Scientific Reports

https://doi.org/10.1038/s41598-026-56721-0

Article in Press

Automatic identification of diagnosis from hospital discharge letters via weakly supervised Natural Language Processing

Received: 23 December 2025

Accepted: 2 June 2026

Published online: 11 June 2026

Cite this article as: Torri V., Barbieri E., Cantarutti A. *et al.* Automatic identification of diagnosis from hospital discharge letters via weakly supervised Natural Language Processing. *Sci Rep* (2026). https://doi.org/10.1038/s41598-026-56721-0

Vittorio Torri, Elisa Barbieri, Anna Cantarutti, Carlo Giaquinto & Francesca Ieva

Automatic identification of diagnosis from hospital discharge letters via weakly supervised Natural Language Processing

Vittorio Torri[1,*], Elisa Barbieri[2], Anna Cantarutti[3,4], Carlo Giaquinto[2], and Francesca Ieva[1,4,5]

[1]MOX - Modelling and Scientific Computing Lab, Department of Mathematics, Politecnico di Milano, Milano, 20133, Italy
[2]Division of Pediatric Infectious Diseases, Department for Woman and Child Health, University of Padova, Padova, 35122, Italy
[3]Unit of Biostatistics, Epidemiology and Public Health, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, 20126, Italy
[4]National Centre for Healthcare Research and Pharmacoepidemiology, Department of Statistics and Quantitative Methods, University of Milano-Bicocca, Milano, 20126, Italy
[5]HDS - Health Data Science Centre, Human Technopole, Milano, 20157, Italy
[*]vittorio.torri@polimi.it

ABSTRACT

Identifying patient diagnoses from hospital discharge letters is essential for large-scale cohort selection and epidemiological research, but traditional supervised approaches require extensive manual annotation, which is often impractical for large textual datasets. We present a weakly supervised Natural Language Processing (NLP) pipeline for classifying Italian discharge letters without document-level manual annotation. The method extracts diagnosis-related sentences, generates semantic embeddings using a transformer model further pre-trained on Italian medical documents, and applies a two-level clustering procedure to derive weak labels that are then used to train a document-level classifier.

The approach was evaluated in a case study on bronchiolitis using 33,176 discharge letters of children admitted to 44 emergency rooms or hospitals in the Veneto Region, Italy, between 2017 and 2020. The best weakly supervised model achieved an AUROC of 77.68% ($\pm$4.30%), an AUPRC of 73.13% ($\pm$4.93%), and an F1-score of 78.14% ($\pm$4.89%) against manually annotated data. Performance surpassed unsupervised baselines and approached fully supervised models, while reducing the need for manual annotation by more than 1,500 hours for a dataset of this size. Similar model rankings were observed in a secondary validation on a smaller bronchitis dataset (3,188 discharge letters, 2020-2025), where the best weakly supervised model achieved an AUPRC of 76.72% ($\pm$5.02%).

These results suggest the potential of weakly supervised NLP methods for scalable disease identification from clinical discharge letters.

Introduction

Electronic Health Records (EHR) contain a vast and continuously growing amount of patient information, but a significant portion of this data remains unstructured in formats like clinical notes, reports, discharge letters and referrals. Despite containing valuable information, this unstructured data is often underutilized in the downstream analysis due to their complexity and heterogeneity[1,2]. While Natural Language Processing (NLP) has made significant advancements in extracting and classifying information from text[3,4], applying these techniques to clinical narratives remain challenging. Medical documents are characterized by a highly specialized lexicon, with ambiguous abbreviations and frequent misspellings, a lack of standard templates across hospitals and a limited amount of manually labelled data[5–7]. Additionally, most NLP tools are optimized for high-resource languages such as English, limiting their transferability to languages like Italian, where datasets and domain-specific linguistic resources are scarce[8].

