Comorbidity Extraction for In-Hospital Mortality Analysis: a Comparison of Regular Expressions and Large Language Models

Ayca Begum Tascioglu
aycabegum.tascioglu2@unibo.it
Dept. of Computer Science and Engineering
University of Bologna
Bologna, Italy

Flavio Bertini
flavio.bertini@unipr.it
Dept. of Mathematical, Physical and Computer Sciences
University of Parma
Parma, Italy

Laura Pistore
laura.pistore@unibo.it
Interdepartmental Centre for Industrial ICT Research
University of Bologna
Cesena, Italy

Andrea Fabbri
andrea.fabbri@auslromagna.it
Emergency Department, Local Health Agency of Romagna
Morgagni-Pierantoni Hospital
Forlì, Italy

Danilo Montesi
danilo.montesi@unibo.it
Dept. of Computer Science and Engineering
University of Bologna
Bologna, Italy

Abstract

Limited hospital resources may prolong patient stays in the Emergency Department (ED), potentially affecting clinical outcomes. This paper investigates the link between overnight Emergency Department (ED) stays and in-hospital mortality, focusing on comorbidity extraction from clinical records. In Italian healthcare records, comorbidities are typically documented using abbreviations and non-standard clinical slang in unstructured free-text fields. We evaluated two approaches for comorbidity extraction: a rule-based method and a Large Language Model approach. Both were assessed against a dataset of 200 clinical records manually annotated by emergency medical staff. This first result showed that the rule-based strategy outperformed Large Language Models in terms of recall, F1-score, consistency, and reliability. Then, to assess the impact of overnight stays on in-hospital mortality and to identify the most significant predictors, 126,696 ED admissions at the Romagna Local Health Agency in Forlì, Italy, between 2017 and 2022 were analysed using several models, with particular emphasis on interpretability. Comorbidity burden, diagnosis severity, age, and infectious, respiratory, and circulatory diseases emerged as the most influential factors.

CCS Concepts

• **Information systems** → *Data analytics*; • **Computing methodologies** → **Information extraction**; • **Applied computing** → **Health informatics**.

Keywords

Real-world healthcare data, Multi-label mining, Clinical expert and artificial benchmarking, Medical prompt engineering

BCB '25, Philadelphia, PA, USA

ACM ISBN 979-8-4007-2200-4/2025/10
https://doi.org/10.1145/3765612.3767202

ACM Reference Format:
Ayca Begum Tascioglu, Flavio Bertini, Laura Pistore, Andrea Fabbri, and Danilo Montesi. 2025. Comorbidity Extraction for In-Hospital Mortality Analysis: a Comparison of Regular Expressions and Large Language Models. In *Proceedings of the 16th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '25), October 11–15, 2025, Philadelphia, PA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3765612.3767202

1 Introduction

Efficient management of resources within the ED plays a pivotal role in delivering timely and effective care, especially for patients requiring urgent admission to hospital wards [1, 24]. The strain imposed by overcrowded environments, prolonged ED stays and constrained resource availability intensifies operational burdens on both healthcare providers and patients [19, 27], affecting care quality and ED performance, with potential implications for elevated in-hospital mortality rates. These issues often worsen during overnight hours, when staff and resources are typically reduced, leading to longer wait times and possible disruptions in care continuity. The problem is exacerbated for older adults, who often have complex profiles, higher risk, and lower resilience.

The relationship between overnight stays in the ED and in-hospital mortality has recently attracted growing attention, yet remains a subject of ongoing debate. A prospective cohort study carried out over a 3-day period across 97 EDs in France assessed 1,598 patients aged 75 years and older to explore the potential effect of remaining overnight in the ED on subsequent in-hospital mortality [20]. In contrast, a retrospective study conducted in Spain across 52 EDs analysed a 7-day cohort of 3,243 elderly patients and found no statistically significant association between overnight ED stays and increased in-hospital mortality [16]. While both investigations benefit from broad institutional participation, they are limited by their short observational windows, which restrict their capacity to capture longer-term trends or variability in patient characteristics. Moreover, neither study accounts for the influence of external variables, such as seasonal fluctuations or extraordinary events like the COVID-19 pandemic, which could impact mortality outcomes.

This work investigates the potential association between overnight stays in the ED and in-hospital mortality by leveraging a real-world dataset routinely collected over a six-year period by the Romagna Local Health Agency in Forlì, Italy. The dataset comprises 126,696 ED admissions recorded between 2017 and 2022, drawn from three distinct sources providing complementary information, including vital signs, diagnostic data, and discharge summaries. Following data cleaning and harmonisation procedures, a total of 20,009 patient records were retained for the final evaluation. A broad set of clinical and contextual variables was employed to characterise the care pathway and identify factors influencing in-hospital mortality. In particular, to assess the role of overnight ED stays, the analysis incorporated the National Early Warning Score (NEWS), Charlson Comorbidity Index (CCI), diagnostic categories, length of stay, patient age groups, trauma-related visits, and temporal variables such as year, month, seasonality, and admission during the COVID-19 period.

The NEWS score is a clinical tool designed to detect early signs of patient deterioration. It is calculated at ED admission based on seven vital signs. The CCI, by contrast, assesses the impact of comorbidities on patient outcomes and estimates long-term survival, and has been recognised in the literature as a key proxy for rapid patient screening. Further details on the NEWS and CCI are provided in Section 2. Unfortunately, in Italian clinical records, comorbidities are documented in unstructured free-text form, which does not include The International Classification of Diseases (ICD) codes or standardised comorbidity checklists. Notably, the CCI was identified as one of the strongest predictors of in-hospital mortality among elderly patients, highlighting the need for reliable comorbidity identification. Thus, two extraction strategies were developed and evaluated: one based on regular expressions, and another leveraging Large Language Model (LLM)-based tools (*i.e.*, ChatGPT 4o and LLaMA 3 8B). Both methods were benchmarked using standard evaluation metrics – *i.e.*, precision, recall, and F1-score – to determine the most effective approach for extracting comorbid conditions. We will show that the rule-based strategy outperformed the LLM-based tools in terms of recall, F1-score, consistency, and reliability. This initial result is crucial, as the subsequent analysis relies heavily on it: the CCI will in fact be one of the key variables considered in the study of overnight stays in the ED and in-hospital mortality.

To investigate the association between overnight stays in the ED and in-hospital mortality, the data analysis pipeline was implemented on a real-world dataset encompassing over 126,696 ED admissions, providing a practical foundation for informing resource allocation decisions by healthcare professionals. In particular, to explore the influence of various clinical variables and to enhance interpretability, multiple models were applied to identify factors associated with in-hospital mortality. Among the models tested, Logistic Regression demonstrated the best overall performance, while also offering the added benefit of improved interpretability. The analysis revealed that overnight stays did not significantly contribute to mortality prediction, thereby shifting attention towards more impactful variables.

