via regional quality assurance systems. These combine information gathered from the *Alitis*® dispatching system, *Mobimed*® ambulance electronic patient care report system, and *Cosmic*® regional medical record system. The availability of these data enables patients to be followed from the point of contact with the dispatch center to discharge from hospital. Data from the dispatching system is first linked deterministically with ambulance journal records based on a common case identifier. The patient's national Personal Identification Number (PIN) as recorded from these sources is then used to search the regional medical record system for the first physical healthcare contact within 72 hours of contact with the prehospital care system.

This approach was used to gather the data for studies II, III and V, while study IV involved the development of a variable specification document which was based on the data extraction performed in Uppsala and Västmanland. This was then distributed to the participating regions and extracted locally by regional staff and delivered to the research group in a pseudonymized format.

Feedback

The provision of routine feedback to care providers was identified as a potential simple and low-cost avenue for improving decisional capacity at both EMD centers and ambulance services. An intervention was thus designed whereby each month, every dispatch nurse at SvLc handling a sufficient number of patient calls (at least 50) received a feedback report via email. The reports contain a set of 14 indicator values divided into descriptive, process, and outcome measures. Each measure was presented visually and in text, with a reference group consisting of other nurses at the same dispatch center. The performance metrics were visualized as seen in Figure 4 below:

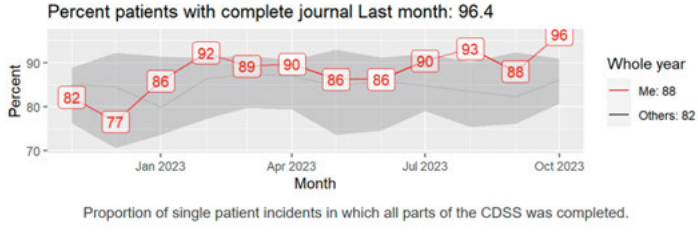

Figure 4 - Example feedback visualization. Red line and values denote results for the individual nurse, grey area and line indicated the interquartile range and median for other nurses.

Whereas process and outcome measures had a desired direction (e.g. high for documentation completeness, and low for under-triage rates), descriptive measures (e.g. the proportion of priority 1 dispatches) were provided without a specified "correct" direction or level, but rather intended to enable providers to gauge their own relative level of risk-taking regarding e.g. priority setting. In selecting the measures, care was taken to achieve balance by including both measures of over- and under-triage. Prior to their automated distribution via email, the reports and the included measures were reviewed and modified in staff meetings. Automatic monthly distribution was preceded by the use of the reports in yearly staff review meetings where care providers were given the opportunity to discuss the measures at each site.

Statistical Learning

The risk prediction models applied in studies III-V of this thesis are generally considered to be a form of Machine Learning (ML). This is a label can refer to a wide variety of methods, which can be generally divided into methods for supervised and unsupervised learning. Supervised learning methods are those which seek to predict the value of an outcome based on some set of inputs. Unsupervised methods instead seek to identify associations within the inputs. [114]

Unsupervised Methods

This thesis is primarily concerned with supervised learning methods, with an important exception being the image printed on its cover. *Unsupervised* statistical methods are commonly employed to e.g. identify clusters in data, reduce the dimensionality (i.e. the number of variables) of data while maintaining information, or generate synthetic datasets [135]. The cover image represents the results of applying a Principal Component Analysis (PCA) on the data used in study IV. PCA is a dimensionality reduction technique which performs a linear transformation of the data into a new coordinate space such that the largest amount of variation in the data is captured by the smallest possible number of dimensions [136].

In the case of the cover image, 44 input variables are represented by 2 dimensions (the x and y axes) which together capture ~20% of the variation in the dataset. Each point represents a single ambulance mission which by virtue of the linear transformation lies near other missions with similar variable values. Finally, to give the figure some pizazz each mission was coloured based on its NEWS score. While omitted from the cover image for the sake of maximizing its beauty, the eigenvalue or "direction" of each included variable may be included for interpretation as in figure 5.

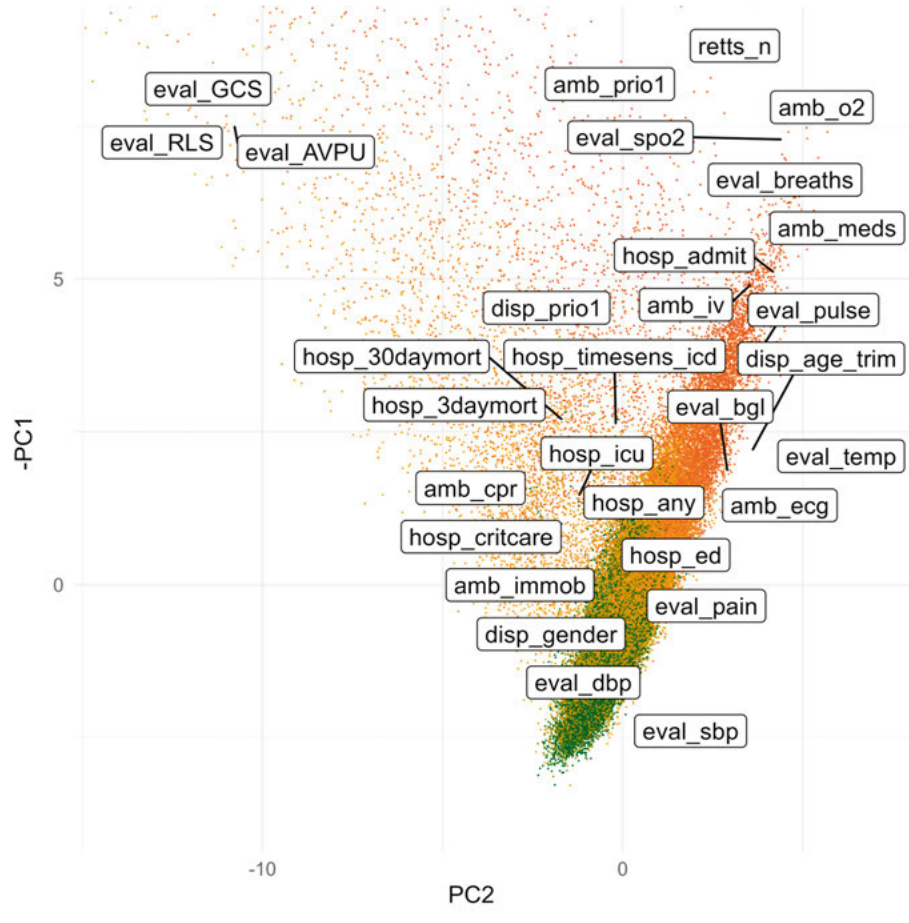

Figure 5 - Principal Component Analysis of data from study IV with labelled eigenvectors for a selection of variables with varimax rotation

As with other unsupervised learning techniques, this plot has no intrinsic meaning, but must be imbued with one by the viewer. There are at least two striking features of this plot worth noting. The first is that the vast majority of ambulance missions are arrayed in a diagonal band from bottom to top right, with a smattering of missions drawn up and to the left. The second is that the variables associated with missions pulled to the left are associated with true emergencies: Reduced patient consciousness, hospital critical care, performance of CPR, and mortality. Variables associated with missions pulled to the right are instead indicators of more substantial demand for definitive care: Hospital admission, administration of medications or oxygen, and patient age, for example. One interpretation of these features could be that patients with true emergencies which require the full capabilities of ambulance care providers are rare as suggested by the sparsity of patients pulled left. Within the population of non-emergency patients, there is furthermore substantial variation

regarding the need for definitive care. While such interpretations are qualitative, this sense of multiple dimensions of patient care need motivates and underpins parts of the outcome compositing approach described later.

Supervised Methods

Within the scope of *supervised* learning, a further differentiation may be made between *parametric* and *algorithmic* methods, using the terminology of Leo Breiman [137]. Parametric methods are those which have a finite set of parameters chosen based on a data model selected *a priori*. These are then adjusted to optimize predictions within the confines of that model, typically through some iterative loss minimization or likelihood maximization method. These methods range from traditional linear regression models to more sophisticated methods including regularization, non-linear parameters, hierarchal data structures, etc. Relatively simple parametric models have the benefit of a high degree of interpretability with respect to their parameter coefficients. They are therefore commonly used in analytical applications where the coefficent estimates themselves are of interest, such as the multiple regression models used in studies I and II.

Algorithmic approaches to machine learning involve methods which no not assume a specific data model, and as such make fewer assumptions regarding the properties of the data operated upon. Rather, the inference process can be said to emerge from the data, making the methods highly flexible and thus well suited to prediction tasks across many domains. This flexibility comes at the cost of less innate interpretability, a potentially reduced ability to generalize, and a greater demand on computational power. While technically parametric, modern neural network architectures have a number of parameters ranging in the billions, rendering them so flexible as to be effectively an algorithmic method. Some of the most popular algorithmic approaches to supervised machine learning for structured data of the type employed in this thesis involve the generation of ensembles of prediction models, typically numbering in the thousands. Random forest models for instance train relatively deep decision trees on random subsamples of data which then "vote" to predict the likely outcome for a given case, using the multiplicity of trees to reduce the variance (i.e. error stemming from the limited sample size) of the estimate. Gradient boosting models on the other hand train relatively shallow decision trees in sequence, with each new tree operating on the errors of the previous trees, thus reducing bias (i.e. error due to lack of flexibility in the model) [114]. The latter approach, specifically as implemented in the XGBoost package [138] is the mainstay of the risk prediction models evaluated in studies III-V.

The methodological gulf between these two approaches has led to what Breiman termed the "two cultures" of statistical modelling. [137] Statisticians tend to be more comfortable assuming that data is generated by a given statistically well-defined process, testing those assumptions, and making predictions based on extending those assumptions to new data. Models based on assumed data structures are thus parsimonious, relatively robust, and useful for answering specific research questions given a correctly assumed data generating process. Fields outside of statistics, in particular computer science, have meanwhile embraced algorithmic modelling, taking advantage of modern computing power to accomplish prediction tasks which had previously not been possible to address using traditional data modelling. Each of these approaches in useful in answering questions of different types, as reflected in their application to various tasks in this thesis.

Implementation

Studies III-V are all focused on the development, validation, and clinical evaluation of a common ML-based risk assessment tool. The general approach has been to first develop a model and perform the study analysis using the R programming language. Once the study was completed, the models were then re-implemented using Python, and published on GitHub as part of openTriage [139]. openTriage is an open source software package developed by the author primarily in Python which implements the functionality necessary to train machine learning models and serve model predictions to third-party systems via an Application programming Interface (API) [139]. It consists of a core set of functionality common to a broad set of risk prediction problems in the healthcare domain, and "frameworks" which can be activated to enable functionality related to a specific application. The methods used in studies III-V of this thesis are thus implemented as frameworks in the openTriage software and are publicly available, although the trained models used to evaluate dispatch data are not currently public due to intellectual property restrictions surrounding the CDS system used at the dispatch center. While a number of methods have been explored, all currently implemented frameworks are based on tree-based gradient boosting algorithm using XGBoost [138]. Additional functionality related to hyperparameter tuning using Bayesian optimization [140] and user interfaces including model prediction explanations using Shapley values [141] are also implemented in openTriage. A user interface demonstrating a risk scoring model based on ambulance vital signs as used in study IV may be found online at https://opentriage.net/ui/vitals. While standard methods and well-validated software tools are used in developing the models themselves, these estimates are then combined into composite risk scores in a novel way which requires some justification.

Outcome Compositing

In this thesis, it is argued that bias due to misalignment can be mitigated through the use of composite outcomes and model aggregation. Composite outcome scores have a long albeit controversial history of use in the evaluation of randomized clinical trials and healthcare quality. While the use of a composite outcome may improve the face validity of an outcome or reduce bias, they are also argued to potentially mask findings in the component variables in randomized control trials [142,143]. A multi-stage process is proposed, whereby composite outcomes are first defined to represent distinct and recognizable aspects of patient risk, with components selected to mitigate misalignment as described previously. After estimating models to predict these outcomes, they are then weighted and aggregated to align with specific decision-making processes based on clinical domain knowledge or correspondence with existing risk assessment tools.

Compositing is proposed at two stages in the modelling process: Pre-estimation and post-estimation. By pre-estimation compositing is meant the common practice of combining multiple variables into a single outcome. Post-estimation compositing is more novel and consists of combining the results of several estimated models into a single composite score. Studies of models based on post estimation compositing are rare, though a simple composite score has been previously used for triage in the emergency department setting [13]. Averaging the predictions of multiple risk prediction models for the same outcome is also a common approach to improving machine learning model performance [114].