Among the various types of clinical documents, hospital discharge letters play a particularly important role: they summarize the patient's condition, diagnoses, and treatments, representing a key source of diagnostic information for both clinical follow-up and population-based studies. Although international standard such as the ICD-CM[9] exists for encoding diagnoses, physicians often prefer to provide textual descriptions of their diagnoses because they are faster to write, more expressive, and often the only option available in document templates[10,11]. Even when ICD-CM codes are included, previous research has shown that

they can be incomplete or inaccurate[12–14]. Automatic methods capable of identifying diagnoses directly from unstructured discharge letters are therefore crucial for cohort selection and epidemiological research[15], reducing the need for manual review, which is labor-intensive and impractical at scale.

In Italy, every hospital or emergency room (ER) admission generates a discharge letter that should contain at least one diagnosis describing the reason for admission or discharge. However, these letters vary widely in format and terminology, often mixing narrative and structured components. Developing robust NLP systems for such data is challenging: traditional approaches rely on supervised classification models that require a substantial amount of manually annotated data[16], an effort that is costly and time-consuming for clinicians. This highlights a significant gap in the field: the need for methods that can effectively exploit large, unannotated datasets to extract clinically meaningful information without the burden of manual labelling.

Previous studies applying NLP to Italian clinical texts have not addressed discharge letters. Lanera et al.[17,18] worked on supervised models to identify varicella cases from notes of family pediatricians, while Sciannameo et al.[19] used a BERT-based model to predict rehospitalizations for elderly patients from medication prescriptions and hospitalization records. Hammami et al.[20] classified pathology reports for cancer morphology using a rule-based approach, and Viani et al.[21,22] extracted entities from cardiology reports through ontologies and recurrent neural networks. Most international work on discharge letters has been conducted in English using resources such as the MIMIC database[23,24] (see[25–28] for some examples), but research has also been conducted in other European languages with limited resources, such as Bulgarian[29,30], Dutch[31,32], Czech[33] and Swedish[34]. These studies, however, all relied on large annotated datasets, making them difficult to reproduce in settings where manual labelling is infeasible.

Recent work has explored weakly supervised approaches to mitigate the need for large annotated clinical corpora. Many of these methods rely on task-specific heuristics, such as disease-related keywords, regular expressions, or the transfer of labels from structured data like diagnostic codes or medication records[35–40]. In parallel, the EHR phenotyping literature has proposed frameworks that combine surrogate signals extracted from structured and textual data to infer patient-level disease status, including approaches such as PheNorm[41], sureLDA[42], Anchor[43], Polar[44] and XPRESS[45]. These methods typically aggregate information across multiple encounters and rely on structured EHR variables (e.g., diagnosis codes, medications, laboratory tests or concept counts extracted from notes) to construct surrogate features for phenotyping. While effective in large longitudinal EHR datasets, such assumptions are not always satisfied in practice. In particular, structured clinical variables may be incomplete or unavailable, especially in emergency department admissions that do not lead to hospitalization, where discharge summaries often constitute the primary source of diagnostic information. In these settings, the task becomes one of document-level diagnosis identification directly from free-text clinical narratives, which remains comparatively less explored. More recently, large language models (LLMs) have also been used to support weak supervision pipelines[46]; however, in clinical settings, effective deployment often still relies on the integration of domain-specific knowledge[47] or the use of proprietary models[48].

To address these limitations, we propose a weakly supervised NLP pipeline for automatically identifying patient diagnoses directly from Italian hospital discharge letters, operating solely on free-text clinical documents without requiring structured EHR variables or patient-level aggregation. Our approach eliminates the dependency on document-level manual annotations and offers a more generalizable solution than previous rule-based methods.

The pipeline first identifies diagnosis-related sentences from a subset of documents and represents them through contextual embeddings generated by a transformer model, previously pre-trained with public data in Italian related to the medical domain. These embeddings are then grouped through a two-level clustering procedure, after which the resulting clusters are semantically summarized and mapped to diseases of interest to produce weak labels. The weakly labelled data are finally used to train a transformer-based classifier capable of recognizing those diseases directly from discharge letters. This design enables large-scale, disease-agnostic classification of clinical narratives while avoiding the costs of large-scale document-level manual labelling.