Our analysis shows that overnight stays in the ED are not a significant predictor of in-hospital mortality. In contrast, stronger associations were observed with variables such as the CCI, NEWS, patient age, and the presence of infectious, respiratory, and circulatory conditions. These results highlight the importance of integrating validated clinical scores, even when dealing with historical datasets, as they provide valuable insight for retrospective analyses. Within this framework, both CCI and NEWS were central to our study. The main contributions of this paper can be summarised as follows:

- **Utilisation of a real-world dataset -** This study relies on routinely collected clinical data from a Local Health Agency in Italy, reflecting real-world healthcare practices rather than data from synthetic or controlled settings.
- **Longitudinal analysis over a six-year period -** The dataset spans from 2017 to 2022, enabling the investigation of time-related variables, including seasonal trends and rare events (*e.g.*, the COVID-19 pandemic).
- **Context-specific characteristics -** ED datasets vary significantly across healthcare systems, particularly in terms of clinical documentation and workflow organisation. Patient data often lacks standardised coding, and eHealth applications must address several related challenges.
- **Zero-shot learning for LLMs -** We evaluate whether state-of-the-art LLMs (*i.e.*, ChatGPT 4o and LLaMA 3 8B), in a zero-shot learning setting, were able to address a highly relevant task in clinical information extraction. Given the considerable attention these models have attracted, it seemed reasonable to expect them to perform strongly even in this domain-specific task. However, the finding that a simple rule-based approach outperformed two of the leading LLMs was not obvious *a priori*.

The structure of the paper is as follows. Section 2 provides an overview of the most relevant literature related to this work. Section 3 describes the dataset in detail along with the preprocessing steps undertaken. Section 4 details the analytical approaches and modelling techniques adopted in the study. Section 5 presents and interprets the main findings, while also discussing the strengths and limitations of the proposed approach. Lastly, Section 6 concludes the paper by summarising the key outcomes and outlining directions for future investigation.

2 Related Work

Several studies have suggested a potential relationship between prolonged ED boarding and increased in-hospital mortality. Valli et al. observed that longer hospital stays and extended waiting times were correlated with ED overcrowding, which itself was linked to elevated mortality rates during periods of high patient volume [25]. A separate research strand has focused on the specific role of overnight ED stays in influencing mortality outcomes [16, 20]. Nonetheless, these investigations have not reached consistent conclusions regarding the impact of overnight admissions and did not comprehensively account for additional clinical variables or external influences.

In order to perform a more in-depth evaluation and to better understand the clinical variables contributing to the relationship between overnight ED stays and in-hospital mortality, both the NEWS and CCI metrics were incorporated into the analysis. The NEWS score, originally developed by the Royal College of Physicians in

Table 1: Scoring criteria for the vital signs included in the NEWS system.

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

the UK, has since been implemented by multiple healthcare systems worldwide [14]. It serves as an early warning system to identify patients at risk of clinical deterioration. The score is determined at the time of ED admission and is based on seven physiological parameters [11]: body temperature, systolic blood pressure, respiratory rate, oxygen saturation, level of consciousness (evaluated through the AVPU scale: Alert, Voice, Pain, Unresponsive), administration of supplemental oxygen (binary variable), and heart rate. As detailed in Table 1, a composite score ranging from 0 to 20 is derived from these measurements. The resulting total stratifies patients into clinical risk categories: scores from 1 to 4 represent low risk; scores from 5 to 6, or a score of 3 in any single parameter, indicate medium risk, warranting clinical reassessment; and scores above 7 reflect high risk, requiring urgent medical intervention.

The CCI is a widely used clinical index designed to quantify the burden of comorbid conditions and their influence on patient outcomes [4]. Originally developed as a prognostic tool for estimating mortality risk among hospitalised individuals, it was subsequently extended to evaluate long-term survival prospects. The index assigns weighted scores to a set of nineteen comorbidities, with each condition receiving a value between 1 and 6. The cumulative score ranges from 0 to 37 and also incorporates an adjustment based on patient age. The scoring of individual comorbidities follows a predefined set of criteria, as detailed below:

- a score of 1 is assigned for the presence of the following conditions: myocardial infarction, congestive heart failure, peripheral vascular disease, cerebrovascular disease, dementia, chronic pulmonary disease, connective tissue disorder, peptic ulcer disease, mild liver disease, and uncomplicated diabetes;
- a score of 2 is assigned for the presence of the following conditions: hemiplegia, moderate or severe renal impairment, diabetes with end-organ complications, solid tumours (non-metastatic), leukaemia, and lymphoma;
- a score of 3 is assigned for moderate or severe hepatic disease;
- the highest score, 6 points, is reserved for metastatic solid tumours and AIDS;
- in addition, for age adjustment, 1 point is added for every decade of life beyond the age of 40.

De Groot et al. carried out a systematic assessment of the validity and reliability of the CCI, relying on a standardised checklist to extract comorbid conditions [7]. A large-scale real-world study involving 50,668 elderly individuals diagnosed with acute myeloid leukaemia reported that elevated CCI scores were independently associated with poorer one-month survival and overall mortality outcomes [8]. Similarly, Jørgensen et al. examined the applicability of the CCI in geriatric oncology, focusing on cancer patients who typically present with high levels of polypharmacy and are frequently underrepresented in clinical research [13]. Within the emergency setting, Murray et al. analysed the predictive value of the CCI for one-year mortality in patients presenting to the ED with suspected infections. Their findings demonstrated that higher CCI values were significantly linked to increased mortality risk, even after controlling for confounders such as age, gender, and 28-day mortality [17]. The index has also been incorporated into a novel two-way clustering methodology aimed at stratifying patient profiles and reducing data dimensionality [9], thereby enhancing predictive performance for outcomes such as hospital length of stay in individuals with cardiovascular conditions. Although many studies have focused on computing the CCI through diagnostic codes from the International Classification of Diseases (ICD-9 and ICD-10) [3], extracting comorbidity information from unstructured clinical narratives remains a considerable challenge—particularly in non-English texts [18]. Such difficulties stem from language-specific expressions, specialised medical vocabulary, and the widespread use of abbreviations by ED clinicians.

In the dataset used for this study, patient documentation was recorded in Italian, which restricted the availability of external tools suitable for CCI extraction. Additionally, standardised resources such as the ICD codes or structured checklists specifying individual comorbidities were not accessible. A number of existing NLP systems, such as CLAMP [23], cTAKES [22], and MetaMap [5], are commonly employed to extract medical concepts from clinical narratives. Nevertheless, these systems do not provide direct mappings to comorbidity indices and encounter significant limitations when applied to texts in languages other than English, unless preprocessed through translation pipelines [6]. Furthermore, the use of clinical abbreviations and terminology tends to vary not only across countries but also among healthcare institutions, with particularly high variability observed in ED settings where time-sensitive decisions are critical. Consequently, tools developed in other linguistic or institutional contexts may not generalise well. This underscores the necessity for models specifically adapted to the linguistic and operational characteristics of the Italian ED environment.