The goal of pre-estimation compositing is to define an unbiased measure of some property of a patient's condition which is recognizable and clinically relevant to the users of the system. By training separate models for each of these elements separately, interpretable risk estimates for each can be provided to the end users of the system. Pre-estimation compositing can also be used to address problems of missing data common in prehospital care [144]. While ambulance care providers may for instance forget to document a single intervention, they are less likely to forget to document all of their interventions. A composite outcome based on the documentation of any one of several interventions is thus argued to be more robust to missing data in the form of omissions resulting in false negative predictors.

Decisions in prehospital care must often be made under substantial time pressure, and care providers do not always have time to carefully weigh the various risks associated with multiple outcomes and consider them in light of the decision at hand. After defining a set of relevant composite outcome measures, a method was therefore devised to quantify value judgements made by the

clinical leadership about the importance of various outcome measures to a given decision. While factor analytical techniques for instance can be useful in understanding the relationships between the available data, ultimately a clinical value judgement about the importance of each outcome must be made. In study III, this was done merely to illustrate the impact of the process on model performance with no specific application in mind. In study V however, we sought to support a specific decision, namely: Out of a given set of low-acuity patients awaiting an ambulance, which patient has the most immediate need of an ambulance response? For this task, a single value representing ambulance need had to be generated which would be valid in the context of comparing risks across the heterogeneous patient population served by the dispatch center. The proposed process of generating composite risk scores is recapitulated below, using the considerations involved in defining the composite outcome applied in study V as an illustration:

1. **Identify relevant outcome measures**
   To quantify the need for an ambulance, outcome measures were selected based on those used in study III. These included both data from ambulance records and hospital outcomes. Outcome measures with a poor correlation with overall patient acuity (e.g. immobilization by ambulance crews), or low rates of documentation (e.g. findings from secondary assessments) were discarded.

2. **Define pre-estimation composite outcomes**
   Exploratory and confirmatory factor analysis combined with domain knowledge was used to generate composite outcome measures based on the selected outcome measures. Variables were combined into four distinct composite outcomes representing coherent aspects of pre-hospital care need: 1) Initial assessment (abnormal airway, breathing, circulation, or consciousness findings by ambulance), 2) ambulance prioritization (lights & sirens transport to hospital or critical initial impression by ambulance), 3) ambulance interventions (administration of medications, IV catheterization, CPR, oxygen administration, or pre-arrival notification), and 4) hospital outcome (hospital admission or mortality within 30-days).

3. **Define post-estimation composite weights**
   A weight was assigned to each of the composite outcomes based on a factor analysis, regression using NEWS scores as the dependent variable, and clinical judgement. We considered the initial assessment of the patient to be the most important factor and assigned it a weight of 4. Ambulance prioritization was considered the next most important

and assigned a weight of 2. Both ambulance interventions and hospital outcomes were assigned a weight of 1.

4. **Define post-estimation composite function**
   The variance of the predicted values for each of the composite measures differed, with the variance of estimates of balanced outcomes (i.e. those with closer to 50% prevalence) exceeding estimates of unbalanced outcomes. As such, we normalized logit-transformed probabilities by subtracting the mean and dividing by the standard deviation of the predictions in a test dataset. After this normalization, the predicted probability of each composite outcome occurring was multiplied by its weight, and the average of the weighted probabilities was taken as the final risk score.

This process resulted in the risk assessment instrument used in study V and described in Figure 6:

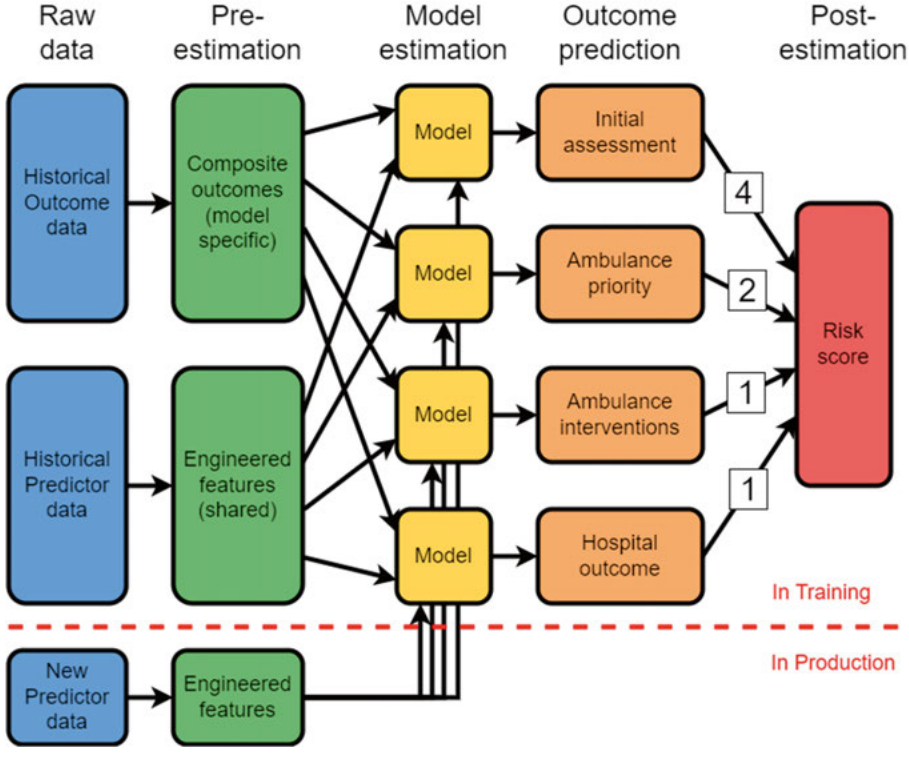

Figure 6- Diagram of compositing process

As illustrated, in training models are developed based on historical outcome and predictor data which are pre-processed to generate the composite outcomes described in step 2, and to generate feature sets suitable for use in the gradient boosting models, respectively. Models are trained based on these data, and raw predictions (logits) are then weighted and averaged to result in an overall risk score. In production, predictor data are pre-processed according to the same method, and predictions are made using the trained models.

While this process is believed to result in a measure which is reasonably well aligned with ambulance care need, no claim is made that it is optimal. There may be important additional outcomes to consider or differing clinical opinions about the appropriate weight to assign to each outcome. There may furthermore exist more sophisticated methods with better properties to estimate these joint likelihoods. Nor is it necessarily the case that the optimal model is a simple weighted average – Perhaps there are interaction, non-linear, or threshold effects regarding the outcomes which should be accounted for?

Study Designs & Statistical Analyses

The studies included in this thesis employ a wide variety of investigative tools to evaluate decisions made by care providers, intervention effects, and the performance of risk prediction models. This section provides a grounding in each of the key study designs and statistical methods used in the studies included in this thesis.

Study I

Study I is a descriptive, observational study examining the characteristics of patients referred by EMD nurses to non-emergency care. Data for study I was collected manually by two nurses abstracting data from hospital records in each region using a guide based on the "trigger tool" method more commonly applied for identifying patient harms in hospital care. [145]

Being a descriptive study, the main results are based on descriptive statistics. In estimating the confidence intervals for the descriptive statistics included in the study, a technique called bootstrapping was used. This highly flexible method entails estimating the distribution of a parameter of interest in a population by repeatedly resampling data from a given sample with replacement many times. The distribution of the parameter estimates in the resampled data can then be used to generate statistics of interest regarding the population, in this case 95% confidence intervals. [146] Bootstrapping confidence intervals for simple descriptive statistics is admittedly a bit overkill. The approach however makes for an elegant analysis as opposed to using confidence intervals based

on varying formulas based on asymptotic assumptions regarding continuous, binomial, and non-parametric estimators. Given the computing power commonly available today, there is little reason not to use bootstrapping when an estimate of the variability of some parameter is desired.

In addition to descriptive analysis, some associational analysis was undertaken. Regression models were used to evaluate the correlation between various patient and call characteristics and the likelihood of the patient visiting an ED following referral to non-emergency care. Given the lack of a clear causal model, it is however important to interpret the results of such analyses conservatively. Some specific tests were further used to evaluate claims made in the manuscript. Cronbach's Alpha was used to assess inter-rater reliability in performing the data abstraction in a sample of overlapping data from one month [147]. Two specific *post hoc* claims regarding patients experiencing patient harms were tested using a non-parametric Wilcoxon rank-sum test.

Study II

Study II is an interrupted time series analysis employing a pragmatic stepped-wedge design. The study aims to evaluate the impact of the feedback intervention described previously, exploiting the staggered implementation of the feedback reports across study sites. By analysing multiple interventions separated in time, the design minimizes the risk of history bias, i.e. the impact of other interventions or changes in population characteristics occurring simultaneously with the intervention under investigation. [148,149]

To estimate the intervention effect, data at the individual nurse-level were aggregated per month, forming a set of panel data. This type of repeated measure data violates assumptions regarding independent and identically distributed random variables necessary for valid inference using simple regression models. There are a number of approaches to handling clustering of this type, of which hierarchical (a.k.a. mixed or multi-level) regression models were chosen due to their flexibility and adaptability to many types of research questions and causal models [150]. To estimate intervention effects, fixed effects were calculated for the intercept, full sample slope, post-intervention effect, and post-intervention slope. Random intercepts were included at the nurse level to account for clustering. In the case of binomial outcomes, the model specification was thus:

$$\log\left[\frac{\hat{P}}{1-\hat{P}}\right] = \alpha_{j[i]} + \beta_1(\text{Month}) + \beta_2(\text{PostIntervention}) + \beta_3(\text{SinceIntervention})$$

$$\alpha_j \sim N\left(\mu_{\alpha_j}, \sigma^2_{\alpha_j}\right), \text{for Nurse j} = 1, \ldots, \text{J}$$

The specification for continuous outcomes was similar, but with a log transformation of the outcome, rather than logit. *Month* was specified based on the calendar months since the start of the study period across all sites, accounting for any common secular trends in the data. The *PostIntervention* and *MonthsSinceIntervention* parameters were estimated based on indicator variables stemming from the month the first feedback reports were delivered at each region (0 or 1 for *PostIntervention*, 0 or # months since intervention for *MonthsSinceIntervention*). Intercepts are assumed to be normally distributed across nurses included in the study. The estimation framework thus assessed both immediate intervention effects and longer-term trends in effects, while accounting for the dependencies induced by including repeated measures from each nurse.

Study III

Study III is a model derivation and internal validation study. This study aimed to assess the performance of a machine learning-based risk prediction model based on the outcome compositing approach described previously in comparison with an existing rule-based triage instrument (NEWS) and human priority-setting.

After evaluating a number of approaches, gradient boosting as implemented in the XGBoost package [138] was used to develop two composite risk prediction models based on dispatch and ambulance data, respectively. These models were then compared with NEWS and the dispatch priorities set by nurses. The models were compared primarily in terms of their overall discriminatory power as measured by the C-index [151]. The C-index as a measure of discrimination has a pleasantly intuitive interpretation: It ranges from 0.5 for random guesses to 1 for perfect predictions. Values between these extremes correspond to the probability that a randomly selected pair of cases will have the higher risk prediction assigned to the positive case. The C-index is equivalent to the area under a receiver operating characteristic curve. This curve, whose archaic name stems from its original application characterizing radar detection during the second world war [152], is generated by plotting all combinations of sensitivity and specificity possible by varying the confidence threshold applied to a continuous predictor of risk. The curves of several models may thus be compared to determine which predictor is superior given any desired trade-off based on the relative costs of false positive and false negative findings.

As the complexity and flexibility of the models used to estimate patient risks increase, overfitting becomes an issue which must be adjusted for. In this study, the primary results were reported based on predictions in a test dataset consisting of data collected from the year after the data used to train the models were collected. Thus, both the issue of overfitting and issues relating to

potential loss of prospective performance over time were accounted for and evaluated.

This was also the first study in which missing data could not be adequately handled using appropriate exclusion criteria to limit the scope of the research question. Missing data in prehospital care is a widespread issue [144,153], and is particularly problematic when missingness is not random, i.e., the underlying data is correlated with the probability of data being missing [144]. Types of missing data are typically classified using the terminology given by Rubin [154]. Data Missing Completely At Random (MCAR), denotes missingness which occurs in an unsystematic manner and can thus be handled without inducing bias using complete-case analysis by excluding observations with missing values. Data Missing At Random (MAR) denotes a pattern of missingness in random conditional on the observed data, i.e. the likely values of the missing data can be inferred from the data which is not missing. Finally, data Missing Not At Random (MNAR) entails that the reasons for missingness are not known, and missing values cannot be inferred.

