

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

Using Large Language Models to Derive Prognostic Indices from Emergency Department Records: A Retrospective Study

Relatore:
Chiar.mo Prof.
Danilo Montesi

Presentata da:
Lorenzo Giarrusso

Correlatore:
Dott.
Flavio Bertini

Sessione I
Anno Accademico 2024/2025

Sommario

Negli ultimi anni, la crescente complessità nella gestione dei pazienti anziani ha evidenziato la necessità di strumenti prognostici multidimensionali in grado di valutare in modo più completo la fragilità geriatrica. Il Multidimensional Prognostic Index (MPI) e la sua versione abbreviata, il Brief MPI, si sono dimostrati strumenti efficaci nella predizione della mortalità entro un anno nei pazienti anziani. Tuttavia, il loro calcolo manuale è spesso poco pratico in contesti di emergenza, a causa della scarsità di tempo e dell'assenza di dati strutturati.

Questa tesi esplora la fattibilità tecnica dell'impiego di Large Language Models (LLM) per la stima automatica del Brief MPI su dati clinici retrospettivi del Pronto Soccorso di Forlì (2017–2022). A partire da un dataset contenente oltre 110.000 accessi di pazienti over 65, è stata sviluppata una pipeline che genera automaticamente dei prompt testuali contenenti informazioni cliniche strutturate e semi-strutturate, successivamente utilizzati per interrogare un LLM (DeepSeek-V3). Il modello ha restituito, per ciascun paziente, stime numeriche dei domini che compongono il Brief MPI.

I risultati mostrano che, pur in assenza di fine-tuning specifico, l'approccio consente di ottenere stime coerenti in molti casi, ma che permangono criticità legate a fenomeni di allucinazione, ambiguità interpretativa e sensibilità alla qualità dei dati. Il lavoro evidenzia le potenzialità dell'uso degli LLM per l'automazione della valutazione prognostica geriatrica e propone linee di sviluppo per migliorare affidabilità e sicurezza in contesti clinici reali.

Contents

1	Introduction	3
1.1	Purpose of the thesis	3
1.2	Motivation and relevance	4
1.3	Overview of the thesis structure	4
2	Context and Objectives	6
2.1	Healthcare context	6
2.1.1	The Emergency Department of Forlì	6
2.1.2	Clinical data source	7
2.2	Objectives of the study	7
2.2.1	Automatic calculation of the Brief MPI	7
3	Materials and Methods	9
3.1	Methodological approach to Data Mining	9
3.1.1	KDD Roadmap: a brief overview	9
3.1.2	How it was applied to the project	10
3.2	The dataset	12
3.2.1	Origin and characteristics of the data	12
3.2.2	Cleansing, pre-processing and anonymization	13
3.3	The Brief MPI	15
3.3.1	Definition and components	15
3.3.2	Relevant data fields used for computing the Brief MPI	16
3.3.3	Clinical interpretation of scores	17
3.4	Chosen Large Language Model	18
3.4.1	Model selected and query dates	18
3.4.2	OpenRouter and interaction with the LLM	19
3.4.3	Prompt engineering strategy	22
4	Experimentation	24
4.1	Operational pipeline and description of the workflow	24
4.2	Output analysis	26

4.2.1	Correct vs. incorrect cases	26
4.2.2	Error classification	26
4.3	Encountered issues	27
4.3.1	Hallucinations: examples and possible causes	28
4.3.2	Prompt constraints and their effect	29
4.3.3	Data errors (missing values, noise, misinterpretation)	29
5	Discussion	31
5.1	Critical evaluation of results	31
5.1.1	Distribution of Brief MPI estimates	31
5.1.2	Distribution of domain estimates	32
5.1.3	Clinical plausibility	32
5.1.4	LLM behavior and robustness	33
5.2	Limitations of the approach	33
5.2.1	Model-related limitations	33
5.2.2	Prompt-related limitations	33
5.2.3	Data quality and preprocessing constraints	33
5.2.4	Evaluation challenges	34
6	Conclusions	35
6.1	Summary of the results	35
6.2	Final considerations	35
6.3	Suggestions for future developments	36
A	Examples of Prompts and Responses	38
B	Brief MPI Indices and Criteria	44
	Acknowledgments	46
	References	47

Chapter 1

Introduction

1.1 Purpose of the thesis

The demographic shift which has characterized the recent decades in developed countries has increased the criticality of care of elderly subjects, thus enhancing the need for assessments which may provide a more comprehensive evaluation for the frailty of a patient. Traditional and simpler tools, such as Clinical Frailty Scale (CFS)[16], focus on only one or few aspects describing the status of a patient, such as their clinical condition. On the other hand, the development of Comprehensive Geriatric Assessment (CGA)[19] tools allows clinicians to obtain a more detailed and comprehensive multidimensional evaluation of the patients, also considering aspects such as their independence in daily activities and their condition in terms of sociability. Among these tools, the *Multidimensional Prognostic Index* (MPI) has been shown to be a valid predictor of 1-year mortality in geriatric patients[13], providing a stronger evaluation of frailty in the elderly when compared to the simpler alternatives mentioned above. Further, an abbreviated form of the MPI, the *Brief MPI*, was found[1] to be a valid substitute for the standard version in contexts where its evaluation might be limited by time constraints or lack of information. As such, usage of the Brief MPI in the context of an Emergency Department may enable us to obtain a more detailed prognostic evaluation of patients and the course of their medical condition during hospitalization.

The objective of this thesis is to evaluate the feasibility of using a *Large Language Model* (LLM) to estimate the Brief Multidimensional Prognostic Index, and in general automatically extract structured information from unstructured free-text data. The model was applied to the retrospective clinical data from the Forlí Emergency Department, collected between 2017 and

2022. Specifically, the model was used to infer the individual components of the Brief MPI using unstructured and semi-structured patient data, with the aim of assessing its potential to support clinical decision-making through automated prognostic scoring. Further, the integration of the MPI fields into the dataset built a foundation for possible future studies focusing on the predictive power of the estimated Brief MPI. These activities were realized as part of my curricular internship at the University of Bologna.

1.2 Motivation and relevance

In recent years, Large Language Models (LLMs) have demonstrated remarkable capabilities in natural language processing (NLP) and generation, opening new frontiers in domains traditionally resistant to automation due to the prevalence of free-text information. Emergency departments (EDs), in particular, generate vast amounts of clinical data that often remain underutilized due to their unstructured nature and time constraints faced by medical staff, leading to a lack of proper standardization in the corresponding data.

Prognostic indices such as the Brief MPI provide essential insight into patient outcomes, yet their manual calculation is rarely feasible in routine practice due to their reliance on information regarding patients that are not obtained during acceptance in the emergency department. Using LLMs to extract and interpret relevant clinical features retrospectively could enable large-scale computation of such indices, facilitating retrospective cohort analyses, and potentially informing future clinical decision support tools.

This work aims to explore whether LLMs can be reliably used to derive prognostic indices from real-world clinical records, contributing both to the development of clinical NLP applications and to the broader goal of medical data automation, while possibly serving as further validation of the Brief MPI.

1.3 Overview of the thesis structure

This thesis is organized into six main chapters, in addition to appendices and references.

1. **Introduction:** introduces the context of the study, including the collaboration with the Emergency Department of Forlì and the overall subject of the work.
2. **Context and Objectives:** describes the context in which the data were collected and outlines the objective of the study.

3. **Materials and Methods:** presents the methodology followed in the study, describes the dataset and its relevant fields, and provides theoretical background of the Brief MPI, outlining its structure, clinical relevance, and scoring system. Finally, we give more details about the chosen Large Language Model and the approach we followed in interacting with it.
4. **Experimentation:** details the adopted pipeline, with further attention given to the output, errors, and issues encountered during the process, with particular regards to hallucinations and their relation with prompts.
5. **Discussion:** reports the results of the experimentation phase, highlighting both successful predictions and the limitations encountered.
6. **Conclusions:** offers a critical discussion of the findings, possible improvements, and future directions for applying Large Language Models in clinical contexts.

The thesis concludes with appendices that provide supporting materials, including examples of prompts, model outputs, and a checklist describing the criteria used to assess the Brief MPI domains.

Chapter 2

Context and Objectives

2.1 Healthcare context

The project is based on a retrospective analysis of emergency department records from the Forlì Hospital, part of the Romagna Local Health Authority (AUSL Romagna), covering patient admissions between 2017 and 2022 (inclusive). These data, originally collected during routine care, offer a unique opportunity to explore the feasibility of automated prognostic scoring using large language models, as they are characterized by a large quantity of precious information represented through often non-standardized free text. The chapter also outlines the data sources, the structure of the ED, and the key collaborators who contributed to the project’s development and validation.

2.1.1 The Emergency Department of Forlì

The Emergency Department of Forlì Hospital is a public healthcare unit serving the Emilia-Romagna region in northern Italy. It is part of AUSL Romagna (Azienda Unità Sanitaria Locale della Romagna). The ED handles a broad spectrum of cases, ranging from minor injuries to critical medical emergencies, and includes a significant proportion of elderly patients, which were the subjects considered for this study. The department generates a substantial volume of clinical documentation, with the analyzed time frame providing 283,385 records for a yearly average of 47,230, including information such as triage assessments, diagnostic reports, discharge summaries, and free-text clinical notes. This makes it an ideal environment to investigate the potential of natural language processing tools in extracting meaningful prognostic indicators from real-world medical records and testing them against the recorded outcomes of a variety of clinical conditions.