Recent advancements in artificial intelligence have significantly influenced the healthcare domain, particularly in areas such as clinical documentation, Electronic Health Record (EHR) processing, and the interpretation of medical imaging [15]. While a growing body

of literature has demonstrated the capability of LLMs to process and extract information from clinical narratives [21, 26], their application to comorbidity extraction from unstructured Italian medical texts remains insufficiently investigated. Prior research has also examined the use of ChatGPT (version 3.5) for answering medical multiple-choice questions [2]. Although the model achieved a satisfactory accuracy rate, inconsistencies were observed in repeated responses to identical prompts. Such variability raises concerns regarding its reliability, especially when deployed in high-stakes clinical scenarios where consistency is critical. In light of the promising developments in generative AI technologies [12], this study investigates the applicability of LLMs for the identification of comorbidities, benchmarking its performance against a conventional rule-based method.

3 The Dataset

In this section, we present a comprehensive overview of the dataset, outlining its structure, origin, and the methodology used for data linkage. This is followed by a description of the preprocessing procedures implemented to prepare the dataset for subsequent analytical tasks. In particular, the process of consolidating the patient cohort used in our analysis is shown in Figure 1.

The dataset employed in this study encompasses patient-level information collected between January 1, 2017, and December 31, 2022, obtained from the official registry of the Romagna Local Health Agency located in Forlì, Italy[1]. The data originate from three separate repositories: diagnostic codes and vital sign measurements retrieved from the ED information system, and mortality records sourced from the inpatient ward system. As these sources are independently maintained, a data linkage procedure was implemented based on patient ID, visit ID, and a temporal proximity rule. In particular, diagnostic codes and vital sign measurements records were matched using both patient and visit identifiers, and repeated arrivals at the ED on different days were treated as distinct cases. Mortality records from the inpatient ward system were then associated with ED visits when a matching patient ID appeared within a one-day interval. This final matching step was required due to the lack of a unified identifier for patients transitioning from the ED to the ward, a limitation inherent in the hospital's information system architecture [10]. From an initial dataset of 126,696 ED visits (*i.e.*, the entry point in Figure 1), a total of 44,905 records were successfully linked.

Following completion of the data linkage procedure, a series of exclusion criteria was applied to define the final analytical cohort. Specifically, 24,896 records were excluded due to missing clinical report data, patient age below 75 years, admission to the ward between 00:00 and 08:00, or absence of ED outcome documentation for transfers to external facilities. We also computed the NEWS score for all remaining 20,009 patients. In particular, the vital signs recorded at ED admission and available in our dataset for this calculation include systolic blood pressure, heart rate, respiratory rate, and temperature, while the other variables listed in Table 1 could not be derived from the available information. However, in the clinicians' opinion, this does not constitute a limitation of the study conducted.

[1]For further information on the dataset, please contact co-author Dr Fabbri.

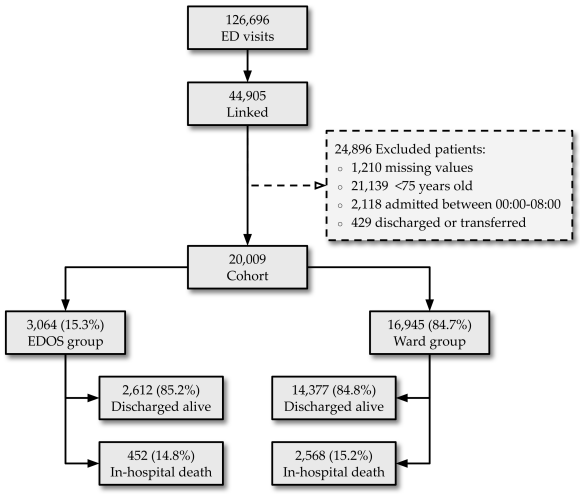

Figure 1: Patient cohort construction.

The analysis concerning overnight stays was carried out on a curated cohort comprising 20,009 patients, stratified into two distinct groups: the Emergency Department Overnight Stay (EDOS) group and the Ward group. The EDOS group consists of patients who were admitted to the ED before 00:00 and remained beyond midnight, while the Ward group includes individuals who presented after 08:00 AM and were discharged before the end of the same calendar day. Initial findings showed that, within the EDOS group, 14.8% of patients died, whereas the mortality rate in the Ward group was higher by 2.7%.

4 Framework Design for Overnight Analysis

This section introduces the analytical framework developed for this study, describing both the methodology used for extracting comorbidities from unstructured Italian clinical text and the structure of the overnight stays analysis performed on a cohort spanning six years.

4.1 Mining Comorbidities from Free Text

Numerous contributions in the literature highlight the relevance of the CCI in evaluating patient health status. Nevertheless, in the Italian healthcare context, the prevalent use of unstructured free-text clinical documentation presents significant barriers to the efficient and accurate computation of the CCI. This issue is further compounded in the ED, where clinical reports are typically compiled manually and lack standardised templates or checklists, thereby increasing the likelihood of inconsistencies and the use of non-standard abbreviations. To address these challenges, we implemented and compared two approaches to comorbidity extraction: a traditional rule-based method using regular expressions, and a more advanced solution based on Large Language Models, specifically ChatGPT 4o and LLaMA 3 8B.

The rule-based method combines Negation Expression (NegEx) logic with a set of customised Regular Expression (RegEx) patterns[2]. The initial phase of this approach involves the detection of negated

[2]The rule set is available at https://github.com/aeyc/ComorbidityExtractionRegEx.

statements, ensuring that subsequent pattern recognition through RegEx operates on semantically relevant content. By isolating and excluding negated expressions early in the pipeline (*e.g.*, "the patient has no diabetes"), the system reduces false positives and increases the precision of the extracted data. The regular expression patterns were developed using a curated collection of domain-specific terminology and abbreviations commonly employed by ED clinicians. The sequential application of negation handling and tailored pattern matching enhances both the accuracy and the reliability of the extracted comorbidity information.