Multiple imputation using chained equations offers a principled approach to handling missingness, which given MAR or MCAR assumptions are fulfilled, results in unbiased point estimates and confidence intervals with good coverage properties [155]. This approach involves imputing missing values based on the values of the data which were captured, and iteratively updating the imputed values based on a selected model, often predictive mean matching or a non-parametric decision tree-based method. This process is then repeated multiple times to create several datasets with varying imputed values [156] The primary drawback of using multiple imputation approaches is the additional complexity it induces in the analysis, requiring necessitating parameters of interest to be estimated in each dataset, and then combined using Rubin's rules to combine within- and between-imputation variance [154,157]. In this study, missing vital parameter data (e.g. pulse, blood pressure, etc) was identified as an issue in calculating NEWS scores, and was handled partly via the application of exclusion criteria, stipulating that at most 2 vital signs could be missing to be included. For patients meeting these criteria, it was considered that remaining missing vital parameters were MAR, and imputation using predictive mean matching was applied. The degree of between-imputation variance in NEWS scores following the application of these criteria was found to be minimal, and single-imputation was used to calculate comparator NEWS scores to reduce the complexity of the analysis.

Study IV

Study IV is an internal-external validation study building on the models developed in study III. The study aimed to characterize the loss in performance experienced by machine learning models when applied to new contexts. The study investigated two outcomes (hospital admission and 3-day mortality) using two ML approaches (Gradient boosting and logistic regression using splines to handle non-linear effects) in a dataset collected from six Swedish regions. This study focused on the use of predictors available to ambulance care providers on-scene prior to conveyance, guided by the hypothetical use case of a risk assessment tool to support non-conveyance decisions.

Given the potential for overfitting to be an issue, k-fold cross-validation was applied in estimating internal model performance in this study. This involves a process whereby a dataset is split into a number of randomly selected "folds" which are successively excluded from the sample used to train the model [114]. Using 10 folds for instance, 10 models are fitted each randomly excluding 10% of the available data. Each excluded fold is then used to evaluate predictions made by the model excluding it. While the approach of using randomly selected folds appropriately accounts for overfitting in the model training process, it does not address sources of systematic variation between settings which may impact the generalizability of a model. Thus, a non-random cross-validation process was used to estimate external performance by using each of the 6 regions as a separate fold. Missing data was handled via single imputation in a manner similar to study III and with a similar justification.

Model calibration was assessed in study III, but given a well-specified prediction model, internal calibration is rarely an issue. In this external validation, assessing the calibration of the models when applied in new contexts played a larger role. A well-calibrated model entails that the predicted probabilities generated by the model correspond to the actual probability of the outcome occurring. This can be assessed "in the large" by calculating the overall mean of the predicted probabilities and outcomes and comparing the two. Such measures (such as the Observed:Expected ratio used in this study) evaluate the level of overall over- or underestimation of risks in the full population. Calibration may however vary across levels of predicted risk, and a more nuanced assessment of calibration may be achieved by fitting calibration slopes or, as done in this study, directly visualizing calibration across strata of predicted values [158].

Study V

Study V is a Randomized Controlled Trial (RCT) of a machine learning-based risk assessment tool for prioritization in Resource Constrained Situations (RCSs). The openTriage risk prediction tool was implemented in the Alitis dispatching system through a Graphical User Interface (GUI) which called upon the risk assessment tool via API as necessary.

Patients were included in the study by dispatchers based on a set of inclusion and exclusion criteria. Inclusion criteria were the documentation of a personal identification number, a priority of 2A or 2B, and an age over 18. In cases where two or more patients meeting these criteria were simultaneously awaiting an ambulance, dispatchers were able to select them for comparison when an ambulance became available and a choice had to be made regarding which patient had the greatest medical need. In addition to the inclusion criteria which were applied automatically, additional exclusion criteria were applied by dispatchers, excluding patients where non-clinical risk factors or mitigators were present (e.g. being exposed outdoors in the cold or having care providers on scene), where logistical factors such as differential transport times played a role in the decision, or where one patient had been waiting much longer (i.e., more than 30 minutes). The study thus sought to assess the accuracy of the clinical decisions made within each of these RCS.

After the selection of patients to be included, each RCS was randomized 1:1 to either a control or intervention arm. In the control arm, risk scores were calculated but not displayed to the user. In the intervention arm, risk scores were calculated and an indication was displayed of which patient had the highest risk score, along with a colored indicator of model confidence. A pop-up window containing additional information about the model prediction could be displayed if desired by the user. Depending on treatment arm assignment, the dispatcher then either followed standard practice and conferred with the nurses assessing the patients about which patient to send the first ambulance to (control), or did so while also considering the risk assessment determination (intervention). The patient in the RCS with the first ambulance assigned to them after inclusion was considered to have been prioritized.

The primary outcome of the study was based on the patient's NEWS, estimated as the difference in the proportion of prioritized patients who had the highest NEWS score across the treatment arms. NEWS was selected party due to its widespread use as a measure of patient risk in emergency care, and due to its relatively objective nature, being based on patient vital signs rather than care provider assessments. As compared to using the outcomes the composite risk score is based on, using NEWS reduces the risk that post-randomization dispatcher behaviour could bias the outcome measure. This also prevents

misalignment of the risk prediction outcome specification from giving the intervention group and unfair advantage.

The required sample size for the study was based on a simulation of data drawn from historical ambulance missions. It was difficult to identify historical RCS where a decision was made purely based on the patient's clinical condition. To estimate dispatcher accuracy, it was instead assumed that a patient prioritized as 2A would always be selected over a patient prioritized as 2B, and patient NEWS values in a random sample of 2A vs 2B patients were compared to the prioritization order proposed by the risk assessment tool in the same population of patients. Taking the difference in accuracy between the risk prediction algorithm and priority 2A vs 2B as the target effect size, a sample size of 1500 RCS was identified as necessary to achieve type I and II error rates of 0.05 and 0.2 respectively.

Whereas patients with substantial missingness in the vital signs necessary to calculate NEWS could be conveniently excluded in studies III and IV, the post-randomization loss to follow-up this would entail in an RCT would introduce an unacceptable risk of bias and loss of power. Given that it was relatively common for patients not prioritized in an RCS to be furnished with alternate transport to the hospital, this had to be addressed in an effective manner. Data was therefore collected from hospital record systems for all patients in Uppsala and for patients missing documented vital signs in Västmanland (as these had to be abstracted manually) and combined with data from ambulance and dispatch records when estimating missing values. While vital signs obtained at the ED are not exactly the same as those which would have been obtained on the ambulance, the additional data was considered sufficient to fulfil MAR assumptions. As between-imputation variance was not negligible in this case, study estimates were pooled using Rubin's rules to generate the coefficient estimates the final study results are based upon.

Results

Study I: Outcome measurement

In study I, a set of patient outcome measures for use in the context of EMD triage to non-emergency care were developed, abstracted from hospital medical records using a Trigger-tool derived methodology, and analysed descriptively. Out of the 1 876 patients referred to non-emergency care in Uppsala and Västmanland over a 4-month period, 1 089 were included in the study. Of these, 258 (24%) visited an ED within 7 days of contacting the EMD. Of those visiting an ED, 208 (81%) had a complaint related to their reason for contacting the EMD center, 143 (55%) received interventions above the primary care level, and 99 (38%) were admitted to the hospital.

Analysis of inter-rater reliability identified a high level of agreement regarding ED visitation and hospital admission (Cronbach's alpha 0.94 for both measures), while agreement regarding the whether the ED visit was related to the EMD contact and the level of interventions was lower (0.71 and 0.70 respectively). The time from EMD contact to ED visit was also analysed, finding that the vast majority (81%) of ED contacts occurred within 3 days, after which the proportion of visits related to the EMD call also dropped off substantially (61% during days 4-7). In performing a multiple logistic regression analysis, two factors were found to be robust predictors of ED visitation: Patient age over 65, and referral of patients onward to another care provider (Adjusted ORs of 1.48 (1.08 to 2.04) and 1.58 (1.15 to 2.19), respectively). Finally, each case resulting in an ED visit was assessed for the presence of avoidable patient harms, with a total of 6 triage errors resulting in a harm to patients. One case involved potentially permanent harms, while several involved delayed pain relief or possibly extended hospital stays. All of these harms occurred within roughly 3 days of EMD contact among patients admitted to the hospital, and none were identified by existing incident reporting systems in the studied regions.

These results confirmed the validity of using quantitative outcome data to guide quality assessment and improvement, and informed decisions made when establishing the criteria for the automated extraction of hospital outcome data in the regions. Based on the results, a 3-day follow-up window was selected as a reasonable compromise between data exposure minimization and

the capture of subsequent related ED contacts following referral. Furthermore, the analysis of avoidable patient harms suggested that reviewing cases in which referred patients were admitted to the hospital within 3 days would capture a majority of patient harms resulting from referral to non-emergency care.

Study II: Feedback

In study II, the effect of a continuous feedback intervention was investigated. Reports on 14 metrics divided into descriptive, process, and outcome categories were delivered to nurses monthly via email. A total of 40 dispatchers in 3 regions were included in the study between March 2020 and October 2023, handling a total of 46 353 incidents. While no significant effects were found for most metrics, documentation-related process measures were found to be improved. The odds of complete documentation increased by 7.5% (95% CI 5.1 to 9.9) and the odds of having a documented contact reason increased by 3.8% (1.5–5.9). These impacts were sustained over time, with slope coefficients suggesting increasing documentation rates compared to the pre-intervention trend. The study results for each metric are displayed visually in figure 7 on the next page. Note that this figure provides a flavour of the analytical results, but does not correspond to the regression models completely, owing the use of a common secular trend parameter and the estimation of the mean value of the transformed measures, rather than medians. The intervention was also found to have resulted in a significant increase in the proportion of manually upgraded patients. While the remainder of effects were not statistically significant, all point estimates were in a direction associated with higher overall priority-setting.

Sensitivity analyses were also conducted to evaluate the impact of excluding time periods associated with the first wave of the covid-19 pandemic, and excluding the months prior to the intervention during which nurses may have been exposed to anticipation effects. These analyses demonstrated intervention effects consistent with the interpretation of the main analysis. We had also originally intended to include a formal analysis of “homogenization effects” expressing themselves as a decrease in between-nurse variability in the descriptive measures. This was found to be more statistically problematic than anticipated, but visual inspection of nurse-level data suggested that this was not occurring to a qualitatively substantial extent.

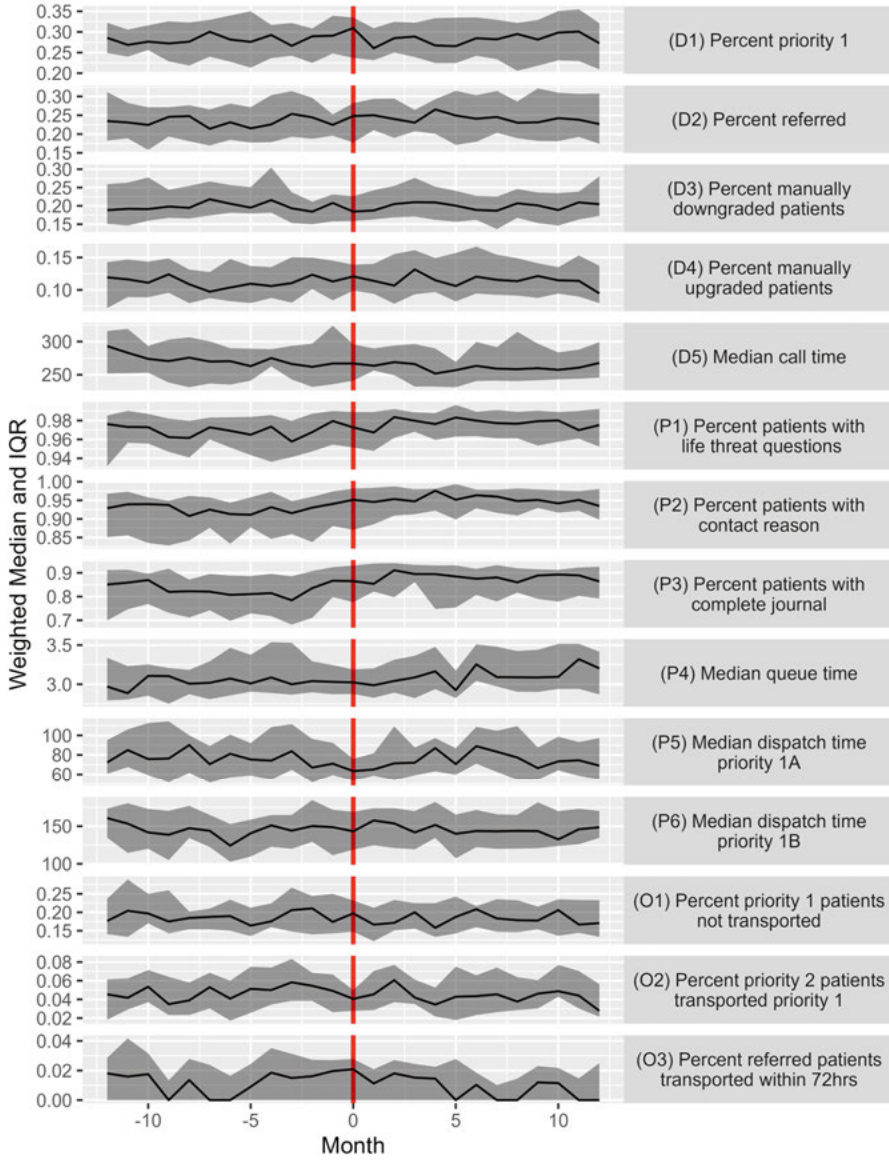

Figure 7 - Median (black line) and interquartile range (grey area) per month before/after intervention (red line) across included nurses for each reported measure

Study III – V: Risk prediction

Studies III – V constitute a series of investigations which progressively increase in rigour, examining the theoretical performance of risk prediction models within the context they are developed (III), in new geographical and organizational settings (IV), and finally the ability to leverage this theoretical performance into enhanced decisional capacity in clinical practice (V).