2.1.2 Clinical data source

The dataset used in this project includes electronic health records (EHRs) from ED admissions that occurred between January 2017 and December 2022. The data were extracted retrospectively from the hospital’s clinical information system and encompass both structured fields (e.g. age, sex, International Classification of Diseases for diagnoses) and semi-structured or unstructured content (e.g. clinical narratives and notes written by nurses and doctors). The records were not originally collected for the purpose of calculating prognostic indices, as such it is not possible to properly calculate the Brief MPI based on the strictly defined criteria (Table B.1) in a deterministic way. However, records often contain enough relevant textual information to allow for a retrospective estimation of Brief MPI scores through the use of a LLM for natural language processing. Prior to analysis, all data were anonymized to protect patient and medical staff confidentiality.

2.2 Objectives of the study

The primary aim of this study is to evaluate the technical feasibility of leveraging large language models to automatically reconstruct Brief Multidimensional Prognostic Index scores from retrospective emergency department clinical data. By assessing the capability of these models to accurately synthesize complex clinical information and generate meaningful MPI estimates, we aim to explore their potential utility as tools for clinical decision support, and their overall ability to extract structured data from unstructured textual information. This foundational evaluation serves as a necessary step toward integrating advanced natural language processing techniques into geriatric patient assessment workflows.

2.2.1 Automatic calculation of the Brief MPI

The primary objective of this study is to assess whether a large language model can be used to automatically estimate the components of the Brief Multidimensional Prognostic Index from unstructured and semi-structured data available in historical emergency department records. By extension, our aim was to evaluate the potential of applying the natural language processing power of LLMs to a clinical context, where it may serve as a new tool capable of providing data derived from traditionally obtained information as material to construct a more comprehensive overview of the conditions of elderly patients. Specifically, the goal was to translate free-text and partially structured clinical information — such as triage notes, physician evaluations, diagnoses,

medications, and test results — into quantitative estimates of the Brief MPI domains using a prompt-based interaction with a LLM. This would offer a way to retrospectively reconstruct prognostic scores for elderly patients, even in the absence of standardized data collection at admission, thus providing an alternative approach to large-scale frailty and mortality risk stratification.

Chapter 3

Materials and Methods

3.1 Methodological approach to Data Mining

3.1.1 KDD Roadmap: a brief overview

Knowledge Discovery in Databases (KDD) can be defined as "the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data"[4]. As such, KDD methodologies enable us to extract information from collections of data to build a foundation for high-level decision making. Within the context of ED data, characterized by a huge amount of potentially meaningful information represented in non-standardized and non-normalized formats, the possibility of utilizing semi-automatic methods to extract useful, higher-level knowledge can be critical to provide domain experts with information in a concise format. Thus, adhering to a well-structured approach to KDD was one of the first steps in the realization of this study.

The *KDD Roadmap*, as defined by Jiawei Han et al.[5], models an approach to KDD based on a roadmap made up of 8 phases, through which many different paths may be followed depending on the context of the data being analyzed. The 8 phases can be summarized as follows.

1. **Problem Specification:** this phase involves processes such as database examination and familiarization and determination of feasibility - in terms of tasks and requirements - in order to move from a loosely defined description to a tightly defined problem specification.
2. **Resourcing:** the purpose of this phase is to elaborate data from multiple eventually heterogeneous sources in order to obtain the resources described in the problem specification. The output is an operational

database that is complete and consistent in terms of structure, format, conventions for missing values, etc.

3. **Data Cleansing:** the purpose of this phase is to prepare the data for the following phases through a process designed to be performed only once per database. The tasks faced by this cleansing process are sampling, balancing, and handling erroneous outliers and missing data.
4. **Pre-Processing:** this phase further elaborates the data to improve the quality and performance of following phases through processes which may be executed multiple times per database throughout the KDD process. Tasks which may be dealt with in this phase are extraction of new features, feature subset selection, discretizations, etc.
5. **Data Mining:** in this phase, the data mining algorithms to be used are determined; that is, we choose the automatic or semi-automatic methods to be applied to the data to extract information from them. This choice is made based on the tasks defined in the problem specification. Contextually, algorithm and problem parameters are set, and results may be subject to a preliminary analysis based on a measurement of their interestingness and on an estimation of the corresponding error.
6. **Evaluation:** in this phase, resulting patterns are evaluated in terms of performance on the testing set, if present, and their generalizability and suitability. In addition, results may be visualized to facilitate statistical analysis of them.
7. **Interpretation:** this phase involves domain experts in order to provide a proper interpretation of the validity of the results, with a contextual integration with pre-existing domain knowledge.
8. **Exploitation:** in this phase, discovered and validated patterns are effectively put to use, applying the newfound knowledge to the field in ways that aim to minimize the risks involved and maximize the potential benefits.

3.1.2 How it was applied to the project

The approach to KDD described above was relevant to our activity in this study in particular with respect to the first four phases, that is from Problem Specification to Pre-Processing, covering the Extraction-Transformation-Loading (ETL) process to create a monolithic data warehouse.

1. With the help of the Head Physician of the ED at Forlí, the dataset was examined and familiarized with in relation to the meaning of its fields, the conventions used for codifications, and the importance we could expect from free text fields. Given this knowledge, we first researched open issues related to ED data, and after some discussion it was agreed upon to focus on Brief MPI to be retrospectively estimated through the use of a large language model. This allowed us to reach a more tightly defined problem specification.
2. The data for the Forlí ED was originally split into multiple datasets, one for each year from 2017 to 2022, with some variance in the naming conventions for some fields. As such, the resourcing phase involved merging and integrating these datasets into a single data warehouse; that is, a subject-oriented, integrated, nonvolatile operational database.
3. Data Cleansing was critical in preparing the data for the following phases, as the original datasets presented many issues such as erroneously inserted records, duplicates, and missing values. In addition to dealing with such issues, in this phase we also made sure to only keep records related to patients 65 years or older, as Brief MPI is an index thought specifically for geriatric patients in that age group. This allowed us to greatly reduce the number of records to be dealt with in the following processing phase. Finally, this phase is where the data were anonymized by removing information such as tax code, names, etc.
4. The Pre-Processing phase involved multiple tasks to be performed on the original data in order to obtain more directly-utilizable information and focus on fields that held more significance to speed up the process, contextually enriching the dataset for future use. Initially, the age of each patient was calculated as the difference between their birth date and their Triage date. Next, we banded the age values according to suggestions by our collaborating clinical expert. Similarly, we also banded the values representing the ED waiting times and the arrival times. Finally, we used the given full ICD-9 codes (International Classification of Diseases - 9th revision) for the diagnosis on arrival and the diagnosis at the moment of hospitalization to identify the corresponding ICD-9 category among the 18 defined by the standard classification.[17]

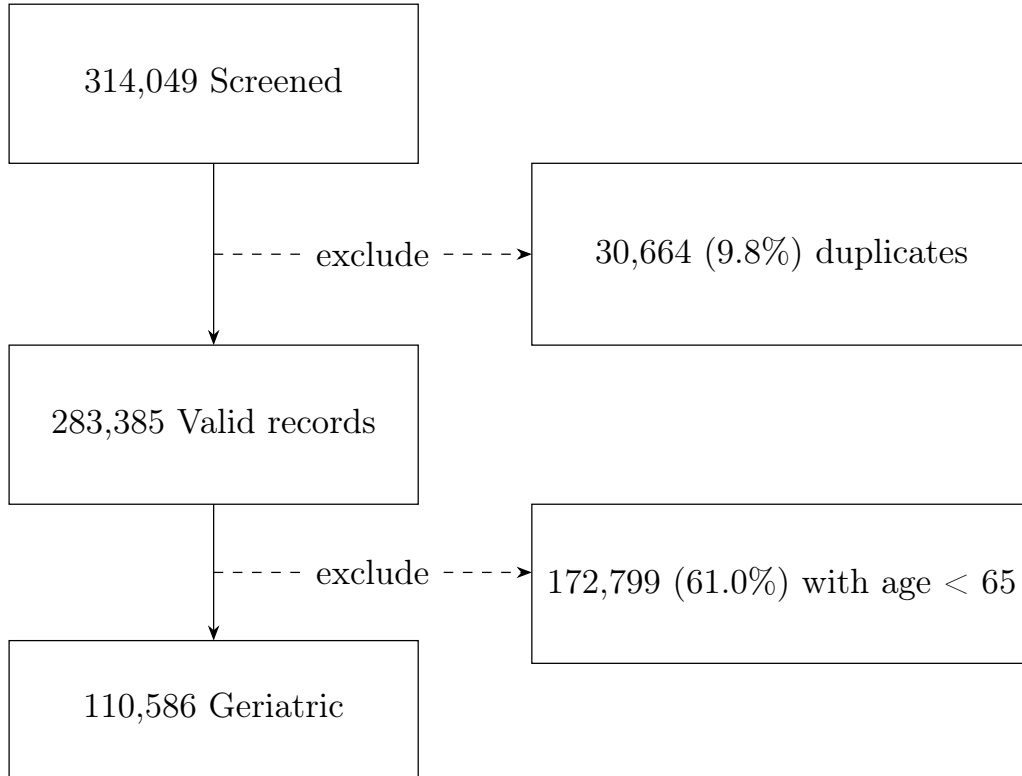


Figure 3.1: Flow diagram of study patients for the study period of 2017-2022.