For comparison purposes, comorbidities were also extracted using two well-known LLMs, namely ChatGPT 4o and LLaMA 3 8B, accessed via an API. ChatGPT 4o is a fourth-generation multimodal language model developed by OpenAI. It was selected for its demonstrated capacity to interpret and process clinical free text, despite inherent linguistic and structural challenges. LLaMA 3 8B is an open-source model developed by Meta, offering notable performance compared to earlier models in the same family. The rationale for selecting both was to evaluate differences between a high-end proprietary model (*i.e.*, ChatGPT 4o) and a cost-effective open-source alternative (*i.e.*, LLaMA 3 8B). The aim of this approach was to retrieve relevant pre-existing comorbid conditions from the unstructured narrative of clinical reports, in order to compute the CCI. A key obstacle in this task stemmed from the presence of non-standard language elements, such as inconsistent punctuation, grammatical and syntactic irregularities, as well as domain-specific abbreviations. Furthermore, the formulation of the prompt used to guide the model's behaviour proved critical in obtaining meaningful and reliable outputs. We deliberately chose to test state-of-the-art LLMs in a zero-shot learning setting. For this reason, we opted not to focus on prompt engineering or to implement other hallucination mitigation strategies such as Retrieval-Augmented Generation, fine-tuning, or model alignment. The prompt initially submitted to both the LLMs is reported below:

(1) `We are analysing texts written by doctors upon patient arrival at the emergency department. From these texts, we need to extract comorbidities for the subsequent calculation of the Charlson Comorbidity Index. Extract the comorbidities found in the text.`

Nonetheless, the models' initial output included not only explicitly documented comorbidities but also conditions inferred from clinical symptoms described in the text. Furthermore, a manual review by clinicians revealed that the models often generated comorbidities aligned with those commonly reported in the literature for specific age groups and related conditions, despite no actual correspondence in the individual patient under analysis. To reduce this overgeneralisation and improve the specificity of the extracted information, a sequence of increasingly restrictive instructions was applied. This refinement process culminated in the final version of the prompt, reported below:

(2) `Return the list of pre-existing comorbidities for the subsequent calculation of the CCI based on a text written by a doctor at the time of the patient's arrival in the emergency department. Return a list of ALL comorbidities with '1' if the patient has the condition or '0' otherwise. The list of comorbidities to be returned is: Myocardial infarction, CHF, Peripheral vascular disease, CVA or TIA, Dementia, COPD, Connective tissue disease, Peptic ulcer disease, Liver Disease, Diabetes mellitus, Hemiplegia, Moderate to severe CKD, Solid tumour, Leukaemia, Lymphoma, AIDS. Return the results with ONLY the comorbidities in the list, written exactly as specified. Respond with ONLY the result in JSON format.`

In order to enhance extraction accuracy and minimise the incidence of false positives, a preparatory step was incorporated to rephrase the clinical narratives. This step aimed to isolate information concerning pre-existing comorbid conditions and ongoing pharmacological treatments, while excluding any symptoms associated with the current ED visit. To achieve this, the following prompt was designed:

(3) `A doctor writes a paragraph about a patient in the ER triage. The text is needed to calculate Charlson's comorbidity index. Can you extract only the preexisting comorbidities and the list of medications taken by the patient (if present), ignoring symptoms related to the emergency visit? Return the modified text without introductions.`

Moreover, as part of the text pre-processing workflow, conversational memory was enabled to promote output consistency across interactions with the model. Once the textual content had been refined, it was subjected to a second processing stage using the comorbidity extraction prompt detailed in Prompt 2.

The performance of both comorbidity extraction approaches was assessed using standard evaluation metrics, namely micro-averaged precision, recall, and F1-score, applied to a reference set of 200 clinical records manually annotated by ED physicians. Following this evaluation, the approach yielding the highest performance was selected and employed to process the full set of 20,009 patient narratives used in the overnight stays analysis.

4.2 Overnight Stays Analysis

In order to examine the influence of overnight stays on in-hospital mortality and to evaluate the role of multiple clinical factors, the study population of 20,009 patients was stratified into two cohorts: the EDOS group and the Ward group. These cohorts were compared across various parameters, including demographic characteristics (*i.e.*, age and sex), comorbidity burden, as measured by the CCI and NEWS scores, length of hospital stay, diagnosis codes according to ICD-9[3], trauma-related presentations to the ED, and temporal variables (*i.e.*, year, month, season, and the COVID-19 period). To streamline the analytical process, selected variables were further categorised as follows:

- length of stay was categorised into four intervals: ≤6 hours, >6 to ≤12 hours, >12 to ≤24 hours, and exceeding 24 hours;
- age was categorised into two groups: ≤85 years and >85 years;
- the CCI score was stratified as ≤4 versus >4;

[3]At the time of writing, the Italian National Health System still uses the 9^{th} revision.

Table 2: Distribution of comorbidities across the 200 ground-truth clinical records and detailed F1-scores.

Comorbidity	#	%	Rule-based	ChatGPT *no pre-proc.*	Llama *no pre-proc.*	ChatGPT *with pre-proc.*	Llama *with pre-proc.*
Acute cerebrovascular accidents	43	21.72	74.00	62.60	65.40	51.80	65.30
AIDS	0	0.00	0.00	0.00	0.00	0.00	0.00
Chronic kidney disease	25	12.63	87.27	83.30	51.00	82.40	54.20
Chronic obstructive pulmonary disease	24	12.12	80.00	72.70	31.80	80.00	35.90
Congestive heart failure	15	7.58	60.00	51.00	32.70	70.30	31.40
Connective tissue disease	11	5.56	44.44	12.50	46.20	13.30	23.10
Dementia	57	28.79	90.00	83.60	85.00	84.90	72.90
Diabetes mellitus	73	36.87	92.00	93.50	87.80	92.10	85.20
Hemiplegia	20	10.10	83.72	24.00	0.00	33.30	0.00
Leukaemia	18	9.09	92.31	66.70	53.70	36.40	70.30
Liver disease	7	3.54	46.16	42.90	40.00	52.20	52.20
Lymphoma	3	1.52	54.54	66.70	50.00	54.50	66.70
Myocardial infarction	43	21.72	87.06	84.80	63.80	81.20	67.20
Peptic ulcer disease	11	5.56	70.00	69.00	50.00	75.00	48.00
Peripheral vascular disease	22	11.11	75.00	83.30	67.80	79.20	61.30
Solid tumour	55	27.78	81.30	83.90	70.20	87.70	63.90

- the NEWS score was divided into three severity categories: mild (1–4), moderate (5–6), and severe (>6);
- diagnosis codes were classified into 17 distinct groups, as established by the Italian Ministry of Health.

For the statistical evaluation, we presented descriptive metrics including absolute frequencies, relative proportions, percentage differences, 95% Confidence Interval (CI), and associated p-values, with statistical significance assessed via the Chi-Square test. An initial comparison was performed among four machine learning algorithms, that is Logistic Regression (LR), Random Forest (RF), Extreme Gradient Boosting (XGB), and Light Gradient Boosting (LGBM). Among these, the LR model exhibited the most favourable balance across evaluation metrics such as sensitivity, specificity and F1-score, which are preferred by clinicians when considering a model for routine clinical use. Thus, LR enabled the investigation of the relationship between overnight ED stays and in-hospital mortality, as well as the identification of other influential predictors. Due to class imbalance within the dataset, specifically, the underrepresentation of the EDOS cohort, a random undersampling strategy was implemented. Sampling ratios were varied from 0.1 to 1.0 in increments of 0.1 to evaluate model robustness across different balancing scenarios. To isolate the most relevant predictors, we applied Recursive Feature Elimination, reporting the resulting odds ratios along with their corresponding 95% CI.