Study III is an internal validation study of a composite hospital outcome prediction model. In this study, each outcome was weighted equally, though a range of alternate outcome weights were examined in sensitivity analyses. The study used data from 45 045 ambulance missions during 2016-2017 to train models which were then tested on 23 623 missions from the year 2018. Two separate models were trained (one based only on data available at the dispatch center, and one also including data collected from ambulance journals). These models were compared with the performance of the patient's dispatched priorities and retroactively calculated NEWS. The results in the test dataset regarding the discrimination of each combination of predictor and outcome in terms of C-index is reported in table 1 below.

Table 1 - C-indexes in test data for various combinations of predictors and outcomes

Outcome	Dispatched priority	NEWS Score	Dispatch risk score	Ambulance risk score
Hospital admission	0.51 (0.50-0.52)	0.66 (0.65-0.67)	0.73 (0.72-0.73)	0.79 (0.78-0.80)
Critical Care	0.57 (0.55-0.59)	0.75 (0.73-0.77)	0.70 (0.68-0.72)	0.79 (0.77-0.81)
Two-day mortality	0.66 (0.62-0.70)	0.85 (0.81-0.88)	0.79 (0.76-0.82)	0.89 (0.87-0.92)

Dispatch risk scores were found to strictly outperform actual dispatched priorities across all outcomes. Ambulance risk scores were found to similarly outperform NEWS for all outcomes. Comparing dispatch risk scores to NEWS, comparative performance was mixed, with NEWS performing better than the dispatch risk score in predicting critical care and mortality, but dispatch risk scores outperforming NEWS in predicting hospital admission, despite not utilizing patient vital sign data. Model calibration was also assessed, with near ideal overall calibration and only slight deviations from ideal calibration as assed using calibration curves.

These results were extended in study IV, where the loss in performance upon generalizing the ambulance risk score models to new settings was investigated.

Data from a total of 864 878 ambulance missions were collected from across six included regions. The predictor data used to develop and validate these models consisted primarily of vital signs, as opposed to information about interventions performed in study III. Given the aim of evaluating differences in generalization loss across various outcomes and models, rather than apply the outcome compositing approach used in studies III and V, each risk prediction was investigated separately. The main results regarding model discrimination are reported in table 2:

Table 2 - C-indexes for combinations of model, outcome, and internal/external validation method (not applicable to NEWS and RETTS scores)

Outcome	Model	Internal	External
Hospital admission	Gradient Boosting	0.757 (0.756-0.759)	0.733 (0.732-0.734)
	Logistic Regression	0.737 (0.736-0.739)	0.725 (0.723-0.726)
	NEWS	0.664 (0.663-0.666)	
	RETTS	0.679 (0.678-0.681)	
Three-day mortality	Gradient Boosting	0.892 (0.888-0.895)	0.888 (0.884-0.892)
	Logistic Regression	0.883 (0.879-0.887)	0.881 (0.877-0.885)
	NEWS	0.831 (0.826-0.836)	
	RETTS	0.761 (0.754-0.767)	

The main results are derived from comparing the C-index for each model between the internally and externally validated scores, and then with the discrimination of NEWS and RETTS for that outcome. Regarding hospital admission outcomes, gradient boosting generally outperformed logistic regressions. While logistic regression models did suffer less generalization loss, the algorithmic method remained slightly superior even in external datasets. Regarding mortality, no significant reduction in performance was identified for

logistic regression, and only a very slight degradation was detected using gradient boosting.

Calibration was also assessed, with significant deviations from ideal calibration identified in the external validation. Deviations from ideal Observed:Expected ratios were identified upon external validation for all models and outcomes. Systematic miscalibration was particularly evident regarding hospital admission outcomes: While miscalibration regarding 3-day mortality appeared randomly distributed, calibration errors for hospital admission were clearly correlated across the models and varied systematically across the regions, with two regions suffering from underestimated risks, and one from overestimated risks as shown in figure 8:

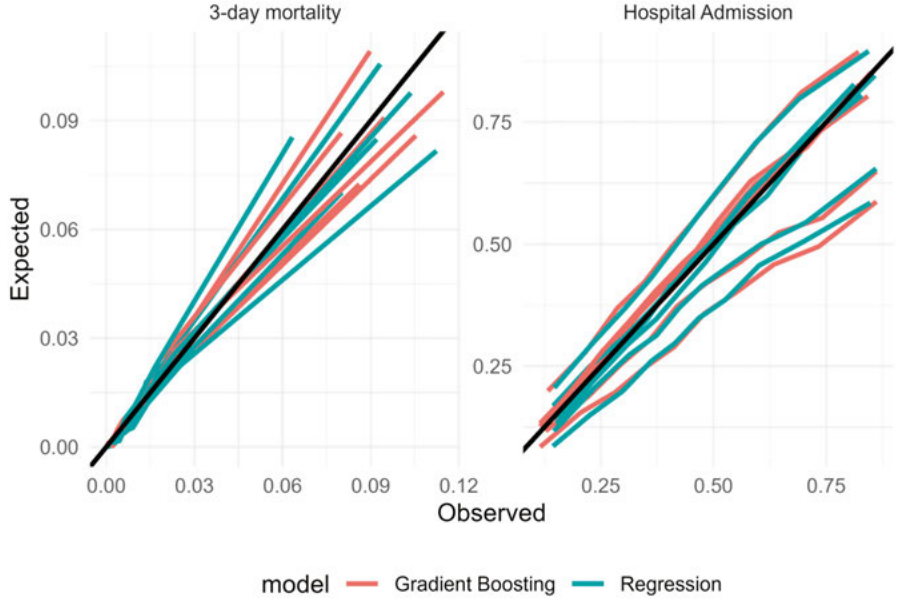

Figure 8 - External calibration curves

In study V, the dispatch model used in study III was further developed, implementing text parsing of free text notes made by dispatchers using a bag-of-words approach including n-grams [139,159]. The composite outcome was also refined to more closely align with ambulance care need as described previously. Over the 34 months between February 2021 and November 2024, 1 845 Resource Constrained Situations (RCS) were identified and included in the study. Of these, 306 contained at least one patient who opted out of the study, leading to post-randomization dropout. An additional 294 cases were identified wherein no ambulance was dispatched prior to the same patients being included in another randomization. In cases where exactly the same patients

were included (for instance due to the dispatcher repeatedly comparing the same patients), this was automatically identified by the randomization software, the patients were assigned to the same trial arm, and only the last randomization was included, thus not impacting randomization. In some cases however, the dispatch situation could change between the time of performing the randomization and dispatching the ambulance (such as a third patient calling in), and a new randomization was performed. In these cases, both randomizations were excluded if they were assigned to different study arms, and if they were assigned to the same arm, only the last randomization was included.

A total of 1 245 RCS were thus included in the final study sample. This was somewhat lower than the 1 500 estimated to be necessary per the power analysis. Of these, 660 were assigned to the intervention arm, and 585 to the Control arm. Patient demographics and call characteristics were similar across both arms. The primary outcome was assessed using a simple logistic regression model with adjustment for between-imputation variance using Rubin's rules. In the control arm, the patient with the highest NEWS value was prioritized in 62.5% of cases, while in the intervention arm the proportion was 68.3%, corresponding to an odds ratio of 1.28 (95% CI 1.00 – 1.63, $p = 0.047$). Thus, despite the smaller than expected sample size, the study was able to identify a statistically significant intervention effect, albeit barely. Secondary outcomes included a hypothesis regarding differentiation based on identifying the patient with the highest composite outcome score, equivalent to the outcome risk prediction score. This analysis identified slightly stronger intervention effects, with an OR of 1.33 (1.04 – 1.71, $p = 0.023$). An alternate estimator based on a Wilcoxon rank-sum test of the difference in NEWS between prioritized and non-prioritized patients however failed to identify a statistically significant difference (mean difference of 0.62 in the control arm vs. 1.08 in the intervention arm, $p = 0.086$).

Prespecified ancillary analysis were also undertaken to investigate various features of the study cohort. An key part of the study protocol was that dispatchers were permitted to deviate from the model recommendation at their discretion. Analysis suggested that dispatchers prioritized the patient suggested by the model in 81% of cases, compared to 54% in the control arm. To evaluate the impact of these deviations, an analysis estimating study effects assuming a 100% compliance rate was performed. This analysis suggested that the intervention effects would have been substantially larger, with an OR of 1.56 (1.21 – 2.00, $p < 0.001$). A function was also included in the risk assessment tool to indicate the certainty with which the model made its determinations, but this was not found to have had an identifiable impact on dispatcher compliance (81.7% in the high confidence group vs 80.3%). In the hypothetical 100% compliance scenario however, there was a substantial difference in prediction accuracy between the groups, with 78.5% correct assessments in the high-confidence group vs 67.4% in the low confidence

group, corresponding to an OR of 1.77 (1.20 – 2.62, p = 0.004). A number of analyses also evaluated changes over time during the nearly 3 year study period, including loss in model predictive performance, changes in compliance with the risk prediction tool, and effects of learning by dispatchers as manifested by increased accuracy in the control arm. No significant effects were however identified. While not a prespecified analysis, effects were also noted in terms of time to dispatch for prioritized patients, with faster dispatch times for prioritized patients in the intervention arm than in the control arm (29.2 vs 32.1 minutes).

Discussion

Result Synthesis

This thesis aims to lay out a roadmap for the evaluation and enhancement of prehospital decisional capacity. This follows the order of the studies in establishing appropriate measures of patient outcome, and then implementing interventions in clinical practice to improve decision-making.

Study I proposes a methodology for collecting quantitative outcome data, even in the absence of automated systems for collecting this data. The study findings suggested that the majority of patient harms caused by dispatcher referral to non-emergency care could be captured by reviewing a relatively small subset of patients; those admitted to a hospital within 72 hours of referral. For the regions in question, this would involve reviewing about 10 cases each per month (about 2% of included non-emergency referrals). While the 72 hour threshold was adopted by the regions in question when extracting hospital outcome data, resources were not committed to performing these case reviews, leaving this potentially fruitful source of organisational learning untapped. Nonetheless, the findings were useful in establishing a baseline level of confidence that the decisions made by nurses at the dispatch center were in parity with those made by nurses on location with patients [4,43].

Study II laid out the first approach to improving measurable outcomes. The intervention had to be implemented across all regions included in SvLc, not all of which could automatically extract outcome data from hospitals. As such, the measures in the study had to be limited to those which could be collected from the dispatching system itself, thus excluding hospital outcomes of the type developed in study I. The study demonstrated that the intervention had the ability to impact some process measures, specifically those regarding documentation completeness. Impacts on the outcome measures we considered most important were however lacking. These results are in line with previous research identifying the ability of feedback provision to impact process-related measures of quality, but difficulties in impacting patient outcomes [124,126]. Redesigning the feedback reports, the intervention itself, or adding additional pedagogical features to the intervention may have the potential to increase the impact on outcomes. For instance, in addition to providing overall statistics, specific cases of potential triage errors could be listed and

mandated for review per the methods used in study I. Such improvements have already been implemented in similar feedback reports delivered to ambulance care providers.

The results of Studies III-V demonstrate that risk prediction models based on vital signs outperform traditional rule-based triage algorithms, even in situations where local training data is not available. However, the systematic deviations from ideal calibration in several regions regarding hospital admission highlight the need to verify the calibration of models prior to their implementation in new settings. The trigger tool methods in study I constitute one approach to gathering this data in a structured manner. The theoretical performance of machine learning models was furthermore found to successfully transfer to the real-world clinical setting of dispatch centres, resulting in more accurate triage decisions in resource constrained situations. This general approach to risk assessment thus appears sound and suitable for broader implementation and trials in additional settings and application.