3.2 The dataset

3.2.1 Origin and characteristics of the data

The data used in this study originates from the Emergency Department of the Forlí Hospital, covering the period from 2017 to 2022. Rather than being provided as a unified dataset, the records were initially distributed across separate files, one for each year, with slight variations in schema and field naming conventions.

This fragmentation required a preliminary data integration phase, where all yearly datasets were merged into a single, subject-oriented data warehouse, as previously described. The resulting dataset includes both structured and semi-structured fields related to ED visits, such as patient demographics, triage information, timestamps, ICD-9 diagnoses, laboratory and radiology orders, prescribed therapies, and free-text notes.

Overall, the dataset comprises a total of 283,385 valid records. Table 3.1 summarizes key descriptive statistics, including sex distribution, triage codes, and age groups.

Category	Count	Percentage
<i>Sex</i>		
Male	145,718	51.42%
Female	137,666	48.58%
Missing	1	<0.01%
<i>Triage code</i>		
White	46,324	16.35%
Green	171,154	60.40%
Blue	44,518	15.71%
Orange	7,829	2.76%
Red	13,560	4.79%
<i>Age group</i>		
0–17	45,690	16.12%
18–29	26,813	9.46%
30–39	23,066	8.14%
40–49	30,053	10.61%
50–59	32,480	11.46%
60–69	29,683	10.47%
70–79	38,476	13.58%
80+	57,124	20.16%

Table 3.1: Summary statistics of the dataset

The mean age of patients is 51.44 years, with a median of 54.00 years.

While the data was not originally collected with prognostic scoring in mind, it contains many of the clinical indicators relevant to estimating the Brief MPI. As such, it serves as a valuable retrospective source for experimentation with automatic index derivation using LLMs.

3.2.2 Cleansing, pre-processing and anonymization

A thorough data cleansing process was essential to prepare the integrated dataset for analysis. This phase involved identifying and removing erroneous records, duplicates, and entries with significant missing information. Additionally, only patients aged 65 or older were retained, in accordance with the geriatric scope of the Brief MPI. This filtering step substantially reduced the dataset size, allowing more focused and efficient processing. The pre-processing phase included several enrichment and transformation steps to make the data more suitable for use with large language models. The patients’ ages were computed by subtracting their birth date from the triage

date, and then categorized into bands following guidance from the collaborating clinician. Similar banding was applied to variables such as ED arrival time and waiting time. Furthermore, diagnosis codes in ICD-9 format were mapped to their corresponding major categories according to the standard 18-group classification, enabling a more abstract understanding of clinical conditions. Anonymization was performed in accordance with privacy and ethical standards. All personally identifying information, such as tax codes and names, was suppressed prior to further processing. In particular, as part of the anonymization process, special attention was given to free-text fields containing nurse and physician notes, which occasionally included the full names of the healthcare professionals who authored such notes. To detect and remove these names, a regular expression pattern was developed and applied using Python. This pattern was designed to identify names in several common formats, including *Firstname Lastname*, *Lastname*, *Firstname*, all-uppercase variants such as *LASTNAME FIRSTNAME*, and names including middle names or compound surnames. The pattern used was the following:

Regular Expression for Name Extraction

```
\b[A-Z] [a-z]+(?:\s[A-Z] [a-z]+)*,?\s[A-Z] [a-z]+\b
|
\b[A-Z]+\s[A-Z]+\b
```

This regular expression is designed to look for two main patterns:

- **Pattern 1: Standard Capitalized Names**

The first part of the regex,

```
\b[A-Z] [a-z]+(?:\s[A-Z] [a-z]+)*,?\s[A-Z] [a-z]+\b
```

aims to capture names that typically include a first name, potentially a middle name, and a last name, in particular:

- A word starting with an uppercase letter, followed by lowercase letters (e.g., “Mary”).
- Optionally followed by one or more similar words separated by spaces (e.g., “Mary Ann”).
- Optionally followed by a comma.
- Followed by a space.
- Followed by another word starting with an uppercase letter and then lowercase letters (the last name, e.g., “Smith”).

- All as a whole word.

- **Pattern 2: All-Caps Names**

The second main part of the regex,

```
\b[A-Z]+\s[A-Z]+\b
```

is designed to capture names that are entirely in uppercase, in particular:

- A sequence of one or more uppercase letters.
- Followed by a space.
- Followed by another sequence of one or more uppercase letters.
- All as a whole word.

Thus, this would match examples like “JOHN DOE”.

3.3 The Brief MPI

The Brief Multidimensional Prognostic Index (Brief MPI) is a validated tool designed to assess frailty and estimate short- and medium-term mortality risk in older adults, especially those admitted to acute care settings such as emergency departments. Derived from the full MPI, the Brief MPI preserves the core multidimensional structure while allowing faster administration and retrospective application when full clinical assessments are unavailable and time is limited. It includes a set of eight domains that cover functional, cognitive, nutritional, and social aspects of a patient’s condition. These domains serve the purpose of providing a multidimensional overview of a patient’s frailty, as opposed to other traditional indices strictly focusing on clinical aspects.

3.3.1 Definition and components

The Brief MPI is composed of eight domains, each capturing a different dimension of a patient’s condition. Each domain is assigned a score of either 0, 0.5 or 1, with each value having a different meaning depending on the domain (Table B.1). The final score, that is the Brief MPI proper, is the arithmetic mean of all domain scores. The eight domains are:

- **Activities of Daily Living (ADL):** based on three items from the Katz ADL index[8], it evaluates independence in dressing and feeding and self-control over urination and defecation;

- **Instrumental Activities of Daily Living (IADL)**: based on three items deemed meaningful by Lawton et al.[9], it evaluates independence in using the telephone, taking medications, and shopping;
- **Mobility**: based on the Barthel mobility scale[15], evaluated inquiring about the abilities to transfer from bed to chair or wheelchair, walking and going up and down the stairs;
- **Cognitive Status**: based on three items from the Short Portable Mental Status Questionnaire (SPMSQ)[12], it evaluates temporal and personal orientation, and computation ability;
- **Nutritional Status**: based on three items from the Mini-Nutritional Assessment (MNA) short form[7], it considers the patient’s BMI, loss of appetite in the last three months, and weight loss in the last three months;
- **Medications**: number of drugs in use as for standard MPI[13];
- **Social Support**: cohabitation status as for standard MPI[13], distinguishing between living alone, with family, and in an institute;
- **Comorbidities**: comorbidities using the Cumulative Illness Rating Scale (CIRS)[11] derived from SELFY-MPI[14] and expressed as number of health diseases requiring drug therapies across 13 health areas.

Further, based on the values of the eight domains, we also calculated two derived fields:

- **Brief MPI**: the index itself, calculated as the average of the eight domains described above; hence, a value in $[0,1]$;
- **Risk**: derived from the Brief MPI score, this field assigns the patient to a risk class according to standard clinical thresholds: low ($\text{MPI} \leq 0.33$), moderate ($0.34\text{--}0.66$), or high ($\text{MPI} > 0.66$).

3.3.2 Relevant data fields used for computing the Brief MPI

In order to generate meaningful and context-rich prompts for the estimation of the Brief MPI domains, a selection of relevant data fields was extracted from the original dataset. These fields were chosen for their potential to provide insight into the patient’s clinical status, functional abilities, cognitive

state, comorbidities, and social context; the dimensions required by the Brief MPI model.

The fields included:

- Age and sex: Basic demographic information necessary for contextualizing the clinical profile.
- Free-text notes written by nurses at triage, often containing first impressions, presenting symptoms, and initial assessments.
- Notes written by physicians, typically containing a structured clinical summary, medical history, and preliminary diagnosis.
- Diagnoses and diagnosis descriptions: ICD-9 codes and their corresponding text descriptions, providing insight into comorbidities and the acute reason for admission.
- The final summary written by the physician and the proposed treatment plan, often reflecting an integrated clinical judgment.
- A miscellaneous free-text field that sometimes included socially relevant information, such as living conditions or family support.
- Laboratory and radiology reports: semi-structured descriptions of tests performed during the ED stay, including findings that could suggest functional or nutritional impairment.
- Prescribed therapies: Information on medications administered or prescribed during the ED stay.

These fields were concatenated into a single prompt structure (see the prompt example), preserving their labels to maintain semantic clarity, and submitted to the LLM to infer the Brief MPI domain values for each patient (see the response example).