5 Results and Discussion

This section presents the results of both the comorbidity extraction task and the analysis of the impact of overnight stays on in-hospital mortality in patients over 75 years.

5.1 Comorbidities Extraction

In order to assess the performance of the two comorbidity extraction approaches, we computed micro-averaged precision, recall, and F1-scores, using as ground truth a reference set of 200 clinical records manually annotated by domain experts (Table 2). It is worth noting that the selection of the 200 clinical notes was undertaken directly by emergency medical staff to ensure coverage of all relevant comorbidities and the variety of abbreviation patterns employed, bearing in mind that a single patient may present with multiple comorbidities simultaneously.

Table 3: Comparison of comorbidities extraction.

	Precision	Recall	F1-score
Rule-based	72.83	**90.65**	**80.77**
ChatGPT *no pre-proc.*	69.63	82.67	75.59
Llama *no pre-proc.*	52.53	74.46	61.60
ChatGPT *with pre-proc.*	**75.64**	76.35	75.99
Llama *with pre-proc.*	54.12	68.97	60.65

Considering the F1-score performance, the rule-based method exhibits greater consistency across all comorbidities (Table 2). Overall, the performance of the rule-based method is reported in the first row of Table 3. While the precision metric indicates a moderate incidence of false positives, the high recall highlights the method's strong capability in capturing relevant comorbidities. Additionally, the F1-score, which balances precision and recall, reinforces the overall effectiveness of this approach for extracting comorbidity information from unstructured clinical narratives.

Both the ChatGPT 4o and LLaMA 3 8B models were initially queried using Prompt 2, producing a JSON output in which each comorbidity was annotated with a binary indicator (1 if present, 0 if absent). Table 3 (*no pre-proc.* rows) presents the corresponding performance metrics. Given the disproportionate number of true negatives in the dataset, accuracy was excluded from the evaluation as it provided limited interpretative value. Although ChatGPT 4o exhibited slightly better performance than LLaMA 3 8B, the results revealed a tendency of the model to assign comorbidities that

were not actually documented in the clinical text, indicating the necessity for methodological refinement. To address this issue, a pre-processing stage was introduced, also based on Prompt 2, which led to an improvement in precision through the reduction of false positives, although it caused a marginal decline in recall (see Table 3, *with pre-proc.* rows). Both models benefited from this prompt improvement, particularly ChatGPT 4o to a greater extent. Nevertheless, even with this enhancement, the models continued to attribute undocumented comorbidities to patients. This pattern illustrates a recurrent limitation observed in LLMs, commonly known as "hallucination", which highlights the critical need for cautious interpretation and the use of supplementary evaluation metrics beyond accuracy. The decision to use two "general-purpose" LLMs was motivated by two main considerations. First, these models represent the current frontier of LLMs, having been trained on vast amounts of data, including, it is reasonable to assume, a significant portion of the medical literature. Second, the clinical text to be processed was mostly authored by nurses and often includes abbreviations and shorthand typical of informal nurse–physician communication within a specific hospital context. Some of these expressions, such as the use of the letter K to indicate the presence of a tumour, are highly idiosyncratic and not documented in the broader literature. Moreover, we tested the most recent version of BioMedLM[4], a medical domain pre-trained model, using identical prompts to ensure a fair comparison. BioMedLM performed significantly worse than the two models we employed: it tends to repeat the given instruction, reproduce the patient-related text, and provide references (line numbers) to articles containing related topics, but it does not return the list of comorbidities from which the patient is supposed to suffer.

Given the comparative evaluation of the two approaches, the rule-based method was selected for subsequent analyses, owing to its higher recall and F1-score. Although the ChatGPT 4o model achieved higher precision using pre-processing, the rule-based technique proved more dependable in consistently identifying comorbidities within unstructured clinical documentation. For completeness, it should be mentioned that other advanced extraction frameworks, such as CLAMP, cTAKES, and MetaMap—were also tested; however, they did not produce meaningful improvements. The likely advantage of the rule-based system lies in its adaptability to the specific linguistic patterns and documentation practices prevalent in individual ED environments, which would otherwise necessitate the bespoke training of more complex language models.

5.2 Overnight Stays and In-Hospital Mortality

The cohort analysed for overnight stay patterns comprised 20,009 individuals, of whom 3,064 (15.3%) were classified under the EDOS group, while the remaining 16,945 (84.7%) were assigned to the Ward group.

Table 4 summarises the comparative analysis between the EDOS and Ward groups. The cohort exhibited a mean age of 85 years, with males representing 45% of the total sample. A CCI score of ≥ 5 was observed in 65.2% of patients, and trauma-related presentations appeared more frequently among those in the EDOS group. No statistically significant differences were detected between the two groups with respect to age distribution, sex, CCI, NEWS categories, or trauma-related emergency visits. Notably, the proportion of patients in the EDOS group was elevated during the COVID-19 period (57.7%) in comparison to the pre-pandemic interval (48.5%).

As shown in Table 5, the most notable distinctions between the EDOS and Ward groups pertain to diagnostic categories including circulatory system disorders, injuries and poisonings, gastrointestinal diseases, genitourinary conditions, infectious and parasitic diseases, as well as haematological disorders.

With respect to hospital length of stay, patients in the EDOS group exhibit a higher incidence of extended durations, which aligns with expectations. Specifically, the proportion of individuals remaining in the emergency department for more than six hours exceeds that of the Ward group by roughly 2%, as illustrated in Table 6.

Given that overnight stays did not exhibit a statistically significant association with in-hospital mortality, we compared four different machine learning algorithms in describing the in-hospital mortality event, namely LR, RF, XGB, and LGBM. To address the imbalance in the dataset, a random undersampling technique was implemented, exploring sampling ratios from 0.1 to 1. This approach was intended to enhance the robustness of the model. Ultimately, we chose a sampling ratio of 1, resulting in equal numbers of instances for both the majority and minority classes. Feature selection was carried out through Recursive Feature Elimination, which isolated the most informative predictors of in-hospital death. As shown in Table 7, among the models, LR demonstrated the best overall performance, offering the added advantage of easier interpretability. For a fair comparison, we also present the precision–recall curve for all models in a single plot (Figure 2). Up to a recall of 0.2, the models perform similarly; beyond this, LR outperforms the others. It is worth noting that, had the objective been to propose a predictive model, the results reported in Table 7 would not be satisfactory. Our primary aim, however, was aligned with a more clinically oriented perspective: to conduct a descriptive analysis and to estimate the effect of independent variables on the outcome, including those whose extraction is non-trivial, such as the CCI. For this reason, we did not test more complex models, which, while potentially offering better classification performance, would have compromised interpretability.