We evaluate the proposed framework through a real-world case study focused on bronchiolitis, one of the leading causes of hospitalizations (20 %), emergency department visits (18 %) and office visits (15 %) among young children and a major driver of seasonal pediatric respiratory epidemics[49]. The dataset includes 33,176 discharge letters of 15,196 pediatric patients admitted to 44 emergency rooms or hospitals across the Veneto Region (Italy) between January 2017 and December 2020, collected through the Pedianet network[50]. To the best of our knowledge, there are no prior studies addressing the identification of bronchiolitis directly from medical text, whereas existing works have relied on structured data for diagnosis prediction[51–53].

To assess the robustness and generalizability of the proposed pipeline, we additionally conduct a secondary validation study on bronchitis using another smaller dataset from the same Pedianet network. This dataset includes 3,188 discharge letters collected between January 2020 and May 2025, involving 2,364 children living in Veneto Region, accessing 31 different hospitals. Although substantially smaller than the bronchiolitis dataset, this additional cohort provides an independent validation setting to assess whether the proposed weakly supervised framework can generalize to a different respiratory condition and data

distribution.

In our evaluation, we test the weakly supervised pipeline against manually labelled data, considering different classification backbones and comparing its performance with both fully unsupervised and fully supervised approaches. The proposed method achieves competitive performance, outperforming other unsupervised techniques and showing only a limited gap compared with supervised classifiers. Its performance is robust to weak labels selection, and the analysis of the intermediate steps indicates a strong potential for generalization across diseases.

Methods

In this section, we describe in detail the pipeline we propose for the weakly supervised diagnosis identification. The pipeline, depicted in Figure 1, can be divided into three main steps:

0 **TPT** - Pre-Training of a Transformer-based model for Italian medical documents.

1 **SAL** - Semi-automatic labelling of the dataset.

2 **WLC** - Classification of discharge letters using the weakly-labelled dataset.

The next sections describe in details the three steps. A summary of the main notation used throughout these sections is reported in Supplementary Material B.

Pre-training of a Transformer-based model for Italian medical documents (TPT)

Statistical and machine learning models for textual data require a mathematical representation of text. Let $\mathscr{X}$ denote the space of textual documents. A text encoder can be formalized as a mapping $f_\theta : \mathscr{X} \to \mathbb{R}^d$, where f_θ is a parametric function with parameters θ and d is the embedding dimension. Classical vocabulary-based representations (e.g., bag-of-words or TF–IDF) produce sparse vectors in $\mathbb{R}^{|V|}$, where V denotes the corpus vocabulary. While effective in certain settings, these representations do not explicitly encode semantic similarity between distinct lexical forms.

Neural network-based embeddings address this limitation. Static word embeddings, such as Word2Vec[54], define a fixed mapping $w \mapsto v_w \in \mathbb{R}^d$ for each token w, where v_w is a dense, continuous vector, used for every occurrence of w, and $d \ll |V|$. In contrast, contextual embeddings compute representations that depend on the full input sequence. Transformer-based models[55], including BERT[56] and RoBERTa[57], define contextual encoders where the embedding of each token depends also on the surrounding tokens, capturing different meanings that can depend on the context.

Several transformer models have been specialized for the biomedical domain in English, such as BioBERT[58] and ClinicalBERT[59]. For Italian, however, most available transformer models are trained primarily on general-domain corpora[60], with limited exposure to biomedical or clinical terminology.

To improve domain adaptation, we perform additional pre-training on the European Clinical Case Corpus (E3C)[61], a multilingual medical dataset containing documents from journal articles, specialty-school admission tests, patient information leaflets, and medical-science thesis abstracts (a summary of its composition is provided in Supplementary Material A.3).