Ethical Considerations

Studies I, III, and IV were observational studies based on the analysis of retrospectively gathered data, and as such, the primary ethical concern regarding the study participants was the integrity of potentially identifiable health data. While data minimization (e.g. the exclusion of direct patient identifiers, encoding variables such as diagnoses and hospital names, etc.) can limit risks, given the granular nature of the data necessary to build effective machine learning models, this can only be a partial remedy for privacy concerns. Re-identification methods are a constantly evolving threat, and correlation attacks in particular constitute a threat to pseudonymised or even aggregated datasets [160–162]. As such, researchers must take care in handling even pseudonymized datasets containing sensitive and legally protected information regarding the health status of individuals. While it would be optimal for the sake of transparency and reproducibility to publicly release the datasets on which the results of the studies are based, this is not possible due to the above privacy concerns. There are approaches being developed to address these issues to enable greater data sharing in this domain (e.g. the development of synthetic datasets and principled data anonymization techniques), but all approaches entail some trade-off between information loss, computational complexity, and re-identification risk [163]. There are however few risks involved in being transparent regarding the methods used and their implementation. All studies in this thesis with the exception of the first therefore include the full R source code used to generate the results as online supplements.

The interventional studies II and V have some further direct ethical considerations. One potential risk involved in study II was detrimentally impacting nurse behaviour, particularly with regards to the descriptive measures. If for instance a relatively inexperienced nurse saw that they referred fewer patients than average to non-emergency care, they might feel pressured to take more risks to align with their colleagues. Regarding the normative measures of quality, poorly designed incentives could induce nurses to "game" the metrics, for instance by over-triaging patients to improve measures of under-triage. At its heart these issues relate to Goodhart's law: When a measure becomes a target, it ceases to be a good measure [164]. Our primary defence against skewing clinical judgement to improve performance metrics was the presentation of the feedback reports as a tool for self-learning, rather than a steering mechanism. We also sought to develop measures which would be difficult to game by ensuring that measures of over- and under-triage were balanced against each other.

Study V involved the implementation of a novel machine learning-based risk prediction tool in clinical practice. This entails clear patient safety risks which had to be mitigated. The foremost of these mitigations was the selection of the study cohort. Priority 1 patients are liable to be substantially harmed by a delayed ambulance response, given the potentially unstable nature of their condition. Patients referred to non-emergency care are on the other hand subject to substantial risks if time sensitive conditions are missed [4,165]. As such, patients assigned a low-priority ambulance response were chosen to constitute the study population. The evaluation of patients in this low-risk "goldilocks zone" of patients was fundamental to designing a study which minimized patient safety hazards. The obverse of the risk minimization coin is that the potential to improve outcomes for patients is correspondingly small, given the previously discussed tenuous association between ambulance response times and e.g. mortality outside of specific high-acuity conditions. The average of 2 minutes in reduced ambulance response time for prioritized patients identified is thus unlikely to have meaningfully impacted definitive patient outcomes.

Study V also necessitated the use of a novel mechanism to obtain patient consent for participation in the study. It was argued successfully to the Swedish Ethics Review Authority that it was not feasible given time constrains at the dispatch center to obtain prospective verbal, much less written, informed consent for participation in the study. Instead, patients were included in the study based on what we termed "assumed consent", and a full informed consent letter was mailed to the patient's home address at a later time. It was further argued that requiring positive written consent would result in an unacceptably high rate of post-randomization loss to follow-up, especially as any RCS containing a single refusal would have to be excluded. As such, patients were instead given the option to opt-out of participation in the study. This can be

considered a compromise between the stringent position of requiring positive, prospective, written, informed consent prior to inclusion of patients in any interventional study, and the position of e.g. Danish ethics review boards, where similar randomized studies have had the requirement for informed patient consent completely waived [67].

Finally, some broader questions relating to the ethics of using ML techniques to support healthcare decisions should be acknowledged. Who, for instance, is responsible for decisions made by or according to an algorithm? It is already quite difficult to disentangle the responsibility of the care provider from the systems in which they are embedded. An experienced care provider today can plainly see why and when it may be appropriate to deviate from a traditional rule-based CDS system. As risk assessment tools become increasingly complex however, it may become difficult to tell when the model is detecting some subtle pattern invisible to the care provider, and when it is making predictions based on e.g. inappropriate extrapolation or unfairly biased outcome data. These issues become particularly severe as one considers the use of large language models such as Chat-GPT for triage, which are capable of not only being disastrously wrong, but also hallucinating plausible-sounding explanations for their incorrect decisions [166,167]. The use of such models thus risks degrading the overall decisional capacity of care providers if providers cease critically appraising their recommendations.

Limitations

For each of the individual studies, there are a number of technical limitations relating to the specific study design and the analyses employed in their execution. There are also some general limitations to the thesis as a whole. The first is that all data used to generate the results are collected in Sweden, and for the most part in a few specific regions in mid-Sweden. To address this, study IV which specifically investigated generalizability was limited to including only relatively objective predictor data which was thought to be common even outside of the Swedish context. These included patient demographics, vital signs, and a set of broad patient chief complaints grouped into 24 general categories. Predictors such as interventions performed by ambulance crews were found to be highly predictive of hospital outcomes, but were considered to be unduly subject to variation across organizations with different clinical guidelines and routines. Given this limited predictor set, the results are likely to generalize well even outside of Sweden, but this can of course only be established through further external validation in international datasets.

A further limitation is the omission in this thesis of studies investigating other important aspects surrounding the use of risk prediction models in triage. Perhaps the foremost is establishing that the risk prediction models do not unduly favor certain classes of individuals, i.e. that the models are fair. This is a particularly difficult property to establish, given that there exist many competing and often mutually exclusive definitions of what it means for a model to be fair [116,117,168].

The approach used to determine composite outcome scores is furthermore admittedly somewhat *ad hoc*. The process for establishing an appropriate composite outcome for study V involved a mix of quantitative and qualitative assessments. The identification of measures suitable for aggregation in the pre-estimation compositing step for instance involved both a factor analysis to identify correlated measures, and consideration of e.g. potential intervention-induced bias. Post-estimation weights were similarly determined partially by evaluating the correlation of each component measure with NEWS and tempering the findings of these analyses with clinical value judgements. It would be desirable to develop a more rigorous process for empirically determining suitable composite outcome measures for given applications.

Further Research

Addressing the last limitation first, a major challenge in selecting outcome weights is that operationalizing patient acuity in terms of measurable outcomes is a rather abstract exercise. As a care provider, it can be difficult to define the properties of an acute patient in quantifiable terms, but to borrow the US Supreme court's definition of pornography, "you know it when you see it" [169]. Discrete Choice Experiments (DCE) are an approach often used in health economics to elicit the value placed on attributes of a situation when the mental constructs underpinning those value judgements are difficult to determine or articulate [170,171]. The method typically involves a survey in which respondents are presented with a decision scenario with two or more alternatives with varying attributes which the respondent must choose between. In the application of identifying appropriate outcome weights for prehospital risk prediction models, a decision scenario corresponding to the precise use case (e.g., "Which of the following two patients should receive an ambulance first?" for the study V scenario) would be described, along with two patient synopses developed either synthetically or drawn from historical data and anonymized. An appropriately selected set of comparisons based on the number of attributes of interest, expected interaction effects, and feasible sample size would then be generated. Ideally, the survey would be delivered to as broad a set of respondents as possible, including ambulance and dispatch care providers, medical leadership, ED clinicians, etc. By analysing the results

using an appropriate model (e.g. a logistic regression model with random respondent-level effects), attribute coefficients can then be extracted and used to define composite outcome weights. Composite risk scores thus established would have some empirical ground to stand on beyond their correspondence with NEWS as a general indicator of patient risk.

Regarding fairness, there are a number of often countervailing definitions of the concept in the literature. Perhaps the most instinctively appealing is the concept of counterfactual fairness, i.e. the conceptualization of a fair risk score as one which is the same in a) reality and b) a counterfactual reality where only a single, sensitive characteristic is different [123]. Applying this conceptualization however requires that only undesirable sources of bias are removed, retaining effects which operate via "fair" causal mechanisms. Thus, addressing bias will likely involve identifying the causal pathways underpinning the appropriate and inappropriate influences of that characteristic on triage decisions, gather the data necessary to estimate their impacts, and implementing methods to disregard effects laying along only the causal pathways associated with undesirable bias. [122,172] Given the difficulty inherent in estimating causal relationships without the ability to manipulate variables, this may be difficult to achieve in purely observational data: It may be that a DCE as described above could provide an opportunity to explore potential biasing paths in a more experimental manner.

The risk prediction models developed in this thesis had a demonstrable impact on care providers, but the overall level of accuracy of the triage process still leaves much to be desired. The implementation reported on in this thesis thus represents a minimum baseline from which to build. One potential avenue to further improving accuracy is to include additional predictor data in the models to improve their performance, primarily from two sources: Structured patient medical record data, and unstructured audio data from the EMD call. Large amounts of comprehensive patient history data are commonly available in Swedish regional health systems which are likely to be highly predictive of outcomes, especially when combined with data regarding the signs and symptoms experienced by the patient as already captured in the models. At scale, these large datasets are typically tokenized and handled using deep learning approaches [173,174]. Utilizing audio data was investigated at the outset of this doctoral project, but at the time, the only feasible approaches to developing useful models for handling Swedish audio data involved contracting development to private companies. This involves patient some patient privacy concerns, and did not align with our principles relating to the development of open source and freely available tools for patient triage. Even if it were to be pursued, attaining sufficiently robust speech-recognition engines then required substantial investment in manually transcribing material to generate training data. Today, the situation has changed and there exist open-source, pre-trained

models (specifically OpenAI's Whisper model and its derivatives) [175] which place the implementation of such tools within reach of less resourced academic and public sector organizations.

Another avenue to increasing the impact the risk prediction models involves designing better user interfaces for the CDS systems they are integrated into. The design of the interface was generally predicated on the authors own conceptualization, albeit with direct input from clinical staff during pilot testing. The same may also be said regarding the design of the feedback reports, which were only somewhat aligned with best practices in the literature. Iterating the design of these tools based on best practices is likely to enhance trustworthiness, usability, and thus compliance. [176]

While study V demonstrated this approach to be effective in the context of dispatching, the models investigated in study IV based on patient vital signs are also ripe for evaluation in a randomized trial. A similar composite risk score aligned with the need for definitive emergency department-level care should be developed based on these models, and a CDS tool implemented either via API to ambulance patient care reporting systems, or via a free-sanding application such as the interface at https://opentriage.net/ui/vitals. Appropriate decision thresholds based on the level of acceptable risk perceived by ambulance service medical directors could then be established to guide care provider non-conveyance decisions. Finally, a randomization module can be integrated into the system to display the risk assessment in only cases randomized to an intervention arm. Given that the models have been validated as outperforming NEWS even in external settings, such a study should be safe to perform even in settings not directly contributing data to the underlying model.

A final area for further work, while not strictly related to research, is the dissemination of these tools such that they can be applied in clinical practice and further developed by other interested parties. It is to this end that the risk prediction tools and feedback reports are provided as open-source software packages [139,177]. Whereas the implementation of retrospective feedback reporting is comparatively uncomplicated and can be easily implemented by organizations with the appropriate data and technical competencies, there are a number of regulatory barriers to the implementation of risk prediction tools of the type investigated in study V. This tool is arguably a class II medical device per the European Medical Device Regulations (MDR), and as such must undergo the processes involved in obtaining and maintaining a CE mark verifying conformity with these regulations [178]. The forthcoming EU AI Act also places additional requirements on the developers of machine learning based risk prediction tools [179].

Ensuring compliance with these regulations constitutes the primary barrier to the immediate adoption of the risk prediction tool described in this thesis. While obtaining and maintaining a CE mark is quite an onerous and expensive process likely exceeding the capacity of a regional ambulance service, there is a glimmer of hope: Article 5.5 of the MDR provides for the use of self-developed medical devices, given that certain conditions are met including the lack of equivalent CE-marked products, and the development of documentation necessary to assess conformity with the MDR by a competent authority. [179] In Sweden, the relevant competent authority is the Health and Social Care Inspectorate, with which all self-developed medical devices must be registered. [180] It is using this exception that we plan pursue the implementation of the tool evaluated in study V in regular clinical practice. For other regions to use these tools based on this exception a key question must be answered which, to the authors knowledge, has no legal precedent: Does the modification of freely available software meet the criteria for self-development as set out in article 5.5 of the MDR?

Conclusion

This thesis demonstrates that the analysis of patient outcomes is a feasible and useful approach in evaluating decision-making with regards to prehospital triage. Providing continuous feedback reports to care providers on their performance is a low-cost intervention with the potential to impact process measures, but further development is required to impact care outcomes. Machine learning-based clinical risk prediction using composite outcome scores was demonstrated to outperform widely used rule-based risk scores in predicting patient outcomes, even when applied to data from new settings. A clinical decision support tool based on this approach was demonstrated to have a beneficial impact on the ability of dispatchers to differentiate between high- and low-risk patients in resource constrained situations. Work remains to establish empirically grounded composite outcome definitions, appropriate methods to evaluate and improve the fairness of these tools, further enhancing their predictive performance, refining the user interface they are implemented in, and addressing regulatory barriers to dissemination.