The LR model yielded an overall accuracy of 0.700 ± 0.007, with an optimal classification threshold identified at 0.472. As shown in Figure 3, the model attained a sensitivity of 74.1% and a specificity of 55.9% at the optimal cutoff. Sensitivity captures the model's effectiveness in detecting patients who experienced in-hospital mortality, while specificity indicates its ability to correctly identify those who survived. These performance metrics suggest that the model is capable of reliably flagging individuals at elevated risk of mortality, while preserving an acceptable level of classification accuracy for non-fatal outcomes.

We also investigated the primary variables contributing to mortality risk. Specifically, among the variables retained, infectious and parasitic diseases (OR = 4.12, 95% CI: 3.19–5.31) and respiratory system disorders (OR = 3.44, 95% CI: 2.91–4.07) represented the most prominent risk factors. These were followed by malignant neoplasms (OR = 3.37, 95% CI: 2.53–4.49) and a CCI equal to or

[4]https://huggingface.co/stanford-crfm/BioMedLM.

Table 4: Patient cohort characteristics summary.

	Total (%)	EDOS group (%)	Ward group (%)	Difference [95% CI]	p-value
Age >85 *years*	9,026 (45.1)	1,398 (45.6)	7,628 (45.0)	0.6 [-1.3,2.5]	0.532
Sex *males*	9,044 (45.2)	1,449 (47.3)	7,595 (44.8)	2.5 [0.6,4.4]	0.011
CCI ≥5	13,046 (65.2)	2,031 (66.3)	11,015 (65)	1.3 [-0.5,3.1]	0.171
NEWS >6	2,317 (11.6)	335 (10.9)	1,982 (11.7)	-0.8 [-2.0,0.5]	0.224
Trauma-related	6,219 (31.1)	1,005 (32.8)	5,214 (30.8)	2.0 [0.2,3.8]	0.025
COVID period	9,992 (49.9)	1,768 (57.7)	8,224 (48.5)	9.2 [7.3,11.0]	<0.001

Table 5: ICD-9 diagnosis profiles of patient groups.

ICD-9 Codes	Total (%)	EDOS group (%)	Ward group (%)	Difference [95% CI]	p-value
Circulatory System Diseases (390-459)	5,121 (25.6)	675 (22.0)	4,446 (26.2)	-4.2 [-5.8,-2.6]	<0.001
Respiratory System Diseases (460-519)	4,033 (20.2)	682 (22.3)	3,351 (19.8)	2.5 [0.9,4.1]	0.002
Injury and Poisoning (800-999)	2,697 (13.0)	306 (10.0)	2,391 (14.1)	-4.1 [-5.3,-2.9]	<0.001
Digestive System Diseases (520-579)	2,403 (12.0)	458 (14.9)	1,945 (11.5)	3.4 [2.1,4.8]	<0.001
Genitourinary System Diseases (580-629)	1,467 (7.3)	281 (9.2)	1,186 (7.0)	2.2 [1.1,3.3]	<0.001
Infectious and Parasitic Diseases (001-139)	1,217 (6.1)	250 (8.2)	967 (5.7)	2.5 [1.4,3.5]	<0.001
Neoplasms (140-239)	972 (4.9)	134 (4.4)	838 (4.9)	-0.5 [-1.3,2.5]	0.175
Blood Diseases (280-289)	531 (2.7)	42 (1.4)	489 (2.9)	-1.5 [-2.0,-1.0]	<0.001
Symptoms, Signs, and Ill-defined Conditions (780-799)	451 (2.3)	74 (2.4)	377 (2.2)	0.2 [-0.4,0.8]	0.514
Endocrine, Nutritional, and Metabolic (240-279)	439 (2.2)	74 (2.4)	365 (2.2)	0.2 [-0.3,0.9]	0.364
Nervous System, Sense Organs Diseases (320-389)	263 (1.3)	35 (1.1)	228 (1.3)	-0.2 [-0.6,0.2]	0.363
Mental Disorders (290-319)	220 (1.1)	29 (0.9)	191 (1.1)	-0.2 [-0.5,0.2]	0.377
Musculoskeletal System Diseases (710-739)	122 (0.6)	15 (0.5)	107 (0.6)	-0.1 [-0.4,0.2]	0.353
Skin Diseases (680-709)	43 (0.2)	8 (0.3)	35 (0.2)	0.1 [-0.1,0.3]	0.548
External Causes (E, V codes)	27 (0.1)	1 (0.03)	26 (0.2)	-0.17 [-0.2,0]	0.094
Congenital Malformations (740-759)	3 (0.01)	0	3 (0.02)	0 [-0.1,0.1]	0.461

Table 6: Length of stay profiles of patient groups.

Length of stay	Total (%)	EDOS group (%)	Ward group (%)	Difference [95% CI]	p-value
≤6 hrs.	17,438 (87.2)	2,491 (81.3)	14,947 (88.2)	-6.9 [-8.4 ,5.5]	<0.001
6 to ≤12 hrs.	2,316 (11.6)	417 (13.6)	1,899 (11.2)	2.4 [1.1,3.7]	<0.001
>12 to ≤24 hrs.	194 (1.0)	95 (3.1)	99 (0.6)	2.5 [1.9,3.2]	<0.001
>24 hrs.	61 (0.3)	61 (2.0)	0	2.0 [1.5,2.5]	<0.001

greater than 4 (OR = 2.75, 95% CI: 1.94–3.91). Advanced age, defined as over 85 years (OR = 1.70, 95% CI: 1.49–1.94), along with circulatory system conditions (OR = 1.89, 95% CI: 1.61–2.23), also demonstrated a moderate association with elevated risk. In addition, elevated NEWS scores exceeding 6 (OR = 1.29, 95% CI: 1.06–1.57) reached statistical significance, although they contributed less to overall model performance. These results highlight the critical role of severe infections, respiratory complications, and comorbid burden in influencing in-hospital mortality, as visualised in Figure 4. The inclusion of CCI among the significant predictors reinforces its clinical utility and supports the relevance of automated extraction techniques such as the one proposed in this research.

In Figure 5, we present a summary of mean positive and negative SHAP values for selected clinical features to provide an interpretable visualisation of their directional impact on in-hospital mortality predictions. This plot highlights how individual characteristics either increase (positive SHAP value) or decrease (negative SHAP value) the predicted risk of in-hospital mortality of the model. In particular, it presents the mean SHAP values for the selected predictors in the model, illustrating how the presence (red bars) or

Table 7: Models performance: in-hospital mortality.