3.3.3 Clinical interpretation of scores

The Brief Multidimensional Prognostic Index (Brief MPI) is a clinically validated tool used to stratify older patients based on their multidimensional frailty status. The final MPI score is a continuous value ranging from 0 to 1, obtained by averaging the normalized scores of its component domains. Based on this score, patients are categorized into three distinct risk classes:

- **Low Risk (0.00–0.33):** Patients in this category are generally robust, with minimal functional, cognitive, or nutritional impairment. They typically have better short-term and long-term survival outcomes.
- **Moderate Risk (0.34–0.66):** These patients exhibit intermediate levels of vulnerability and may benefit from closer clinical monitoring and tailored interventions to prevent deterioration.
- **High Risk (0.67–1.00):** Patients in this group are characterized by significant levels of frailty across multiple domains. They are more likely to experience adverse outcomes, such as hospitalization, functional decline, or mortality, and may require more intensive or palliative care planning.

Clinicians may use these categories to inform decision-making, such as hospital admission, treatment aggressiveness, discharge planning, and coordination of post-acute care services. In emergency settings, where time and data availability are limited, having an estimate of the Brief MPI can assist in prioritizing geriatric patients and personalizing their care trajectory.

3.4 Chosen Large Language Model

3.4.1 Model selected and query dates

After testing multiple large language models (LLMs) for the task of estimating the Brief MPI domains based on structured and semi-structured clinical data, we selected DeepSeek-V3-Chat 0324 as the most suitable model for this study.

DeepSeek is a family of large language models developed by the DeepSeek company, with a strong focus on instruction-following, reasoning, and code generation capabilities. The **V3-Chat 0324** variant, released in March 2024, thus coinciding with the start of our study, is optimized for conversational tasks and supports multi-turn reasoning and context tracking, making it particularly effective for tasks that require interpretation of clinical narratives and data summarization.

Model interaction was conducted via the OpenRouter platform, a third-party service that provides unified API access to various LLMs from different providers. OpenRouter enabled seamless access to DeepSeek’s latest version without requiring direct infrastructure deployment, while also offering stable response times and manageable request logging. All model prompts and responses were handled programmatically through this interface.

DeepSeek V3 was chosen as the LLM to be used for this study after experimentation on multiple LLMs highlighted some unwanted behaviors in its alternatives. Table 3.2 reports the responses provided by DeepSeek V3, Claude 3.7 Sonnet, GPT-4o, and Gemini 2.0 Flash on the same prompt; although this is only one example, it encapsulates the general behaviors observed on a larger set of test samples. These models, at the time of our experimentation, represented the state of the art.

This example highlights general tendencies observed across the models. Claude 3.7 consistently produced responses with multiple `null` values, particularly when patient information was ambiguous or incomplete. GPT-4o, on the other hand, rarely deviated from middle-ground estimates (typically returning 0.5 across most fields), indicating a conservative averaging behavior even when some data would justify more polarized judgments. Gemini 2.0 Flash, despite producing reasonable numerical outputs, frequently returned verbose explanations alongside the requested JSON object, complicating the parsing pipeline and violating the strict response format. In contrast, DeepSeek V3 reliably returned clean, JSON-formatted responses without additional text, and its estimates showed sensitivity to clinical cues in the input, balancing precision with format compliance.

These patterns, consistent across a larger set of test samples, motivated the final decision to adopt DeepSeek V3 as the model for Brief MPI estimation.

Overall, DeepSeek V3 stood out for its:

- Robust adherence to the prompt format, that being a JSON object with a key-value pair for each Brief MPI domain;
- Consistent inclusion of estimates for all requested domains;
- Balanced interpretative approach, neither too speculative nor overly cautious, although with some strong limitations in its assumptions (see Section 4.3).

The definitive queries to obtain the Brief MPI estimates were executed in batches, based on the year of the original split datasets, with one or more dataset years processed on each date, as detailed in Table 3.3.

3.4.2 OpenRouter and interaction with the LLM

To interact with the selected language model, DeepSeek-V3-Chat 0324, we relied on OpenRouter, an online platform that provides API-based access to a wide variety of large language models from different providers, including OpenAI, Anthropic, Mistral, Cohere, and DeepSeek.

Model	Response to prompt example	Observations
DeepSeek V3	<pre>{ "Codice di Servizio": "2022_000002", "ADL": 0, "IADL": 0, "Barthel": 0, "SPMSQ": 0.5, "MNA": 0, "Comorb": 0.5, "Farmaci": 0.5, "Stato Coabitativo": 0 }</pre>	-
Claude 3.7	<pre>{ "Codice di Servizio": "2022_000002", "ADL": null, "IADL": null, "Barthel": null, "SPMSQ": 0.5, "MNA": null, "Comorb": 0.5, "Farmaci": null, "Stato Coabitativo": 0 }</pre>	Tendency to estimate "null"
GPT-4o	<pre>{ "Codice di Servizio": "2022_000002", "ADL": 0.5, "IADL": 0.5, "Barthel": 0.5, "SPMSQ": 0.5, "MNA": 0.5, "Comorb": 0.5, "Farmaci": 0.5, "Stato Coabitativo": 0 }</pre>	Rarely estimates anything else besides 0.5
Gemini 2.0	See appendix A	Verbose text despite requests

Table 3.2: Comparison of Brief MPI domain estimates across four LLMs on the same patient input

Dataset Year	Query Execution Date
2017	March 27, 2025
2018	March 28, 2025
2019	March 31, 2025
2020	March 31, 2025
2021	April 1, 2025
2022	April 1, 2025

Table 3.3: Dates on which the LLM queries were executed for each dataset year.

OpenRouter acts as a proxy interface that enables standardized and unified communication with various LLMs through a RESTful API. This abstraction simplifies development and experimentation, as the same codebase can be used to test multiple models by simply switching configuration parameters.

The interaction pipeline with the model followed these steps:

1. **Prompt composition:** For each patient record, a structured textual prompt was created, including relevant clinical information as previously described. The format and phrasing of the prompt were designed in collaboration with a domain expert to provide the most clinically meaningful information.
2. **Request submission:** The prompt was sent as a POST request to the OpenRouter API endpoint, specifying the target model (`deepseek-chat-v3-0324`), along with the necessary headers and authentication token.
3. **Response parsing:** The model returned a JSON object containing the estimated values for each requested domain. These values were extracted and added to the corresponding records in the original database for later analysis.
4. **Validation and logging:** Responses were validated for completeness and format consistency. When inconsistencies or invalid values were detected, the corresponding record was flagged and automatically discarded or manually reviewed.

This programmatic approach allowed for batch processing of thousands of records in a reproducible manner, minimizing manual effort and ensuring a standardized evaluation of model outputs.

Advantages of using OpenRouter in this context included:

- Easy switching between LLM providers during the choice of the model;
- Transparent billing and monitoring, without the need for a subscription;
- Fast response times and high availability, thanks to the distributed infrastructure based on the ability to fall back on other providers when one goes down.

Overall, OpenRouter served as an effective bridge between our local data pipeline and the remote model, enabling efficient experimentation and integration of LLMs into the clinical data mining workflow.

3.4.3 Prompt engineering strategy

We implemented a prompt engineering strategy designed to provide the DeepSeek LLM with all necessary information in a structured and clear format. Our goal was to maximize the accuracy of the estimations and facilitate automated result extraction.

The development of the final prompt structure followed an iterative refinement process. Initial versions of the prompt were simpler and less structured: they included patient data in a raw textual format, followed by a loosely formulated question asking the model to estimate the Brief MPI scores. These early versions often led to incomplete, verbose, or inconsistently formatted responses, which made automatic parsing and analysis difficult.

Through empirical testing and manual inspection of sample outputs, we gradually introduced improvements such as clearer criteria for the values, better selection of the fields included from the original dataset, and stricter formatting instructions. Most notably, we transitioned from open-ended textual prompts to a more standardized template that clearly separated input data from the instructions and enforced a predefined JSON response format. This evolution significantly improved the consistency and machine-readability of the model outputs, while also reducing hallucinations and off-topic completions.

The core of the final strategy involved generating a personalized prompt for each patient using a "one-shot prompting" approach; that is, for every patient, we supplied the model with a single prompt containing all relevant clinical data. Each prompt was dynamically constructed from a pre-processed dataset, where missing values were explicitly stated as such to prevent ambiguity and provide the model with a complete context, even in the absence of specific data points (see the prompt example).

Each prompt included the following patient information categories:

- Demographic data: age and sex;
- Clinical notes: taken by nurses and doctors.
- Diagnoses: primary and secondary diagnoses.
- Additional information: miscellaneous details, laboratory results, radiology reports, and prescribed therapies.

Following the patient data, the prompt explicitly requested the model to estimate values for the eight Brief MPI index domains of interest, as described in Table B.1.

A crucial aspect of our strategy was the precise definition of the desired output format and value scale for each parameter. We instructed the model to provide only the final numerical value for each parameter and specified the interpretation scale (e.g., 0, 0.5, 1) along with its meaning (e.g., "0 indicates little difficulty and 1 indicates much difficulty" for ADL, IADL, and Barthel). This standardization of model responses facilitated the following quantitative analysis.

Finally, to simplify automated result processing, we required the model's response to be a JSON object. This structured format ensured that the extracted data was easily parsable and ready for subsequent automatic analysis.

Chapter 4

Experimentation

4.1 Operational pipeline and description of the workflow

The operational pipeline designed for this study consists of several sequential stages, each performing a specific transformation or processing task on the data. These steps lead to the final estimation of the Brief MPI for each patient using a Large Language Model.