Sammanfattning på Svenska

Utvecklingen inom den moderna prehospitala vården har gått mot en ökad diversifiering av de vårdinsatser som kan erbjudas patienterna. Därmed finns idag stora möjligheter att öka effektiviteten genom att tillhandahålla vård nära patienterna och på en vårdnivå som är anpassad efter deras behov. Att fatta patientsäkra beslut om vilka vårdinsatser som ska erbjudas, och var de ska utföras, ställer dock höga krav på vårdgivarnas beslutsförmåga. För att möta dessa krav behövs en vidareutveckling av de metoder som används för att utvärdera och stödja beslut som fattas av de prehospitala vårdgivarna. Syftet med denna avhandling har således varit att främja nyttjandet av mått baserade på patientutfall vid utvärdering och förbättring av prehospitala bedömningsprocesser för patienter med mindre akuta tillstånd.

I delstudie I utvecklades en uppsättning utfallsbaserade mått genom vilka vårdskador som missades av det befintliga avvikelsesystemet kunde identifieras. I delstudie II utvärderades en intervention för kontinuerlig återkoppling av kvalitetsmått till sjuksköterskor vid larmcentraler. Interventionen ledde till viss förbättring avseende processmått rörande fullständig dokumentation, men effekter på bedömningsförmåga uteblev. I delstudie III internvaliderades ett riskbedömningsinstrument baserat på maskininlärningsmetoder vilket påvisades ha bättre prestanda än traditionella regelbaserade instrument. I delstudie IV undersöktes hur väl riskbedömningsmodeller baserade på ambulansdata kunde generaliseras till nya omständigheter genom en extern validering med data från sex regioner. Resultatet visade att modellerna fortfarande presterade bra, men att det fanns brister i dess kalibrering avseende hur riskbedömningar stämde överens med verkliga utfall. Delstudie V utgjordes av en randomiserad klinisk prövning av ett riskbedömningsinstrument avsett att användas vid ambulansbrist. Studien påvisade att användningen av instrumentet förbättrade larmcentralens förmåga att prioritera patienter med större vårdbehov.

Framöver finns möjligheter att vidareutveckla modellernas prestanda, testa dem i nya kliniska kontexter, ta fram empiriskt grundade kompositutfall, samt säkerställa att modellerna bidrar till en rättvis resursfördelning. Den källkod som krävs för att implementera de beskrivna interventionerna tillhandhålls öppet för att främja transparens och möjliggöra spridning och vidareutveckling.

Acknowledgements

One of the first birthday parties I remember involved pipetting and gel electrophoresis at the lab where my dad Rudy worked, and many weekends were spent surrounded by centrifuges, Eppendorf tubes, and RNAse free zones. Having been thusly molded by science in my formative years, I suppose writing a PhD thesis was inevitable. He was also, in a way, responsible for my interest in prehospital care – At least in the sense that his lung cancer diagnosis prompted my dropping out of college and getting a Real Job as an emergency medical technician. So for all that, your tips from the streets of Brooklyn, your skills in circumventing bureaucracy, and your dedication to principle no matter the cost, thanks. I'm sure you're looking up at us and smiling. I thank my mom, Maria for constancy in seeking the truth, your sense for beauty and adventure, and your capacity to do what is best regardless of what other people think. I love you. I thank my awesome friends, Adam and Johan in particular – It's a comforting thing knowing you have friends for life. My lovely Mimi, I thank you for your drive, intelligence, your dedication to the greater good, for putting up with my random absent gazes as my mind drifts off to thesisland, and letting me play with Olivia (sötast i hela världen ❤️🐈🦄😻).

I thank my many excellent supervisors for their guidance developing and executing the studies in this thesis, and for their input in refining its form. David, your drive to build a research group from scratch is inspiring. Ulrika, your ability to juggle a thousand tasks and still grasp the details of a study is amazing. Carl, your insight into machine learning has been immensely useful, and your willingness and ability to jump into a new field is impressive. Sten, you are truly the Don Corleone of clinical research, and this would not have been possible without you. Finally Hans, you may be the most kindhearted, wise, and thoroughly good person I have met. I know that you always have my back, and that gives me enormous peace of mind. I thank everybody in the Health Services Research group for many hours of deep discussion, fika, beers, and board games. The ambulance service attracts individuals who are twisted in just the right way, and I've thus always been blessed with fantastic colleagues from my first days at the Park Slope volunteer ambulance corps, to now at the Uppsala regional ambulance service. I thank you all for rendering my sense of humor completely inappropriate for any other workplace.

References

1. Pozner, C. N., Zane, R., Nelson, S. J. & Levine, M. International EMS Systems: The United States: past, present, and future. *Resuscitation* **60**, 239–244 (2004).
2. Lindström, V., Bohm, K. & Kurland, L. Prehospital care in Sweden. *Notf. Rettungsmedizin* **18**, 107–109 (2015).
3. Jensen, J. L. *et al.* Alternatives to Traditional EMS Dispatch and Transport: A Scoping Review of Reported Outcomes. *Can. J. Emerg. Med.* **17**, 532–550 (2015).
4. Ebben, R. H. A. *et al.* A patient-safety and professional perspective on non-conveyance in ambulance care: a systematic review. *Scand. J. Trauma Resusc. Emerg. Med.* **25**, 71 (2017).
5. Snooks, H. A., Dale, J., Hartley-Sharpe, C. & Halter, M. On-scene alternatives for emergency ambulance crews attending patients who do not need to travel to the accident and emergency department: a review of the literature. *Emerg. Med. J.* **21**, 212–215 (2004).
6. Alpert, A., Morganti, K. G., Margolis, G. S., Wasserman, J. & Kellermann, A. L. Giving EMS Flexibility In Transporting Low-Acuity Patients Could Generate Substantial Medicare Savings. *Health Aff. (Millwood)* **32**, 2142–2148 (2013).
7. Eastwood, K., Morgans, A., Smith, K. & Stoelwinder, J. Secondary triage in prehospital emergency ambulance services: a systematic review. *Emerg. Med. J.* **32**, 486–492 (2015).
8. Benjamin, E. R. *et al.* The Age of Undertriage: Current Trauma Triage Criteria Underestimate The Role of Age and Comorbidities in Early Mortality. *J. Emerg. Med.* **55**, 278–287 (2018).
9. Höglund, E., Schröder, A., Möller, M., Andersson-Hagiwara, M. & Ohlsson-Nevo, E. The ambulance nurse experiences of non-conveying patients. *J. Clin. Nurs.* **28**, 235–244 (2019).
10. Lederman, J., Lindström, V., Elmqvist, C., Löfvenmark, C. & Djärv, T. Non-conveyance in the ambulance service: a population-based cohort study in Stockholm, Sweden. *BMJ Open* **10**, e036659 (2020).
11. Munjal, K. G. *et al.* Patient Perspectives on EMS Alternate Destination Models. *Prehosp. Emerg. Care* **20**, 705–711 (2016).
12. Burgess, A., van Diggele, C., Roberts, C. & Mellis, C. Feedback in the clinical setting. *BMC Med. Educ.* **20**, 460 (2020).
13. Levin, S. *et al.* Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann. Emerg. Med.* **71**, 565-574.e2 (2018).
14. Blomberg, S. N. *et al.* Machine learning as a supportive tool to recognize cardiac arrest in emergency calls. *Resuscitation* **138**, 322–329 (2019).

15. Pirneskoski, J. *et al.* Random forest machine learning method outperforms prehospital National Early Warning Score for predicting one-day mortality: A retrospective study. *Resusc. Plus* **4**, 100046 (2020).
16. Fraess-Phillips, A. J. Can Paramedics Safely Refuse Transport of Non-Urgent Patients? *Prehospital Disaster Med.* **31**, 667–674 (2016).
17. Yeung, T. *et al.* Review article: Outcomes of patients who are not transported following ambulance attendance: A systematic review and meta-analysis. *Emerg. Med. Australas. EMA* **31**, 321–331 (2019).
18. Höglund, E. *et al.* Ambulance quality and outcome measures for general non-conveyed populations (AQUA): A scoping review. *PLOS ONE* **19**, e0306341 (2024).
19. Pollock, A. Historical Perspectives in the Ambulance Service. in *Ambulance Services* 17–28 (Springer, Cham, 2015). doi:10.1007/978-3-319-18642-9_2.
20. A Brief History of Emergency Medical Services in the United States. https://www.emra.org/about-emra/history/ems-history.
21. Lerner, E. B. & Moscati, R. M. The Golden Hour: Scientific Fact or Medical "Urban Legend"? *Acad. Emerg. Med.* **8**, 758–760 (2001).
22. Rogers, F. B., Rittenhouse, K. J. & Gross, B. W. The golden hour in trauma: Dogma or medical folklore? *Injury* **46**, 525–527 (2015).
23. Veser, A., Sieber, F., Groß, S. & Prückner, S. The demographic impact on the demand for emergency medical services in the urban and rural regions of Bavaria, 2012–2032. *Z. Gesundheitswissenschaften* **23**, 181–188 (2015).
24. Lowthian, J. A. *et al.* The challenges of population ageing: accelerating demand for emergency ambulance services by older patients, 1995–2015. *Med. J. Aust.* **194**, (2011).
25. Hagihara, A., Hasegawa, M., Hinohara, Y., Abe, T. & Motoi, M. The aging population and future demand for emergency ambulances in Japan. *Intern. Emerg. Med.* **8**, 431–437 (2013).
26. Huffstetler, A. & Topmiller, M. Primary Care Access in Areas of the United States With Aging Populations. *Am. Fam. Physician* **109**, 308–309 (2024).
27. Jones, C. H. & Dolsten, M. Healthcare on the brink: navigating the challenges of an aging society in the United States. *NPJ Aging* **10**, 22 (2024).
28. Krůtilová, V. Unmet Need For Health Care – A Serious Issue for European Elderly? *Procedia - Soc. Behav. Sci.* **220**, 217–225 (2016).
29. Lehm, K. K., Andersen, M. S. & Riddervold, I. S. Non-urgent Emergency Callers: Characteristics and Prognosis. *Prehosp. Emerg. Care* **21**, 166–173 (2016).
30. National Association of Emergency Medical Technicians. 2023 National Survey on MIH-CP Programs. https://naemt.org/docs/default-source/community-paramedicine/toolkit/mihcp-presentation_may-2023.pdf (2023).
31. Center for Medicare & Medicaid Services. Emergency Triage, Treat, and Transport (ET3) Model | CMS. https://www.cms.gov/priorities/innovation/innovation-models/et3.
32. Hansson, C. Drivers of Emergency Medical Services (EMS) Utilization in the United States in 2019–2020. (Northcentral University, 2022).
33. Couturier, K. *et al.* EMS Non-Transport of Low-Risk COVID-19 Patients. *Prehosp. Emerg. Care* **27**, 310–314 (2023).
34. Satty, T., Ramgopal, S., Elmer, J., Mosesso, V. N. & Martin-Gill, C. EMS responses and non-transports during the COVID-19 pandemic. *Am. J. Emerg. Med.* **42**, 1–8 (2021).
35. NHS England. Safely Reducing Avoidable Conveyance to Hospital. https://www.england.nhs.uk/urgent-emergency-care/improving-ambulance-services/reducing-avoidable-conveyance/.