	Sensitivity	Specificity	F1-Score
LR	**68.20**	62.90	**65.50**
LGBM	67.80	61.90	64.90
XGB	66.40	**63.00**	64.40
RF	60.50	61.00	59.80

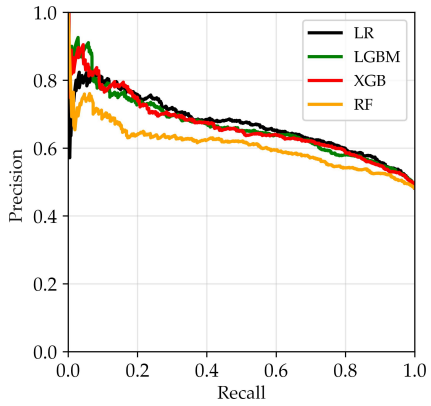

Figure 2: Precision–Recall curves of all models.

absence (blue bars) of each variable influences the predicted risk. As all variables are binary and categorical, the red bars indicate the average increase in predicted risk when the variable is present. In contrast, the blue bars reflect the average decrease when it is absent. For instance, the presence of respiratory diseases, being 85 years or older, or circulatory diseases contributes most strongly to a higher predicted risk, whereas their absence is associated with a reduced risk. Other variables, such as infectious diseases, a CCI of 4 or more, neoplasms, and a NEWS greater than 6, also increase risk when present and reduce it when absent, albeit to a lesser extent. Notably, the impact of neoplasm absence (blue bar) appears smaller than that of other variables, suggesting a more limited protective effect. This visualisation supports the interpretation of both the direction and the relative strength of each feature's contribution to the model's predictions. In our case, the CCI is one of the key variables associated with an increased risk of in-hospital mortality, showing a positive impact at higher values. For example, the in-hospital mortality rate among patients with CCI < 4 was 5.08%, whereas it rose to 15.77% among patients with CCI ≥ 4, which increases the in-hospital mortality risk.

This study focuses on a specific Italian region. At present, no publicly available datasets exist in Italy that would allow us to extend the analysis further. This limitation is particularly relevant given the potential presence of local "dialects" in nurse–physician communication, which may vary considerably across institutions and regions. Although the conclusions cannot be considered definitive, we believe our work provides a meaningful contribution, particularly when compared with prior studies based on datasets covering only 3 [20] or 7 [16] days. Moreover, some variables, such as triage priority, were tested but showed no significant association, while others (*e.g.*, transfer delays) were not available. By contrast,

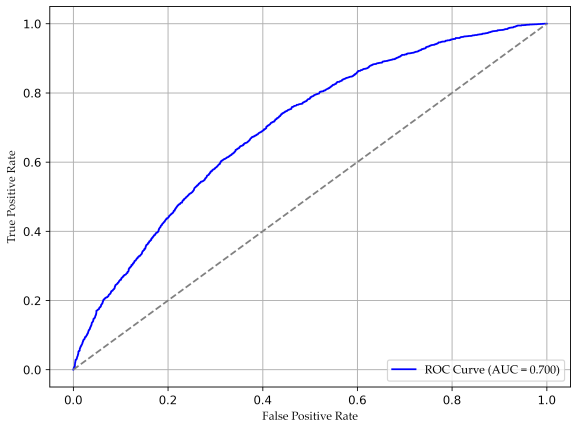

Figure 3: Receiver operating characteristic curve illustrating the trade-off between sensitivity and specificity.

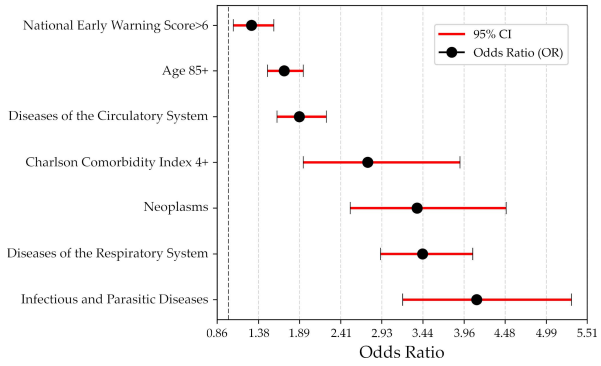

Figure 4: Estimated odds ratios with corresponding 95% CI for the principal variables associated with in-hospital mortality.

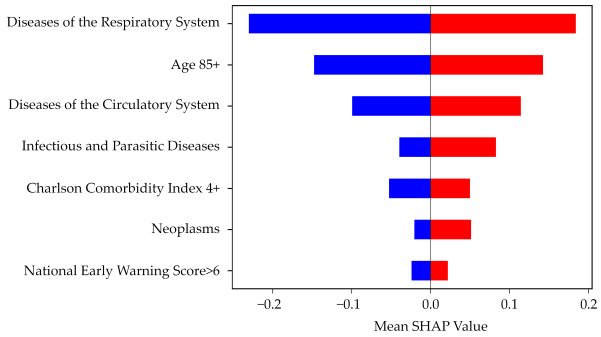

Figure 5: Mean positive (*red*) and negative (*blue*) SHAP values.

factors like socioeconomic status are less relevant in the Italian publicly funded healthcare system, while ward bed occupancy, which could influence the outcome, is mitigated by protocols enabling transfer to a dense network of affiliated facilities. For these reasons, we believe the scientific community can still benefit from this initial effort, which is based on a real-world dataset with substantial temporal depth.

6 Conclusion

In clinical settings where time constraints and resource allocation critically influence patient outcomes, recognising conditions that may negatively affect the care trajectory is of paramount importance. However, existing research does not offer a unified perspective regarding the effect of extended ED stays during nighttime hours on the likelihood of in-hospital mortality.

This study found no evidence of an association between overnight stays in the ED and in-hospital mortality, based on a substantial six-year dataset. Stronger predictors included comorbidity burden, severity at presentation, age, and infectious, respiratory, and circulatory system diseases, whereas overnight stay duration was not statistically significant. Given the predictive value of the CCI, careful extraction of comorbidity data was performed. Comparing a rule-based approach with LLMs for extracting comorbidities from unstructured narratives, the rule-based method outperformed the LLM-based approach. This highlights the importance of deriving structured clinical indicators, such as the CCI and NEWS, from unstructured text, especially with non-English records and absent standardised coding. It also suggests that LLM-based models may not yet achieve the level of reliability required for this task, given the high accuracy needed, which cannot be guaranteed by relying solely on statistical patterns learned from training corpora.

As part of future developments, we plan to integrate a multidimensional geriatric evaluation tool (*e.g.*, the Brief Multidimensional Prognostic Index), which will necessitate a dedicated extraction method tailored for raw data, given its usual absence in standard patient records. With respect to comorbidity extraction, further work will focus on fine-tuning LLMs by enlarging the dataset to encompass greater variability in documentation styles and terminologies.

Acknowledgments

This work was supported by the MIPEPS project (E45J19000810002), co-funded by FIN-RER 2020 of the Emilia-Romagna Region.

References

[1] Samer Badr, Andrew Nyce, Taha Awan, Dennise Cortes, Cyrus Mowdawalla, and Jean-Sebastien Rachoin. 2022. Measures of Emergency Department Crowding, a Systematic Review. How to Make Sense of a Long List. *Open Access Emergency Medicine* 14 (2022), 5–14.