Let $\mathscr{C} = \bigcup_{\ell \in \mathscr{L}} \mathscr{C}_\ell$ denote the full E3C corpus, where $\mathscr{L}$ is the set of languages represented in the dataset and $\mathscr{C}_\ell$ denotes the subset of documents in language ℓ. Let $\ell = it$ denote Italian. Since the corpus contains documents in multiple languages, we translate all non-Italian subsets into Italian using the Google Translate API. Automatic translation may introduce lexical or syntactic noise. However, the translated documents are used only for masked language modeling during domain-adaptive pre-training. As a result, occasional translation errors are unlikely to systematically bias the downstream task, while the enlarged corpus helps expose the model to a broader range of medical terminology.

Formally, for $\ell \neq it$, we define $T_\ell : \mathscr{C}_\ell \to \mathscr{C}_{it}$ (Step 0.1 in Figure 1.a).

The resulting Italian medical corpus used for domain-adaptive pre-training is

$$\mathscr{C}^{it} = \mathscr{C}_{it} \cup \bigcup_{\ell \in \mathscr{L} \setminus \{it\}} T_\ell(\mathscr{C}_\ell). \tag{1}$$

Starting from a pre-trained transformer encoder f_{θ_0}, we perform additional masked language modeling (MLM) pre-training on $\mathscr{C}^{it}$. The objective function is

$$\mathscr{L}_{\text{MLM}}(\theta) = -\mathbb{E}_{(c,t)} \log P_\theta(w_t \mid c_{\setminus t}), \tag{2}$$

where w_t denotes a masked token in document c, and $c_{\setminus t}$ denotes the surrounding context.

The optimized parameters are obtained as $\hat{\theta} = \arg\min_\theta \mathscr{L}_{\text{MLM}}(\theta)$ (Step 0.2 in Figure 1 .a), resulting in the domain-adapted encoder $f_{\hat{\theta}} : \mathscr{X} \to \mathbb{R}^d$ where $d = 768$ for the BERT-base encoders that we use. This encoder is subsequently used both in the the semi-automatic labelling (SAL, Step 1) and in the weakly supervised classification (WLC, Step 2), which are described in the following sections.