36. Snooks, H. *et al.* Towards primary care for non-serious 999 callers: results of a controlled study of 'Treat and Refer' protocols for ambulance crews. *Qual. Saf. Health Care* **13**, 435–443 (2004).
37. Marks, P. J., Daniel, T. D., Afolabi, O., Spiers, G. & Nguyen-Van-Tam, J. S. Emergency (999) calls to the ambulance service that do not result in the patient being transported to hospital: an epidemiological study. *Emerg. Med. J.* **19**, 449–452 (2002).
38. NHS England. Ambulance Quality Indicators Data 2024-25. https://www.england.nhs.uk/statistics/statistical-work-areas/ambulance-quality-indicators/ambulance-quality-indicators-data-2024-25/.
39. Mason, S. *et al.* Effectiveness of paramedic practitioners in attending 999 calls from elderly people in the community: cluster randomised controlled trial. *BMJ* **335**, 919 (2007).
40. NHS England. Paramedic prescribing. https://www.england.nhs.uk/ahp/med-project/paramedics/.
41. Williams, B. The National Early Warning Score: from concept to NHS implementation. *Clin. Med.* **22**, 499–505 (2022).
42. Ebben, R. H. A., Castelijns, M., Frenken, J. & Vloet, L. C. M. Characteristics of non-conveyance ambulance runs: A retrospective study in the Netherlands. *World J. Emerg. Med.* **10**, 239 (2019).
43. Magnusson, C., Källenius, C., Knutsson, S., Herlitz, J. & Axelsson, C. Prehospital assessment by a single responder: The Swedish ambulance nurse in a new role: A pilot study. *Int. Emerg. Nurs.* **26**, 32–37 (2015).
44. Wennlund, K. T. *et al.* Emergency medical dispatchers' experiences of managing emergency calls: a qualitative interview study. *BMJ Open* **12**, e059803 (2022).
45. Wennlund, K. T. *Emergency Medical Dispatching: Protocols, Experiences and Priorities*. (Karolinska Institutet (Sweden), 2023).
46. Milbrett, P. & Halm, M. Characteristics and Predictors of Frequent Utilization of Emergency Services. *J. Emerg. Nurs.* **35**, 191–198 (2009).
47. Edwards, M. J., Bassett, G., Sinden, L. & Fothergill, R. T. Frequent callers to the ambulance service: patient profiling and impact of case management on patient utilisation of the ambulance service. *Emerg. Med. J.* **32**, 392–396 (2015).
48. Ko, M., Lee, Y., Chen, C., Chou, P. & Chu, D. Prevalence of and Predictors for Frequent Utilization of Emergency Department. *Medicine (Baltimore)* **94**, (2015).
49. Höglund, E., Andersson-Hagiwara, M., Schröder, A., Möller, M. & Ohlsson-Nevo, E. Characteristics of non-conveyed patients in emergency medical services (EMS): a one-year prospective descriptive and comparative study in a region of Sweden. *BMC Emerg. Med.* **20**, 61 (2020).
50. Platts-Mills, T. F., Leacock, B., Cabañas, J. G., Shofer, F. S. & McLean, S. A. Emergency Medical Services Use by the Elderly: Analysis of a Statewide Database. *Prehosp. Emerg. Care* **14**, 329–333 (2010).
51. Ryynänen, O.-P., Iirola, T., Reitala, J., Pälve, H. & Malmivaara, A. Is advanced life support better than basic life support in prehospital care? A systematic review. *Scand. J. Trauma Resusc. Emerg. Med.* **18**, 62 (2010).
52. Brink, A. *et al.* Predicting mortality in patients with suspected sepsis at the Emergency Department; A retrospective cohort study comparing qSOFA, SIRS and National Early Warning Score. *PLOS ONE* **14**, e0211133 (2019).

53. Kothari, R. U., Pancioli, A., Liu, T., Brott, T. & Broderick, J. Cincinnati Prehospital Stroke Scale: Reproducibility and Validity. *Ann. Emerg. Med.* **33**, 373–378 (1999).
54. Sasser, S. M. *et al.* Guidelines for Field Triage of Injured Patients: Recommendations of the National Expert Panel on Field Triage, 2011. *Morb. Mortal. Wkly. Rep. Recomm. Rep.* **61**, 1–20 (2012).
55. Hardeland, C. *et al.* Comparison of Medical Priority Dispatch (MPD) and Criteria Based Dispatch (CBD) relating to cardiac arrest calls. *Resuscitation* **85**, 612–616 (2014).
56. Feldman, M. J. *et al.* Comparison of the Medical Priority Dispatch System to an Out-of-hospital Patient Acuity Score. *Acad. Emerg. Med.* **13**, 954–960 (2006).
57. Ellensen, E. N., Hunskaar, S., Wisborg, T. & Zakariassen, E. Variations in contact patterns and dispatch guideline adherence between Norwegian emergency medical communication centres - a cross-sectional study. *Scand. J. Trauma Resusc. Emerg. Med.* **22**, 2 (2014).
58. Torlén, K., Kurland, L., Castrén, M., Olanders, K. & Bohm, K. A comparison of two emergency medical dispatch protocols with respect to accuracy. *Scand. J. Trauma Resusc. Emerg. Med.* **25**, 122 (2017).
59. *AMLS: Advanced Medical Life Support*. (Jones & Bartlett Learning, LLC, 2020).
60. Holmström, I. K., Kaminsky, E., Lindberg, Y., Spangler, D. & Winblad, U. Registered Nurses' experiences of using a clinical decision support system for triage of emergency calls: A qualitative interview study. *J. Adv. Nurs.* **76**, 3104–3112 (2020).
61. Widgren, B. R. & Jourak, M. Medical Emergency Triage and Treatment System (METTS): A New Protocol in Primary Triage and Secondary Priority Decision in Emergency Medicine. *J. Emerg. Med.* **40**, 623–628 (2011).
62. Pirneskoski, J., Kuisma, M., Olkkola, K. T. & Nurmi, J. Prehospital National Early Warning Score predicts early mortality. *Acta Anaesthesiol. Scand.* **63**, 676–683 (2019).
63. Silcock, D. J., Corfield, A. R., Gowens, P. A. & Rooney, K. D. Validation of the National Early Warning Score in the prehospital setting. *Resuscitation* **89**, 31–35 (2015).
64. Brangan, E. *et al.* Using the National Early Warning Score (NEWS) outside acute hospital settings: a qualitative study of staff experiences in the West of England. *BMJ Open* **8**, e022528 (2018).
65. Levin, S. *HopScore: An Electronic Outcomes-Based Emergency Triage System - Final Report*. https://digital.ahrq.gov/sites/default/files/docs/citation/r21hs023641-levin-final-report-2018.pdf (2018).
66. Taylor, R. A. *et al.* Impact of Artificial Intelligence–Based Triage Decision Support on Emergency Department Care. *NEJM AI* **2**, AIoa2400296 (2025).
67. Blomberg, S. N. *et al.* Effect of Machine Learning on Dispatcher Recognition of Out-of-Hospital Cardiac Arrest During Calls to Emergency Medical Services: A Randomized Clinical Trial. *JAMA Netw. Open* **4**, e2032320 (2021).
68. Amann, J. *et al.* To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digit. Health* **1**, e0000016 (2022).
69. Kang, D.-Y. *et al.* Artificial intelligence algorithm to predict the need for critical care in prehospital emergency medical services. *Scand. J. Trauma Resusc. Emerg. Med.* **28**, 17 (2020).

70. Kim, D. *et al.* A data-driven artificial intelligence model for remote triage in the prehospital environment. *PLOS ONE* **13**, e0206006 (2018).
71. Tollinton, L., Metcalf, A. M. & Velupillai, S. Enhancing predictions of patient conveyance using emergency call handler free text notes for unconscious and fainting incidents reported to the London Ambulance Service. *Int. J. Med. Inf.* **141**, 104179 (2020).
72. Shirakawa, T. *et al.* Institution-Specific Machine Learning Models for Prehospital Assessment to Predict Hospital Admission: Prediction Model Development Study. *JMIR Med. Inform.* **8**, e20324 (2020).
73. Tamminen, J., Kallonen, A., Hoppu, S. & Kalliomäki, J. Machine learning model predicts short-term mortality among prehospital patients: A prospective development study from Finland. *Resusc. Plus* **5**, 100089 (2021).
74. Ebrahimian, S. *et al.* FDA-regulated AI Algorithms: Trends, Strengths, and Gaps of Validation Studies. *Acad. Radiol.* **29**, 559–566 (2022).
75. Benjamens, S., Dhunnoo, P. & Meskó, B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. *Npj Digit. Med.* **3**, 1–8 (2020).
76. Donabedian, A. The Quality of Care: How Can It Be Assessed? *JAMA* **260**, 1743–1748 (1988).
77. Toplica, L. Region Stockholm lever inte upp till målet för väntetid vid prio 1-larm. *SVT Nyheter* (2023).
78. Campanile, C. NYC ambulance response times for life-threatening emergencies jump as EMTs see staffing crunch. *New York Post* (2025).
79. Duncan, P., Campbell, D. & García, C. A. Ambulance response times worst on record as A&E waits hit new highs. *The Guardian* (2023).
80. Pons, P. T. *et al.* Paramedic Response Time: Does It Affect Patient Survival? *Acad. Emerg. Med.* **12**, 594–600 (2005).
81. Blackwell, T. H., Kline, J. A., Willis, J. J. & Hicks, G. M. Lack of Association Between Prehospital Response Times and Patient Outcomes. *Prehosp. Emerg. Care* **13**, 444–450 (2009).
82. Zayed Khalifah Al-Shaqsi, S. Response time as a sole performance indicator in EMS: Pitfalls and solutions. *Open Access Emerg. Med.* **2**, 1–6 (2010).
83. Pons, P. T. & Markovchick, V. J. Eight minutes or less: does the ambulance response time guideline impact trauma patient outcome?1. *J. Emerg. Med.* **23**, 43–48 (2002).
84. Blanchard, I. E. *et al.* Emergency Medical Services Response Time and Mortality in an Urban Setting. *Prehosp. Emerg. Care* **16**, 142–151 (2012).
85. Pell, J. P., Sirel, J. M., Marsden, A. K., Ford, I. & Cobbe, S. M. Effect of reducing ambulance response times on deaths from out of hospital cardiac arrest: cohort study. *BMJ* **322**, 1385–1388 (2001).
86. Lee, D. W., Moon, H. J. & Heo, N. H. Association between ambulance response time and neurologic outcome in patients with cardiac arrest. *Am. J. Emerg. Med.* **37**, 1999–2003 (2019).
87. Holmén, J. *et al.* Shortening Ambulance Response Time Increases Survival in Out-of-Hospital Cardiac Arrest. *J. Am. Heart Assoc.* **9**, e017048 (2020).
88. Kim, J. *et al.* Does Prehospital Time Influence Clinical Outcomes in Severe Trauma Patients?: A Cross Sectional Study. *Prehosp. Emerg. Care* **21**, 466–475 (2017).
89. Harmsen, A. M. K. *et al.* The influence of prehospital time on trauma patients outcome: A systematic review. *Injury* **46**, 602–609 (2015).
90. Turner, J. *et al.* Identifying potential measures to assess ambulance service performance and quality of care. in *Developing new ways of measuring the*

quality and impact of ambulance service care: the PhOEBE mixed-methods research programme (NIHR Journals Library, 2019).
91. Perkins, G. D. & Cooke, M. W. Variability in cardiac arrest survival: the NHS Ambulance Service Quality Indicators. *Emerg. Med. J.* **29**, 3–5 (2012).
92. Derlet, R. W. & Richards, J. R. Overcrowding in the nation's emergency departments: Complex causes and disturbing effects. *Ann. Emerg. Med.* **35**, 63–68 (2000).
93. Di Somma, S. *et al.* Overcrowding in emergency department: an international issue. *Intern. Emerg. Med.* **10**, 171–175 (2015).
94. Perona, M., Rahman, M. A. & O'Meara, P. Paramedic Judgement, Decision-Making and Cognitive Processing: A Review of the Literature. *Australas. J. Paramed.* **16**, 1–12 (2019).
95. Sheffield, K., O'Meara, P. & Verrinder, G. Decision-making processes when paramedics refer patients away from hospital: a scoping review. *Ir. J. Paramed.* **1**, (2016).
96. Ball, S. J. *et al.* Association between ambulance dispatch priority and patient condition. *Emerg. Med. Australas.* **28**, 716–724 (2016).
97. Khorram-Manesh, A., Montán, K. L., Hedelin, A., Kihlgren, M. & Örtenwall, P. Prehospital triage, discrepancy in priority-setting between emergency medical dispatch centre and ambulance crews. *Eur. J. Trauma Emerg. Surg.* **37**, 73–78 (2010).
98. Koivulahti, O., Tommila, M. & Haavisto, E. The accuracy of preliminary diagnoses made by paramedics – a cross-sectional comparative study. *Scand. J. Trauma Resusc. Emerg. Med.* **28**, 70 (2020).
99. Purabdollah, M., Khorrami Markani, A., Moghaddam Tabrizi, F. & Mokhtari, M. Comparison of the Agreement and Accuracy Between Paramedic and Hospital Diagnosis. *Air Med. J.* **41**, 228–232 (2022).
100. Taylor, J., Yudkowsky, E., LaVictoire, P. & Critch, A. Alignment for advanced machine learning systems. *Ethics Artif. Intell.* 342–382 (2016).
101. Bostrom, N. *Superintelligence: Paths, Dangers, Strategies*. (Oxford University Press, 2014).
102. Aaronson, E., Borczuk, P., Benzer, T., Mort, E. & Temin, E. 72h returns: A trigger tool for diagnostic error. *Am. J. Emerg. Med.* (2017) doi:10.1016/j.ajem.2017.08.019.
103. Kansagara, D. *et al.* Risk Prediction Models for Hospital Readmission: A Systematic Review. *JAMA* **306**, 1688–1698 (2011).
104. Brown, L. H. *et al.* Paramedic Determinations of Medical Necessity: A Meta-Analysis. *Prehosp. Emerg. Care* **13**, 516–527 (2009).
105. Torlén Wennlund, K., Kurland, L., Olanders, K., Castrén, M. & Bohm, K. A registry-based observational study comparing emergency calls assessed by emergency medical dispatchers with and without support by registered nurses. *Scand. J. Trauma Resusc. Emerg. Med.* **30**, 1 (2022).
106. Hernán, M. A., Hsu, J. & Healy, B. A Second Chance to Get Causal Inference Right: A Classification of Data Science Tasks. *CHANCE* **32**, 42–49 (2019).
107. Prosperi, M. *et al.* Causal inference and counterfactual prediction in machine learning for actionable healthcare. *Nat. Mach. Intell.* **2**, 369–375 (2020).
108. Seol, J.-W., Yi, W., Choi, J. & Lee, K. S. Causality patterns and machine learning for the extraction of problem-action relations in discharge summaries. *Int. J. Med. Inf.* **98**, 1–12 (2017).
109. Dami, F. *et al.* Prehospital triage accuracy in a criteria based dispatch centre. *BMC Emerg. Med.* **15**, (2015).