[2] Silvia Casola, Tiziano Labruna, Alberto Lavelli, and Bernardo Magnini. 2023. Testing ChatGPT for stability and reasoning: a case study using Italian medical specialty tests. (2023).

[3] Mary E Charlson, Danilo Carrozzino, Jenny Guidi, and Chiara Patierno. 2022. Charlson comorbidity index: a critical review of clinimetric properties. *Psychotherapy and psychosomatics* 91, 1 (2022), 8–35.

[4] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. 1987. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases* 40, 5 (1987), 373–383.

[5] Emma Chiaramello, Francesco Pinciroli, Alberico Bonalumi, Angelo Caroli, and Gabriella Tognola. 2016. Use of "off-the-shelf" information extraction algorithms in clinical informatics: A feasibility study of MetaMap annotation of Italian medical notes. *Journal of biomedical informatics* 63 (2016), 22–32.

[6] Claudio Crema, Tommaso Mario Buonocore, Silvia Fostinelli, Enea Parimbelli, Federico Verde, Cira Fundarò, Marina Manera, Matteo Cotta Ramusino, Marco Capelli, Alfredo Costa, et al. 2023. Advancing Italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application. *Journal of Biomedical Informatics* 148 (2023), 104557.

[7] Vincent De Groot, Heleen Beckerman, Gustaaf J Lankhorst, and Lex M Bouter. 2003. How to measure comorbidity: a critical review of available methods. *Journal of clinical epidemiology* 56, 3 (2003), 221–229.

[8] Prajwal Dhakal, Valerie Shostrom, Zaid S Al-Kadhimi, Lori J Maness, Krishna Gundabolu, and Vijaya Raj Bhatt. 2020. Usefulness of Charlson comorbidity index to predict early mortality and overall survival in older patients with acute myeloid leukemia. *Clinical Lymphoma Myeloma and Leukemia* 20, 12 (2020), 804–812.

[9] Debopriya Ghosh, Javier Cabrera, Tarek N Adam, Petros Levounis, and Nabil R Adam. 2016. Comorbidity patterns and its impact on health outcomes: two-way clustering analysis. *IEEE transactions on big data* 6, 2 (2016), 359–368.

[10] The Lancet Regional Health-Europe. 2025. The Italian health data system is broken. 101206 pages.

[11] Mark Holland and John Kellett. 2023. The United Kingdom's National Early Warning Score: should everyone use it? A narrative review. *Internal and Emergency Medicine* 18, 2 (2023), 573–583.

[12] Mohd Javaid, Abid Haleem, and Ravi Pratap Singh. 2023. ChatGPT for healthcare services: An emerging stage for an innovative perspective. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations* 3, 1 (2023), 100105.

[13] Trine Lembrecht Jørgensen, Jesper Hallas, Lotte Holm Land, and Jørn Herrstedt. 2010. Comorbidity and polypharmacy in elderly cancer patients: The significance on treatment outcome and tolerance. *Journal of geriatric oncology* 1, 2 (2010), 87–102.

[14] R London. 2012. National early warning score (NEWS): standardising the assessment of acute-illness severity in the NHSdreport of a working party. *London: Royal College of Physicians* (2012).

[15] Manal Makram and Ammar Mohammcd. 2024. AI Applications in Medical Reporting and Diagnosis. In *2024 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC)*. IEEE, 185–192.

[16] Òscar Miró, Sira Aguiló, Aitor Alquézar-Arbé, Cesáreo Fernández, Guillermo Burillo, Sergio Guzmán Martínez, María Esther Martínez Larrull, Andrea B Bravo Periago, Claudia Lorena Amarilla Molinas, Carolina Rangel Falcon, et al. 2024. Overnight stay in Spanish emergency departments and mortality in older patients. *Internal and Emergency Medicine* 19, 6 (2024), 1653–1665.

[17] Scott B Murray, David W Bates, Long Ngo, Jacob W Ufberg, and Nathan I Shapiro. 2006. Charlson Index is associated with one-year mortality in emergency department patients with suspected infection. *Academic Emergency Medicine* 13, 5 (2006), 530–536.

[18] Samuli T Niiranen, Jari M Yli-Hietanen, and Larry A Nathanson. 2008. Toward reflective management of emergency department chief complaint information. *IEEE Transactions on Information Technology in Biomedicine* 12, 6 (2008), 763–767.

[19] Jesse M. Pines and Richard T. Griffey. 2015. What We Have Learned From a Decade of ED Crowding Research. *Academic Emergency Medicine* 22, 8 (July 2015), 985–987.

[20] Melanie Roussel, Dorian Teissandier, Youri Yordanov, Frederic Balen, Marc Noizet, Karim Tazarourte, Ben Bloom, Pierre Catoire, Laurence Berard, Marine Cachanado, et al. 2023. Overnight stay in the emergency department and mortality in older patients. *JAMA internal medicine* 183, 12 (2023), 1378–1385.

[21] Neha Sathe, Vaibhav Deodhe, Yash Sharma, and Anand Shinde. 2023. A Comprehensive Review of AI in Healthcare: Exploring Neural Networks in Medical Imaging, LLM-Based Interactive Response Systems, NLP-Based EHR Systems, Ethics, and Beyond. In *2023 International Conference on Advanced Computing and Communication Technologies*. IEEE, 633–640.

[22] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5 (2010), 507–513.

[23] Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association* 25, 3 (2018), 331–336.

[24] Antonia S Stang, Jennifer Crotts, David W Johnson, Lisa Hartling, and Astrid Guttmann. 2015. Crowding measures associated with the quality of emergency department care: a systematic review. *Academic Emergency Medicine* 22, 6 (2015), 643–656.

[25] Gabriele Valli, Elisabetta Galati, Francesca De Marco, Chiara Bucci, Paolo Fratini, Elisa Cennamo, Carlo Ancona, Nicola Volpe, and Maria Pia Ruggieri. 2021. In-hospital mortality in the emergency department: clinical and etiological differences between early and late deaths among patients awaiting admission. *Clinical and Experimental Emergency Medicine* 8, 4 (2021), 325.

[26] Amadeo Jesus Wals Zurita, Hector Miras del Rio, Nerea Ugarte Ruiz de Aguirre, Cristina Nebrera Navarro, Maria Rubio Jimenez, David Muñoz Carmona, and Carlos Miguez Sanchez. 2025. The Transformative Potential of Large Language Models in Mining Electronic Health Records Data: Content Analysis. *JMIR Medical Informatics* 13 (2025), e58457.

[27] Shuang Wang, Jun-Yi Gao, Xiang Li, Yu Wu, Xiao-Xia Huo, Chao-Xia Han, Meng-Jie Kang, Hong Sun, Bao-Lan Ge, Yu Liu, et al. 2020. Correlation between crowdedness in emergency departments and anxiety in Chinese patients. *World Journal of Clinical Cases* 8, 13 (2020), 2802.