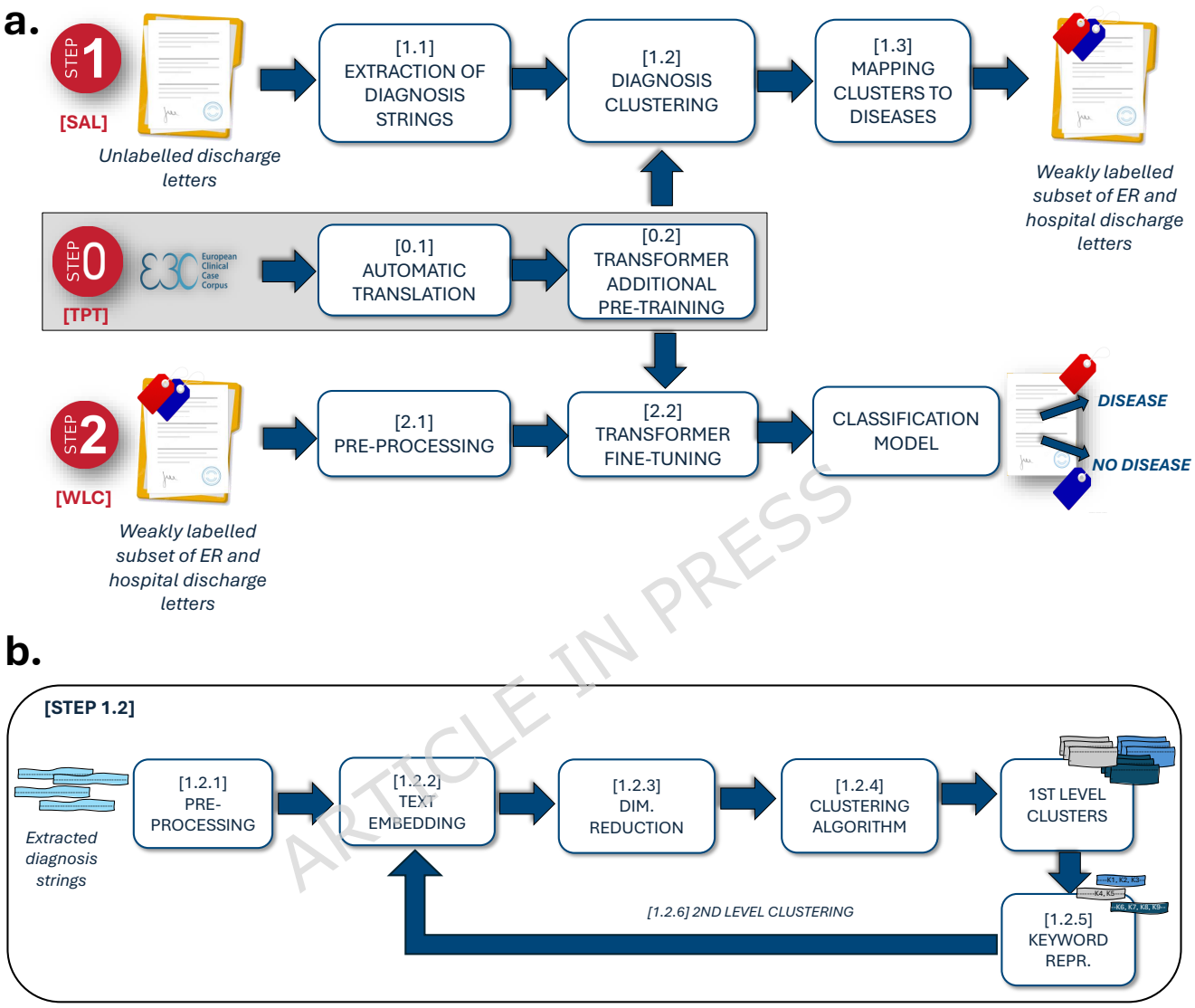

Figure 1. a. Diagram of the full pipeline for weakly supervised classification of discharge letters. Step 0 (TPT) consists in additional pre-training of a transformer-based model on a dataset of medical documents in Italian, including documents auomatically translated into Italian. Step 1 (SAL) is the core of the pipeline, consisting in the automatic extraction of diagnosis strings from the letters, their clustering (exploiting the model from TPT) and the mapping of the clusters to the disease labels. This procedure leads to a weakly-labelled subset of data that is used in Step 2 (WLC) to train a weakly supervised classificaton model. b. Detailed diagram of the clustering step for diagnosis strings, corresponding to Step 1.2 of the full pipeline. The extracted diagnoses are embedded with a transformer-based model, and, after a dimensionality reduction step, clustered. Clusters are summarized with relevant keywords and the clustering is repeated on these clusters' representations.

Semi-automatic labelling pipeline (SAL)

Developing a weakly supervised classification pipeline for discharge letters requires a procedure to derive weak-labels for at least a subset of data. Weak-labels are labels that can be automatically associated to the data but that might contain some noise. This section presents a methodology to construct these weak labels by extracting diagnosis strings from discharge letters and clustering them. This allows the creation of groups of diagnoses related to the same disease or symptoms that can be exploited as labels. This step corresponds to Step 1 (SAL) of Figure 1.a.

Our approach combines regular expressions for extracting diagnosis strings and a customized clustering pipeline based on a transformer model. In the following sections, we describe in detail these sub-steps.