110. Holmström, I. K., Kaminsky, E., Lindberg, Y., Spangler, D. & Winblad, U. Better safe than sorry: Registered nurses' strategies for handling difficult calls to emergency medical dispatch centres – An interview study. *J. Clin. Nurs.* **31**, 2486–2494 (2021).
111. Nagendran, M., Chen, Y. & Gordon, A. C. Real time self-rating of decision certainty by clinicians: a systematic review. *Clin. Med.* **19**, 369–374 (2019).
112. Myrstad, M. *et al.* National Early Warning Score 2 (NEWS2) on admission predicts severe disease and in-hospital mortality from Covid-19 – a prospective cohort study. *Scand. J. Trauma Resusc. Emerg. Med.* **28**, 66 (2020).
113. Vickers, A. J. & Elkin, E. B. Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Med. Decis. Making* **26**, 565–574 (2006).
114. Friedman, J., Hastie, T. & Tibshirani, R. *The Elements of Statistical Learning*. vol. 1 (Springer series in statistics New York, 2001).
115. Webb, G. I., Hyde, R., Cao, H., Nguyen, H. L. & Petitjean, F. Characterizing concept drift. *Data Min. Knowl. Discov.* **30**, 964–994 (2016).
116. Corbett-Davies, S., Pierson, E., Feller, A., Goel, S. & Huq, A. Algorithmic Decision Making and the Cost of Fairness. in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 797–806 (Association for Computing Machinery, New York, NY, USA, 2017). doi:10.1145/3097983.3098095.
117. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann. Intern. Med.* **169**, 866–872 (2018).
118. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. & Weinberger, K. Q. On Fairness and Calibration. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
119. Byrnes, J. P., Miller, D. C. & Schafer, W. D. Gender differences in risk taking: A meta-analysis. *Psychol. Bull.* **125**, 367–383 (1999).
120. Hiller, J., Schatz, K. & Drexler, H. Gender influence on health and risk behavior in primary prevention: a systematic review. *J. Public Health* **25**, 339–349 (2017).
121. Newman, J. D. *et al.* Gender Differences in Calls to 9-1-1 During an Acute Coronary Syndrome. *Am. J. Cardiol.* **111**, 58–62 (2013).
122. Chiappa, S. Path-Specific Counterfactual Fairness. *Proc. AAAI Conf. Artif. Intell.* **33**, 7801–7808 (2019).
123. Kusner, M. J., Loftus, J., Russell, C. & Silva, R. Counterfactual Fairness. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
124. Eaton-Williams, P., Mold, F. & Magnusson, C. Effective clinical feedback provision to ambulance clinicians: a literature review. *J. Paramed. Pract.* **12**, 109–117 (2020).
125. Carey, R. N. *et al.* Behavior Change Techniques and Their Mechanisms of Action: A Synthesis of Links Described in Published Intervention Literature. *Ann. Behav. Med.* **53**, 693–707 (2019).
126. Wilson, C., Janes, G., Lawton, R. & Benn, J. Types and effects of feedback for emergency ambulance staff: a systematic mixed studies review and meta-analysis. *BMJ Qual. Saf.* **32**, 573–588 (2023).
127. Brehaut, J. C. *et al.* Practice Feedback Interventions: 15 Suggestions for Optimizing Effectiveness. *Ann. Intern. Med.* **164**, 435–441 (2016).
128. Brown, B. *et al.* Clinical Performance Feedback Intervention Theory (CP-FIT): a new theory for designing, implementing, and evaluating feedback in

health care based on a systematic review and meta-synthesis of qualitative research. *Implement. Sci.* **14**, 40 (2019).
129. Greenes, R. A. *et al.* Clinical decision support models and frameworks: Seeking to address research issues underlying implementation successes and failures. *J. Biomed. Inform.* **78**, 134–143 (2018).
130. Sutton, R. T. *et al.* An overview of clinical decision support systems: benefits, risks, and strategies for success. *Npj Digit. Med.* **3**, 1–10 (2020).
131. Belle, V. & Papantonis, I. Principles and Practice of Explainable Machine Learning. *Front. Big Data* **4**, (2021).
132. McElreath, R. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. (Chapman and Hall/CRC, Boca Raton, 2016).
133. Hofstadter, D. *Gödel, Escher, Bach: An Eternal Golden Braid*. (Basic Books, New York City, 1979).
134. Lerner, E. B., Weik, T. & Edgerton, E. A. Research in Prehospital Care: Overcoming the Barriers to Success. *Prehosp. Emerg. Care* **20**, 448–453 (2016).
135. Eckhardt, C. M. *et al.* Unsupervised machine learning methods and emerging applications in healthcare. *Knee Surg. Sports Traumatol. Arthrosc.* **31**, 376–381 (2023).
136. Greenacre, M. *et al.* Principal component analysis. *Nat. Rev. Methods Primer* **2**, 1–21 (2022).
137. Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Stat. Sci.* **16**, 199–231 (2001).
138. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (ACM, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
139. Spangler, D. openTriage. https://github.com/dnspangler/openTriage.
140. Snoek, J., Larochelle, H. & Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. in *Advances in Neural Information Processing Systems* vol. 25 (Curran Associates, Inc., 2012).
141. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).
142. Cordoba, G., Schwartz, L., Woloshin, S., Bae, H. & Gøtzsche, P. C. Definition, reporting, and interpretation of composite outcomes in clinical trials: systematic review. *BMJ* **341**, c3920 (2010).
143. Schang, L., Hynninen, Y., Morton, A. & Salo, A. Developing robust composite measures of healthcare quality – Ranking intervals and dominance relations for Scottish Health Boards. *Soc. Sci. Med.* **162**, 59–67 (2016).
144. Newgard, C. D. The Validity of Using Multiple Imputation for Missing Out-of-hospital Data in a State Trauma Registry. *Acad. Emerg. Med.* **13**, 314–324 (2006).
145. Classen, D. C., Lloyd, R. C., Provost, L., Griffin, F. A. & Resar, R. Development and Evaluation of the Institute for Healthcare Improvement Global Trigger Tool. *J. Patient Saf.* **4**, 169–177 (2008).
146. Efron, B. Bootstrap Methods: Another Look at the Jackknife. *Ann. Stat.* **7**, 1–26 (1979).
147. Heale, R. & Twycross, A. Validity and reliability in quantitative studies. *Evid. Based Nurs.* **18**, 66–67 (2015).
148. Bernal, J. L., Cummins, S. & Gasparrini, A. Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int. J. Epidemiol.* **46**, 348–355 (2017).

149. Ting Fok, C. C., Henry, D. & Allen, J. Research Designs for Intervention Research with Small Samples II: Stepped Wedge and Interrupted Time-Series Designs. *Prev. Sci. Off. J. Soc. Prev. Res.* **16**, 967–977 (2015).
150. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (Cambridge university press, 2006).
151. Harrell, F. E. *Regression Modeling Strategies*. (Springer International Publishing, 2015).
152. Marcum, J. A statistical theory of target detection by pulsed radar. *IRE Trans. Inf. Theory* **6**, 59–267 (1960).
153. Coster, J. *et al.* Outcomes for Patients Who Contact the Emergency Ambulance Service and Are Not Transported to the Emergency Department: A Data Linkage Study. *Prehosp. Emerg. Care* **23**, 566–577 (2019).
154. Rubin, D. B. Inference and Missing Data. *ETS Res. Bull. Ser.* **1975**, i–19 (1975).
155. Buuren, S. van & Groothuis-Oudshoorn, K. Multivariate Imputation by Chained Equations in R. *J. Stat. Softw.* **45**, (2011).
156. Azur, M. J., Stuart, E. A., Frangakis, C. & Leaf, P. J. Multiple imputation by chained equations: what is it and how does it work? *Int. J. Methods Psychiatr. Res.* **20**, 40–49 (2011).
157. Marshall, A., Altman, D. G., Holder, R. L. & Royston, P. Combining estimates of interest in prognostic modelling studies after multiple imputation: current practice and guidelines. *BMC Med. Res. Methodol.* **9**, 57 (2009).
158. Van Calster, B. *et al.* Calibration: the Achilles heel of predictive analytics. *BMC Med.* **17**, 230 (2019).
159. Schonlau, M., Guenther, N. & Sucholutsky, I. Text Mining with n-gram Variables. *Stata J.* **17**, 866–881 (2017).
160. Podoliaka, O., Mushkatblat, V. & Kaplan, A. Privacy Attacks Based on Correlation of Dataset Identifiers: Assessing the Risk. in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* 0808–0815 (2022). doi:10.1109/CCWC54503.2022.9720885.
161. Dwork, C., Smith, A., Steinke, T. & Ullman, J. Exposed! A Survey of Attacks on Private Data. *Annu. Rev. Stat. Its Appl.* **4**, 61–84 (2017).
162. Emam, K. E., Jonker, E., Arbuckle, L. & Malin, B. A Systematic Review of Re-Identification Attacks on Health Data. *PLOS ONE* **6**, e28071 (2011).
163. Eze, B. & Peyton, L. Systematic Literature Review on the Anonymization of High Dimensional Streaming Datasets for Health Data Sharing. *Procedia Comput. Sci.* **63**, 348–355 (2015).
164. Goodhart, C. Problems of monetary management : the U.K. experience. *Pap. Monet. Econ. 1975 1* **1**, (1975).
165. Dale, J. *et al.* Safety of telephone consultation for "non-serious" emergency ambulance service patients. *Qual. Saf. Health Care* **13**, 363–373 (2004).
166. Zaboli, A. *et al.* Chat-GPT in triage: Still far from surpassing human expertise - An observational study. *Am. J. Emerg. Med.* **92**, 165–171 (2025).
167. Kim, Y. *et al.* Medical Hallucinations in Foundation Models and Their Impact on Healthcare. Preprint at https://doi.org/10.48550/arXiv.2503.05777 (2025).
168. Balayn, A., Lofi, C. & Houben, G.-J. Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. *VLDB J.* **30**, 739–768 (2021).
169. Gewirtz, P. On 'I Know It When I See It'. *Yale Law J.* **105**, 1023–1047 (1996).

170. Ryan, M., Bate, A., Eastmond, C. J. & Ludbrook, A. Use of discrete choice experiments to elicit preferences. *BMJ Qual. Saf.* **10**, i55–i60 (2001).
171. Amaya-Amaya, M., Gerard, K. & Ryan, M. Discrete Choice Experiments in a Nutshell. in *Using Discrete Choice Experiments to Value Health and Health Care* vol. 11 13–46 (Springer Netherlands, Dordrecht, 2008).
172. Zhang, L., Wu, Y. & Wu, X. A Causal Framework for Discovering and Removing Direct and Indirect Discrimination. *Proc. Twenty-Sixth Int. Jt. Conf. Artif. Intell.* 3929–3935 (2017).
173. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *Npj Digit. Med.* **1**, 18 (2018).
174. Ayala Solares, J. R. *et al.* Deep learning for electronic health records: A comparative review of multiple deep neural architectures. *J. Biomed. Inform.* **101**, 103337 (2020).
175. Radford, A. *et al.* Robust Speech Recognition via Large-Scale Weak Supervision. Preprint at https://doi.org/10.48550/arXiv.2212.04356 (2022).
176. Gulliksen, J. *et al.* Key principles for user-centred systems design. *Behav. Inf. Technol.* **22**, 397–409 (2003).
177. Spangler, D. feedback_reports. https://github.com/dnspangler/feedback_reports (2025).
178. Swedish Medical Products Agency. Risk classification. https://www.lakemedelsverket.se/en/medical-devices/manufacture/risk-classification.
179. *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance)*. (2024).
180. Health and Social Care Inspectorate. Medicintekniska produkter. https://www.ivo.se/vard-omsorgsgivare/anmal-handelse-lamna-underrattelse/medicintekniska-produkter/.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine 2149

Editor: The Dean of the Faculty of Medicine

A doctoral dissertation from the Faculty of Medicine, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Medicine".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-553889

ACTA UNIVERSITATIS
UPSALIENSIS
2025