The main components of the pipeline are as follows:

1. **Data Acquisition and Integration:** Original patient data were distributed across yearly datasets (2017–2022). These datasets were merged into a single data warehouse after harmonizing field names and formats. Only records for patients aged 65 or older were retained.
2. **Data Cleaning and Anonymization:** The datasets were cleaned to remove duplicates, erroneous entries, and incomplete records. Personally identifiable information (PII), such as names and tax codes, was removed. A regular expression-based script was also used to remove healthcare personnel names from free-text fields.
3. **Preprocessing and Feature Engineering:** Additional fields were computed, such as patient age (from birthdate and triage date), and binned values for age and waiting time. Diagnoses expressed in ICD-9 codes were mapped to standard clinical categories.
4. **Prompt Generation:** For each patient, a detailed prompt was automatically generated using a Python script. It included all relevant clinical fields and a final instruction asking the model to return numerical estimates of the Brief MPI domains in JSON format.

5. **Interaction with the Language Model:** The prompts were sent to the DeepSeek-V3 LLM via the OpenRouter API. The model returned structured JSON responses with the requested scores.
6. **Postprocessing and Validation:** Model responses were parsed and stored. A sample of the results was reviewed by a clinical expert for plausibility and accuracy. Distributions of the computed values were also analyzed.

A schematic overview of the pipeline is shown in Figure 4.1.

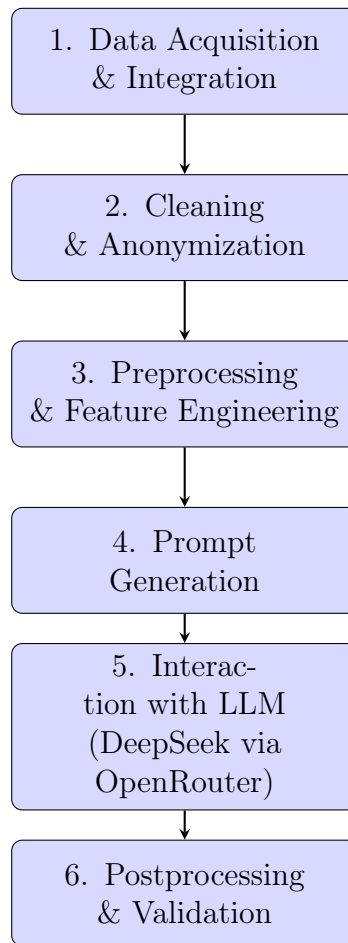


Figure 4.1: Schematic overview of the operational pipeline.

4.2 Output analysis

4.2.1 Correct vs. incorrect cases

To evaluate the reliability of the language model’s outputs, a manual review was conducted on a representative subset of records, selected across various years and clinical profiles. On these representative results, the estimates were compared to the underlying clinical information by an expert clinician to determine whether the inferred scores were reasonable, plausible, or clinically inaccurate.

During the obtaining of estimates, responses were classified as incorrect if they met any of the following conditions:

- At least one of the estimated domains was missing or left empty;
- One or more values fell outside the expected numerical range for their respective domains (i.e., values not in $\{0, 0.5, 1\}$);
- The response did not follow the expected JSON format, making automatic parsing or interpretation impossible.

Conversely, responses were considered correct if all fields were present and syntactically valid.

Overall, the majority of predictions were considered clinically acceptable. The model performed well in domains where explicit cues were present in the input (e.g., cohabitation status, or functional impairments described in physician or nurse notes).

Incorrect cases were typically associated with ambiguous or incomplete documentation in the original records, misinterpretations of negations, or excessive extrapolations by the model. The latter case was particularly prevalent where the textual information did not contain any insight into one of the expected domains.

These findings suggest that the model can perform surprisingly well given rich context, but remains very sensitive to the clarity, explicitness, and completeness of the input information.

4.2.2 Error classification

To better characterize the nature of the model’s inaccuracies, we may categorize identified errors into the following classes:

- **Hallucinations:** The model occasionally invented medical details not present in the prompt, such as stating that a patient was institutionalized or taking medications that were never listed. These errors highlight the generative nature of LLMs and their tendency to fill in perceived gaps.
- **Omissions:** Some outputs failed to account for clearly present data. For example, despite not explicitly mentioning any ongoing therapy, the corresponding score was incorrectly set to a value higher than 0. This points to weaknesses in attention or reasoning over long inputs.
- **Semantic misunderstandings:** The model sometimes misinterpreted language, such as interpreting "mild difficulty walking" as full independence, or confusing negated statements as affirmatives. These are more subtle errors that affect clinical interpretation and fall in line with observed limitations in the behavior of LLMs with reference to negations in the prompts[18].
- **Format inconsistencies:** Some outputs failed to follow the strict JSON structure or provided qualitative instead of numeric values. These required post-processing corrections or deletion.
- **Boundary Violation:** The LLM sometimes provided scores outside the accepted range (e.g., 0.3 instead of 0, 0.5, or 1).

Understanding these error types and their causes is essential for guiding future prompt refinement, model selection, and possible post-processing correction pipelines. It also reinforces the importance of human-in-the-loop validation in clinical applications.

4.3 Encountered issues

The use of a general-purpose language model for estimating the Brief Multidimensional Prognostic Index led to several critical challenges. These issues reflect both intrinsic limitations of current LLMs in handling clinical reasoning and the complexity of interpreting heterogeneous, unstructured, real-world hospital data. We classified the encountered problems into three broad categories: hallucinations, prompt limitations, and input data quality. Each of these impacted model performance, reliability, and interpretability.

4.3.1 Hallucinations: examples and possible causes

A *hallucination* occurs when the LLM produces content that is not grounded in the input data. These hallucinations may be plausible-sounding, but are clinically invalid and potentially misleading.

- **Medication misestimation:** In one case, the patient (a 67-year-old man) had no listed therapies in the provided records. Nonetheless, the LLM assigned a score of 0.5 (corresponding to 4–6 medications per day); upon further inspection and research, it seems plausible to assume this estimate was given due to 5+ medications per day being the average for 65+ year old men in Italy[3]. This illustrates an inference based on population-level heuristics, not case-specific evidence.
- **Cognitive status misinterpretation (SPMSQ):** In several instances, the model assigned a perfect cognitive score (1.0) despite ambiguous or absent mention of orientation or memory in the medical notes. Conversely, in other cases, vague references to confusion led to extremely poor scores (0), even when the episode was transient or explained by medication.
- **Comorbidity inflation:** The LLM often inferred the presence of comorbidities based solely on the number of diagnoses, even if the listed conditions were interrelated or non-severe.
- **Bias in guessing cohabitation status:** In the case of missing cohabitation data, the model often assumed family living, rarely selecting either of the other two alternatives (see Figure 5.2). This suggests an implicit prior built from earlier prompts or generalized assumptions about elderly care.

Following further analysis and discussions among the research team, the most likely causes of this behavior were identified as:

- **Incomplete prompts:** Lack of information in key fields often forced the model to fill in gaps with assumptions. This inevitably led to inconsistencies.
- **Overgeneralization:** The LLM tends to apply statistical priors — i.e., what is “typical”, even when individual data suggest otherwise.
- **Training limitations:** The model was not trained specifically for structured clinical estimation tasks and lacks a mechanism to distinguish between “not enough information” and “low score”.

- **Population-level heuristics:** When domain-specific information was missing, the LLM often defaulted to assumptions based on general population trends (e.g., age-related averages), rather than assuming no issues were present.

This latter aspect was a particularly relevant point of discussion with our collaborating expert clinician, as it was deemed critical to establish whether the LLM would behave consistently in the face of missing information (e.g., always overestimating or always underestimating) or not.

4.3.2 Prompt constraints and their effect

The prompt was carefully engineered to reduce ambiguity by:

- Presenting the patient data in a structured bullet-point format;
- Explicitly requesting only numerical values in JSON format;
- Defining the Score and semantics for each domain.

Despite these constraints, the LLM occasionally deviated:

- Returned explanatory text along with JSON (e.g., “Here is the result: ...”);
- Assigned invalid values (e.g., 0.3) despite the specification;
- Left fields blank when unsure.

Interestingly, making the prompt longer (with more context) did not improve accuracy. Rather, prompts overloaded with information resulted in much more lackluster responses, with most of the domains not being assigned an estimate. A potential solution could involve breaking down the estimation into separate prompts per domain, though this would increase complexity and query costs notably.

4.3.3 Data errors (missing values, noise, misinterpretation)

Despite careful prompt design, limitations inherent to the original ED datasets, collected from real-world clinical workflows, posed significant challenges to accurate estimation:

- **Missing data:** Information regarding aspects such as nutritional status or cohabitation status was not always present.
- **Free-text ambiguity:** Clinical notes were not standardized and often included shorthand, inconsistent terminology, or vague statements.
- **Noise due to anonymization:** Personal references (e.g., physician names, institutional roles) were removed, sometimes breaking context in the notes.
- **Misinterpretation by LLM:** The model sometimes misunderstood negations and often assumed frailty in domains where no problems were explicitly pointed out, and thus where it would have been reasonable to assume no relevant problems existed.