Extraction of diagnosis strings

The first step (1.1 in Figure 1.a) consists of extracting diagnosis strings from discharge letters. Let $\mathscr{D} = \{x_i\}_{i=1}^{N}$ denote the set of discharge letters, where each $x_i \in \mathscr{X}$ is a textual document. Italian discharge letters do not follow a standardized template, as each hospital has its own, and even different departments within the same hospital may personalize it further. Nevertheless, we have observed that these templates often include a designated section for diagnoses, where physicians report the conditions related to the current hospitalization. This section may contain one or multiple diagnosis entries (e.g., primary and secondary diagnoses referring to the present admission), which typically contain the information required to determine the document label.

We formalize the extraction step as a deterministic function $g : \mathscr{X} \to \mathscr{S} \cup \{\emptyset\}$, where $\mathscr{S}$ denotes the space of diagnosis strings. For each document x_i, the output is $s_i = g(x_i)$, where $s_i \in \mathscr{S}$ if a diagnosis section is detected and $\emptyset$ otherwise. Three main challenges arise in handling diagnosis sections:

1. Not all discharge summaries include this indication, as it may be absent or left empty.
2. Different templates use varying terminologies and position this indication differently in the text.
3. Diagnosis strings are typically written as free text, with the same disease described by different strings and levels of detail

The first point is the primary reason why our pipeline focuses on classifying the text of a discharge letter rather than relying solely on the extracted diagnosis strings. To address the second point, the function g is implemented using a predefined set of regular expressions that identify and extract the diagnosis section across heterogeneous templates. Regular expressions provide a flexible and reproducible pattern-matching mechanism for detecting structured fields with limited syntactic variation. The full set of patterns is reported in Supplementary Material D.

After this step, for the majority of the letters we obtain a free-text diagnosis string describing the conditions associated with the current hospitalization. This extraction step focuses on diagnoses explicitly reported in the structured diagnosis section and does not aim to capture conditions mentioned exclusively in other narrative sections (e.g., past medical history). These strings are not directly used as labels due to their heterogeneity (point 3 above); instead, they are clustered, as detailed in the following section.

Diagnosis Clustering

Once the diagnosis strings are extracted, we cluster them to identify semantically coherent groups corresponding to diseases or symptoms . This corresponds to Step 1.2 in panel a of Figure 1 and is expanded in panel b of the same figure.

Let $I = \{i \in \{1,\dots,N\} : s_i \neq \emptyset\}$ denote the subset of documents for which a diagnosis string was successfully extracted. In this clustering step, we consider only the set $\{s_i\}_{i\in I}$, ignoring documents with $s_i = \emptyset$.

Each diagnosis string undergoes minor normalization, including removal of punctuation, dates, and time expressions, through a deterministic pre-processing function $p_s : \mathscr{S} \to \mathscr{S}$. The normalized string is denoted $\bar{s}_i = p_s(s_i)$ (Step 1.2.1).

Each normalized diagnosis string is embedded using the domain-adapted transformer encoder $f_{\hat{\theta}}$ defined in Section Pre-training of a Transformer-based model for Italian medical documents (TPT). The resulting embedding is

$$z_i = f_{\hat{\theta}}(\bar{s}_i) \in \mathbb{R}^d, \quad i \in I. \tag{3}$$

Let $Z = \{z_i\}_{i\in I}$ denote the set of all diagnosis embeddings (Step 1.2.2).

Since high-dimensional embeddings are not directly suitable for density-based clustering, we apply Principal Component Analysis (PCA) to Z (Step 1.2.3). Let $W \in \mathbb{R}^{d\times k}$ denote the projection matrix obtained by fitting PCA on Z, where in our case studies $d = 768, k = 100$, retaining more than 80% of the variance in the embedding space. The projected embeddings are $\tilde{z}_i = W^\top z_i \in \mathbb{R}^k$.