Chapter 5

Discussion

5.1 Critical evaluation of results

The results obtained through the application of a general-purpose large language model to estimate the Brief MPI show both promise and significant limitations.

5.1.1 Distribution of Brief MPI estimates

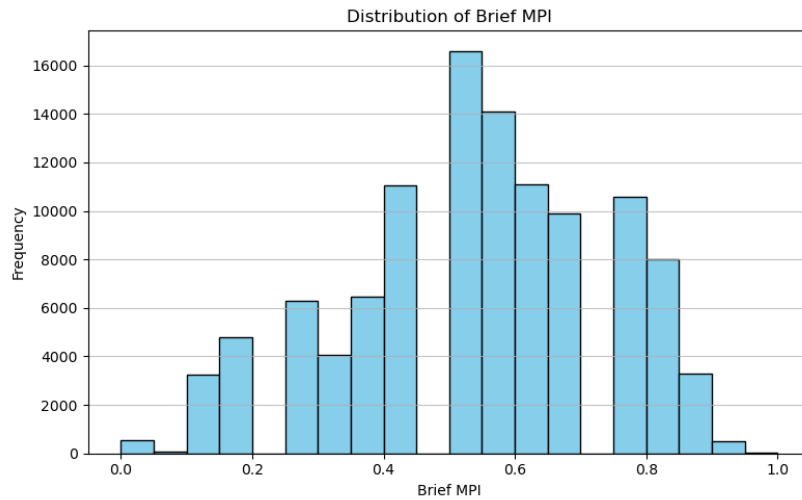


Figure 5.1: Distribution of LLM-estimated Brief MPI scores.

Figure 5.1 shows the distribution of the resulting Brief MPI estimates across all patient records. Most score values cluster around intermediate values, suggesting that the model often defaults to conservative mid-range

estimations. This may indicate an averaging heuristic in the absence of strong signals.

5.1.2 Distribution of domain estimates

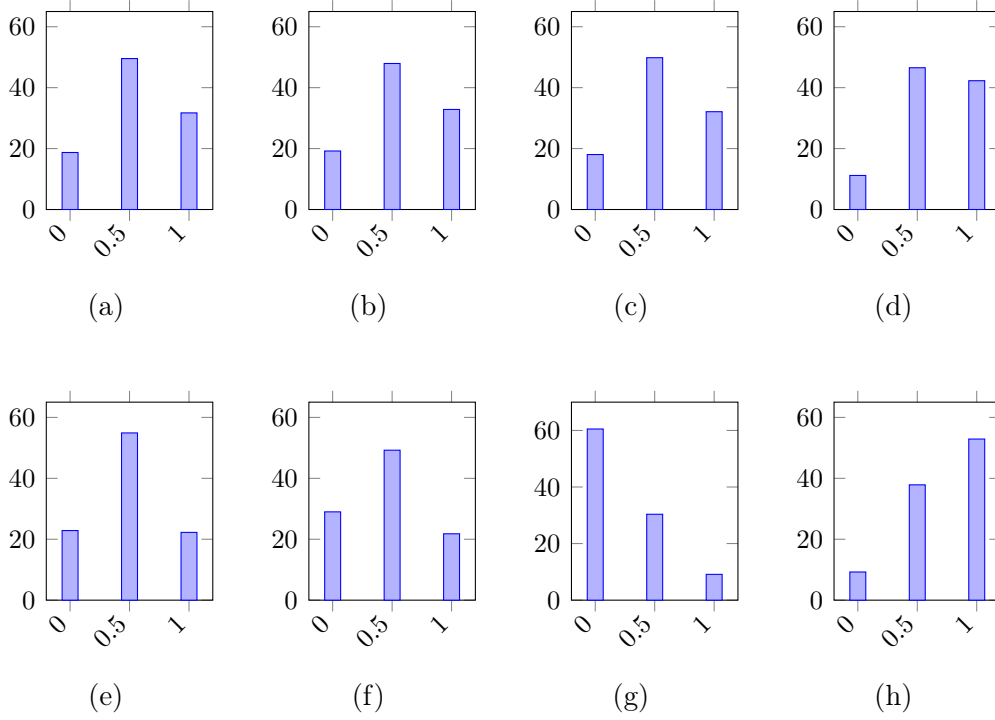


Figure 5.2: Distribution of scores across different domains (in percentage). Each subfigure represents a specific domain: (a) ADL, (b) IADL, (c) Mobility (Barthel), (d) SPMSQ, (e) MNA, (f) Medications, (g) Cohabitation, (h) Comorbidities. The x-axis indicates the score assigned (0, 0.5, or 1), and the y-axis shows the relative frequency of each score in the dataset.

Descriptive statistics reveal that certain domains (e.g., ADL and IADL) tend to be more uniformly distributed, while others such as Comorbidities and Medications show significant skewness toward higher impairment scores. This might reflect a model bias toward assuming greater frailty, particularly in the absence of contrary evidence.

5.1.3 Clinical plausibility

From a clinical perspective, many estimates appeared superficially plausible and, in some cases, aligned with expected patterns in geriatric populations.

Our clinical collaborator noted that the model often inferred higher frailty levels, which is not entirely unexpected given the age group under study. However, deeper inspection revealed that these assessments were not always well-grounded in the patient-specific information available in the notes. In particular, when documentation was sparse or ambiguous, the model sometimes produced responses that seemed disconnected from the clinical context. This suggests a tendency to generalize based on population-level priors rather than case-specific evidence.

5.1.4 LLM behavior and robustness

The Brief MPI estimation task exposes the fragility of current general-purpose LLMs in clinical reasoning tasks. While models like DeepSeek V3 show promising abilities in synthesizing data, they remain sensitive to prompt design and prone to bias or oversimplification. Domain-specific fine-tuning, more structured inputs, or ensemble querying strategies may be required to ensure safety and reliability in future applications.

5.2 Limitations of the approach

5.2.1 Model-related limitations

General-purpose LLMs are not specifically designed for structured clinical inference; they may apply broad statistical heuristics or prior knowledge when specific evidence is lacking, and cannot inherently recognize when information is insufficient to make a safe judgment. Their tendency to not express uncertainty is particularly problematic in high-stakes clinical contexts.

5.2.2 Prompt-related limitations

Although prompt engineering significantly constrained output format and semantics, it could not prevent the model from introducing invalid values, hallucinations, or omissions. Attempts to increase prompt context did not consistently improve performance, and in some cases led to lacking responses. Alternative strategies such as single-domain prompting or modular reasoning were not explored but may be beneficial.

5.2.3 Data quality and preprocessing constraints

The underlying emergency department dataset introduced considerable noise and ambiguity. Missing or generic entries, shorthand or unstructured free-

text notes, and context loss due to anonymization all likely contributed to LLM misinterpretation. These data limitations undermine model performance and reliability regardless of prompt structure.

5.2.4 Evaluation challenges

Model outputs were evaluated qualitatively, based on plausibility and structural correctness, rather than against a known ground truth. The absence of true Brief MPI scores for the dataset prevented a quantitative assessment. Furthermore, human review was very limited and lacked formal inter-rater reliability assessment.

Chapter 6

Conclusions

6.1 Summary of the results

This study explored the use of large language models for estimating the Brief MPI based on unstructured clinical notes. Quantitative and qualitative analyses revealed that current general-purpose models demonstrate a limited yet intriguing capacity for clinical inference. Models such as DeepSeek V3 exhibited the most promising performance, particularly in synthesizing complex patient information into structured outputs. However, even the best-performing models showed sensitivity to prompt phrasing, a tendency toward oversimplification, and variable alignment with the true clinical context.

Although performance varied across dimensions of the Brief MPI, domains such as comorbidity, nutrition, and number of medications appeared particularly challenging for the models, either due to often lacking explicit data for those domains or due to the LLM’s misinterpretation of free-text information. Plausibility ratings by human review often revealed surface-level coherence, but deeper inspection highlighted a reliance on general assumptions rather than patient-specific evidence.

6.2 Final considerations

The Brief MPI estimation task exposes the fragility of current LLMs in performing nuanced clinical reasoning. While the use of LLMs for clinical decision support remains conceptually promising, this study highlights important limitations. General-purpose LLMs, even at their current level of sophistication, lack the precision and reliability necessary for safe deployment in real-world geriatric care settings.

Nonetheless, the results underscore the feasibility of using LLMs as as-

sistive tools in clinical workflows, especially if appropriately constrained, guided, or fine-tuned. A recurring challenge is the balance between leveraging their generative capacity and mitigating their propensity to fabricate or misinterpret medical information. This work reaffirms the necessity for rigorous validation, clinical oversight, and domain adaptation in any future LLM-based applications in healthcare.

6.3 Suggestions for future developments

Improving the reliability and clinical applicability of LLMs in tasks like Brief MPI estimation will require several enhancements:

- **Domain-specific fine-tuning:** Training or adapting models on curated geriatric datasets can improve contextual relevance and reduce hallucination. Some LLMs trained specifically in the context of healthcare, such as MedGemma[2], BioBERT[10] and ClinicalBERT[6], may provide interesting results.
- **Structured input formats:** Providing models with semi-structured summaries or templates derived from electronic health recording systems could anchor responses more firmly in the patient data.
- **Ensemble prompting strategies:** Querying multiple models or using chain-of-thought prompting may help surface uncertainty and reduce overconfidence in single-point predictions.
- **Transparent uncertainty handling:** Future systems should express confidence estimates and indicate when input data is insufficient to support reliable inference.
- **Clinical validation:** Continuous evaluation through clinician-in-the-loop studies is essential to align model outputs with real-world expectations and safety requirements.
- **Testing the estimates by using them as predictors:** To better assess the clinical utility of the estimates, it could be useful to evaluate their predictive power using outcomes for which reliable ground truth data exist, such as patient mortality or readmission rates. Such validation would provide more objective metrics of model performance and support the safe integration of LLM-based tools into clinical decision-making processes.

In summary, while general-purpose LLMs hold potential, substantial technical and methodological advances are necessary before they can be entrusted with tasks as sensitive as prognostic scoring in vulnerable populations. Further studies could improve on our work by more extensively testing different approaches to prompt engineering, utilizing LLMs specifically trained for the healthcare context, and collaborating with a larger number of domain experts to expand and improve the human evaluation of the results.

Appendix A

Examples of Prompts and Responses

Example Prompt for Patient Record Inference

The following is an illustrative example of a prompt used for inferring Brief MPI index values for a patient's emergency department admission. The patient's sensitive information has been anonymized, and only the data deemed relevant for the estimation of the Brief MPI index domains has been extracted and included. This structured input provides the DeepSeek LLM with the necessary clinical context to generate the required numerical estimations for the eight domains of the Brief MPI for the patient. This example demonstrates the "one-shot prompting" approach, where all pertinent patient details and the specific estimation request, along with the desired output format and value scales, are encapsulated within a single prompt.

Paziente con Codice di Servizio 2022_000002:

- Eta': 72
- Sesso: F
- Note degli infermieri: Trauma cranico in NAO con ematoma palpebrale destro + emicostato anteriore destro. Dubbia PDC, amnesia dell'evento, no neurodeficit H 1.15, Rivalutata: assopita, risvegliabile al minimo stimolo verbale. Vario codice (CAF).
- Note dei medici: Vive in famiglia. Autonoma. Non allergie note. Dislipidemia, precedenti di fibrillazione atriale parossistica. In tp con xarelto 20 la sera alle 19 (assunto oggi), almarytm 1/2 cpx2, crestor 10mg a di' alterni. Questa sera dopo cena riferita dispepsia, in seguito riferisce di

- essersi trovata sulla poltrona davanti la televisione, con gli occhiali lontano e dolore al torace ed al capo, con ematoma dell'occhio dx ed escoriazione del gomito dx; telefonava quindi alla figlia e poi veniva accompagnata in PS dal 118. EO: vigile ma rallentata, persiste amnesia per l'evento. Non neurodeficit focali, azione ritmica con S2 vibrato. Eseguita dopo la TC ricoagulazione con complesso protrombinico a 25 UI/Kg. *segue in Varie
- Diagnosi: TRAUMA CRANIO-FACCIALE FRONTO-ORBITARIO DX COMPLICATO DA NOTE DI ESA. FRATTURA COMPOSTA ZIGOMO DX.
 - Diagnosi principale: EMORRAGIA SUBARACNOIDEA CONSECUTIVA A TRAUMATISMO SENZA MENZIONE DI FERITA INTRACRANICA ESPOSTA, STATO DI COSCIENZA NON SPECIFICATO
 - Diagnosi secondarie: CONTUSIONE DELLA FACCIA, DEL CUIOIO CAPELLUTO E DEL COLLO ESCLUSO L'OCCHIO; FRATTURA CHIUSA DI ALTRE OSSA DELLA FACCIA; USO A LUNGO TERMINE (IN CORSO) DI ANTICOAGULANTI; Informazione non disponibile; Informazione non disponibile
 - Suggerimenti: Informazione non disponibile
 - Varie: *segue da Dati Clinici. Durante l'osservazione in Medicina d'Urgenza eseguito monitoraggio ECGrafico con esclusione di eventi aritmici, valutazione neurologica e TC encefalo di controllo. Dal punto di vista sintomatologico ed obiettivo non sintomi o segni evolutivi neurologici. Ricovero per il completamento della osservazione e organizzazione dei controlli a breve.
 - Esami di laboratorio: Informazione non disponibile
 - Esami di radiologia: RX SCHELETRO COSTALE BILATERALE; RX TORACE ; TC CRANIO/ENCEFALO; RX BACINO; TC CRANIO/ENCEFALO
 - Terapie: KEDCOM*FL 500UI+FL 20ML+SET [1] 1500 UI

Basandoti su queste informazioni, puoi stimare i seguenti parametri per il paziente?

- ADL (Mangiare, Vestirsi, Controllo)
- IADL (Telefono, Farmaci, Acquisti)
- Barthel (Alzarsi, Camminare, Scale)
- SPMSQ (Data, Anni, Calcolo)
- MNA (BMI, Perdita Appetito, Perdita Peso)
- Comorb
- Farmaci
- Stato Coabitativo

Per ogni parametro, vorrei ottenere solo il valore numerico finale.

In particolare: per ADL, IADL e Barthel un valore tra 0, 0.5 e 1 dove 0 indica poca difficolt  e 1 molta difficolt ;

per SPMSQ un valore tra 0, 0.5 e 1 dove 0 indica molta difficolt  e 1 nessuna;

per MNA un valore tra 0, 0.5 e 1 dove 0 indica nessuna difficolt  e 1 molta;

per Comorb 0 se il numero di comorbidit  e' 0, 0.5 se e' minore di 3, 1 se maggiore o uguale a 3;

per Farmaci 0 se il numero di farmaci nella terapia attuale e' fra 0 e 3 inclusi, 0.5 se e' minore di 7, 1 se maggiore o uguale a 7;

per Stato Coabitativo, 1 se vive da solo, 0 se con la famiglia, 0.5 se in un istituto.

La risposta dovrebbe essere semplicemente un oggetto JSON con i campi Codice di Servizio, ADL, IADL, Barthel, SPMSQ, MNA, Comorb, Farmaci, Stato Coabitativo.

Example DeepSeek V3 Output

The following is the estimate provided by the DeepSeek V3 LLM in response to the prompt shown above.

```
{
  "Codice di Servizio": "2022_000002",
  "ADL": 0,
  "IADL": 0,
  "Barthel": 0,
  "SPMSQ": 0.5,
  "MNA": 0,
  "Comorb": 0.5,
  "Farmaci": 0.5,
  "Stato Coabitativo": 0
}
```

Example Gemini 2.0 Flash Output

The following is the response provided by the Gemini 2.0 Flash model to the same prompt, showcasing its overall tendency to provide responses full of

verbose text despite our requests to only provide a JSON object containing the estimates.

To estimate the requested parameters, I will analyze the provided patient information based on the typical interpretations of these scales and definitions.

Parameter Estimation

Here's the breakdown of how each parameter is estimated:

ADL (Activities of Daily Living): The patient is described as "Autonoma" (autonomous) and able to perform activities like telephoning her daughter and being accompanied to the PS. However, she is "vigile ma rallentata" (awake but slowed) and has "amnesia dell'evento" (amnesia of the event), suggesting some temporary impairment in her daily functioning. Given the instructions where 0 is "poca difficoltà" (little difficulty) and 1 is "molta difficoltà" (much difficulty), and considering the acute trauma and current slowed state, there's likely some impact, but not total dependence. A value of 0.5 seems appropriate.

IADL (Instrumental Activities of Daily Living): The patient was able to telephone her daughter, which is an IADL. While her current state is "rallentata" and she has amnesia, the fact she initiated a call suggests a baseline ability. However, the acute event and current hospital stay would limit her ability to perform other IADLs like managing medications or shopping independently. A value of 0.5 reflects this partial difficulty/dependence in the current acute phase.

Barthel (Mobility and Self-Care): Similar to ADL, the patient is "vigile ma rallentata" and had a trauma. While she was able to get on the sofa, the notes mention "ematoma palpebrale destro + emicostato anteriore destro" and "escoriazione del gomito dx", indicating some physical injury that would likely impede full independence in mobility and self-care tasks like dressing or getting up/walking without some difficulty in the acute phase. A value of 0.5 is chosen.

SPMSQ (Short Portable Mental Status Questionnaire): This scale assesses cognitive function (Date, Age, Calculation). The patient has "amnesia dell'evento" and is described as "

assopita, risvegliabile al minimo stimolo verbale" and "rallentata." This points to a cognitive impairment, even if temporary. Given the scoring where 0 indicates "molta difficoltà" (much difficulty) and 1 indicates "nessuna" (none), a value of 0.5 (indicating some difficulty but not severe) is reasonable, as she is responsive to verbal stimuli.

MNA (Mini Nutritional Assessment): This includes BMI, appetite loss, and weight loss. There is no information provided regarding her BMI, appetite, or weight loss. Therefore, it's not possible to estimate this parameter reliably; the value should be null.

Comorb (Comorbidities): The patient has a history of "Dislipidemia" and "fibrillazione atriale parossistica." This accounts for two comorbidities.
0 if the number of comorbidities is 0
0.5 if it is less than 3
1 if greater than or equal to 3 Since the number of comorbidities (2) is less than 3, the value is 0.5.

Farmaci (Medications): The patient is currently on "xarelto 20 la sera alle 19 (assunto oggi)," "almaryl 1/2 cpx2," and "crestor 10mg a di' alterni." This totals three medications.
0 if the number of drugs in the current therapy is between 0 and 3 inclusive
0.5 if it is less than 7
1 if greater than or equal to 7 Since the number of medications (3) is between 0 and 3 inclusive, the value is 0.

Stato Coabitativo (Living Arrangement): The notes state, "Vive in famiglia" (Lives with family).
1 if lives alone
0 if with family
0.5 if in an institution Therefore, the value is 0.

```
{  
  "Codice di Servizio": "2022_000002",  
  "ADL": 0.5,  
  "IADL": 0.5,  
  "Barthel": 0.5,  
  "SPMSQ": 0.5,
```

```
"MNA": null,  
"Comorb": 0.5,  
"Farmaci": 0,  
"Stato Coabitativo": 0  
}
```

Appendix B

Brief MPI Indices and Criteria

Domain	Tool / Variable	Values / Scoring
Activities of Daily Living (ADL)	Three items from the Katz ADL index (dressing, feeding, continence)	0 denotes low difficulty in the three activities considered, 0.5 medium difficulty, 1 high difficulty
Instrumental Activities of Daily Living (IADL)	Three items to evaluate independence in using the telephone, taking medications, and shopping	0 denotes low difficulty in the three activities considered, 0.5 medium difficulty, 1 high difficulty
Mobility	Barthel mobility scale, evaluated inquiring about the abilities to transfer from bed to chair or wheelchair, walking and going up and down the stairs	0 denotes low difficulty in the three activities considered, 0.5 medium difficulty, 1 high difficulty
Cognitive Status	Short Portable Mental Status Questionnaire (SPMSQ)	0 = poor cognitive state 0.5 = average cognitive state 1 = good cognitive state
Nutritional Status	Mini-Nutritional Assessment (MNA) short form	0 = good nutritional state 0.5 = average nutritional state 1 = poor nutritional state
Medications	Number of drugs (active substances) currently taken	$0 \Leftrightarrow n \leq 3$ $0.5 \Leftrightarrow 3 < n \leq 6$ $1 \Leftrightarrow n > 6$
Social Support	Cohabitation status (alone, with family, in institute)	0 = with family 0.5 = institutionalized 1 = alone
Comorbidities	Number of diseases requiring chronic therapies	$0 \Leftrightarrow n = 0$ $0.5 \Leftrightarrow 0 < n \leq 2$ $1 \Leftrightarrow n > 2$
Brief MPI	Brief MPI	Average of the eight fields above; hence a value in $[0,1]$
Risk	Based on Brief MPI value	$1 \Leftrightarrow MPI \leq 0.33$ $2 \Leftrightarrow 0.33 < MPI \leq 0.66$ $3 \Leftrightarrow MPI > 0.66$

Table B.1: Domains and scoring criteria used for Brief MPI fields.

Acknowledgments

I am profoundly thankful to my academic supervisors, Professors Danilo Montesi and Flavio Bertini, for their mentorship, patience, and expert advice. Their insights into data analysis, information retrieval, and the responsible application of AI models in healthcare were essential in shaping this thesis.

I would also like to express my sincere gratitude to Dr. Andrea Fabbri, Head Physician of the Emergency Department at Forlì Hospital, for his invaluable guidance and support throughout this project. His deep knowledge of emergency medicine and geriatric care, as well as his continuous feedback and great previous experience in research, greatly enriched both the clinical and methodological aspects of this work.

Finally, I would like to thank my family and friends for their continuous support during my studies. I have never taken for granted the sacrifices made by my parents to allow me to continue my studies, for which I am grateful from the bottom of my heart. Similarly, I want to thank my brother Alex and my sister Chiara for bringing me joy every day and inspiring me to try and become the best person I can be. I feel truly privileged to be able to call them family.

References

Bibliography

- [1] Alberto Cella, Nicola Veronese, Carlo Custodero, Alberto Castagna, Lisa A Cammalleri, Walter M Capitano, Luisa Solimando, Luca Caruba, Carlo Sabbà, Giovanni Ruotolo, Mario Barbagallo, and Alberto Pilotto. Validation of abbreviated form of the multidimensional prognostic index (mpi): The brief-mpi project. *Clinical Interventions in Aging*, Volume 17:789–796, May 2022.
- [2] Google DeepMind. Medgemma. <https://medgemma.org/>, 2025. [Online; accessed 01-June-2025].
- [3] Agenzia Italiana del Farmaco. L’uso dei farmaci in italia - rapporto nazionale anno 2023. <https://www.aifa.gov.it/-/1-uso-dei-farmaci-in-italia-rapporto-osmed-2023>, 2024. [Online; accessed 04-April-2025].
- [4] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37, Mar. 1996.
- [5] Jiawei Han, Jian Pei, and Hanghang Tong. *Data Mining: Concepts and Techniques, Fourth Edition*. Elsevier, 2024.
- [6] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

- [7] M.J. Kaiser, J.M. Bauer, C. Ramsch, W. Uter, Y. Guigoz, T. Cederholm, D.R. Thomas, P. Anthony, K.E. Charlton, M. Maggio, A.C. Tsai, D. Grathwohl, B. Vellas, and C.C. Sieber. Validation of the mini nutritional assessment short-form (mna-sf): A practical tool for identification of nutritional status. *The Journal of nutrition, health and aging*, 13(9):782–788, November 2009.
- [8] S. Katz, T. D. Downs, H. R. Cash, and R. C. Grotz. Progress in development of the index of adl. *The Gerontologist*, 10(1 Part 1):20–30, March 1970.
- [9] M. P. Lawton and E. M. Brody. Assessment of older people: Self-maintaining and instrumental activities of daily living. *The Gerontologist*, 9(3 Part 1):179–186, September 1969.
- [10] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.
- [11] Bernard S. Linn, Margaret W. Linn, and Lee Gurel. Cumulative illness rating scale. *Journal of the American Geriatrics Society*, 16(5):622–626, May 1968.
- [12] Eric Pfeiffer. A short portable mental status questionnaire for the assessment of organic brain deficit in elderly patients. *Journal of the American Geriatrics Society*, 23(10):433–441, October 1975.
- [13] Alberto Pilotto, Luigi Ferrucci, Marilisa Franceschi, Luigi P. D’Ambrosio, Carlo Scarcelli, Leandro Cascavilla, Francesco Paris, Giuliana Placentino, Davide Seripa, Bruno Dallapiccola, and Gioacchino Leandro. Development and validation of a multidimensional prognostic index for one-year mortality from comprehensive geriatric assessment in hospitalized older patients. *Rejuvenation Research*, 11(1):151–161, February 2008.
- [14] Alberto Pilotto, Nicola Veronese, Katerin Leslie Quispe Guerrero, Sabrina Zora, An L.D. Boone, Matteo Puntoni, Angela Giorgeschi, Alberto Cella, Ines Rey Hidalgo, Yves-Marie Pers, Alberto Ferri, Jose Ramon Hevia Fernandez, Marta Pisano Gonzalez, Marta Pisano Gonzalez, An L.D. Boone, Inés Rey, Raquel Ochoa Gonzalez, J.R. Hevia Fernandez, Y.M. Pers, Christian Jorgensen, Valsecchi Verushka, Rosanna Ferreira, Adrien Durand, Cristina Balaguer, Céline Cardoso, Manuel

- Ramos, Graham Baker, Danni Brown, Suzanne Lucas, S.S. Tan, X. Zhang, P. de Vries, H. Raat, Alberto Pilotto, Sabrina Zora, Alberto Ferri, Alberto Cella, Alessandra Argusti, Bárbara Branchini, and Jorge Garcés. Development and validation of a self-administered multi-dimensional prognostic index to predict negative health outcomes in community-dwelling persons. *Rejuvenation Research*, 22(4):299–305, August 2019.
- [15] Alberto Pilotto, Nicola Veronese, Giacomo Siri, Stefania Bandinelli, Toshiko Tanaka, Alberto Cella, and Luigi Ferrucci. Association between the multidimensional prognostic index and mortality during 15 years of follow-up in the inchianti study. *The Journals of Gerontology: Series A*, 76(9):1678–1685, September 2020.
- [16] K. Rockwood. A global clinical measure of fitness and frailty in elderly people. *Canadian Medical Association Journal*, 173(5):489–495, August 2005.
- [17] Vergil N. Slee. The international classification of diseases: Ninth revision (icd-9). *Annals of Internal Medicine*, 88(3):424–426, March 1978.
- [18] Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. Language models are not naysayers: An analysis of language models on negation benchmarks, 2023.
- [19] Gill Turner and Andrew Clegg. Best practice guidelines for the management of frailty: a british geriatrics society, age uk and royal college of general practitioners report. *Age and Ageing*, 43(6):744–747, November 2014.