After this step, a clustering algorithm can be applied to the resulting embedding vectors (Step 1.2.4). We selected HDBSCAN[62] because it does not require specifying the number of clusters a priori, allows leaving part of data unclustered, and can identify clusters with heterogeneous shapes and densities. These properties are well suited to transformer-based text embeddings, which often lie on complex, nonlinear manifolds and may form clusters with varying densities, irregular

geometries, and overlapping boundaries. Several commonly used clustering algorithms, such as K-Means, Gaussian Mixture Models, and standard hierarchical clustering, are less suitable for this type of data[63,64], as further supported by a comparative analysis reported in Supplementary Material E.

We apply HDBSCAN to the set $\{\tilde{z}_i\}_{i\in I}$. The clustering defines an assignment function $h : \mathbb{R}^k \to \{1,\dots,K\} \cup \{-1\}$, where $h(\tilde{z}_i) = k$ indicates membership in cluster C_k, and $h(\tilde{z}_i) = -1$ denotes noise points left unclustered. HDBSCAN is parameterized by a minimum cluster size and a minimum number of samples, selected via grid search as detailed in Supplementary Material E.

The first-level clustering may produce multiple compact clusters corresponding to the same disease category. To merge semantically related clusters, we construct a keyword-based representation for each cluster.

For each cluster C_k, we define a set of representative keywords $\text{Keywords}(C_k)$ obtained by applying frequency-based filtering to the tokens appearing in C_k after tokenization and stopword removal. These keywords provide a compact textual summary of the cluster. The complete keyword extraction procedure is described in Supplementary Material E.

Each cluster is then represented by embedding its keyword summary using the same encoder $f_{\hat{\theta}}$, yielding a cluster-level embedding $u_k \in \mathbb{R}^d$. A second-level clustering procedure is applied to $\{u_k\}$ to merge clusters that are semantically close (Step 1.2.6).

Mapping clusters to diseases

Step 1.3 consists of mapping clusters to weak labels that will be used to train the classification model.

Let $\{C_k\}_{k=1}^K$ denote the clusters obtained in Step 1.2 and let $\text{Keywords}(C_k)$ be the keyword set associated with cluster C_k.

For each disease of interest, we specify one or more definitions. Each definition d consists of:

- a set of positive keywords $P_d = \{p_1,\dots,p_r\}$ that must be present,
- optionally, a set of negative keywords $N_d = \{n_1,\dots,n_s\}$ that must be absent.

A cluster C_k is selected as positive for the disease if there exists at least one definition d such that $P_d \subseteq \text{Keywords}(C_k)$ and $N_d \cap \text{Keywords}(C_k) = \emptyset$.

Let $L = \{k \in \{1,\dots,K\} : C_k \text{ satisfies at least one } d\}$ denote the set of clusters selected for the disease, $h(i)$ the cluster assignment of diagnosis string s_i (Step 1.2), and $I_c = \{i \in I : h(i) \neq -1\}$ the subset of documents whose diagnosis strings were assigned to a cluster (i.e., excluding noise points identified by HDBSCAN). The weak label is defined as

$$\tilde{y}_i = \begin{cases} 1 & \text{if } h(i) \in L, \\ 0 & \text{if } h(i) \notin L, \end{cases} \quad i \in I_c. \tag{4}$$

Documents with $h(i) = -1$ (i.e., diagnosis strings left unclustered) or $i \notin I$ (no diagnosis string extracted) do not receive a weak label and are excluded from the weakly-labelled training subset.

Thus, among the documents in I_c, discharge letters whose extracted diagnosis strings belong to selected clusters are assigned a positive weak label, while the remaining clustered documents receive a negative weak label.

This procedure is fully deterministic once the disease definitions are specified and does not depend on gold annotations. The only domain input required is the specification of disease-related keywords, which is substantially less demanding than document-level manual annotation. In practice, keyword definitions are specified in collaboration with clinicians, with attention to the terminology typically used in discharge letters for the disease of interest. This step does not require document-level review and is performed independently of gold labels. While the coverage of weak labels depends on the completeness of the keyword definitions, the cluster-level matching strategy reduces sensitivity to minor lexical variations.

A keyword-based selection at the cluster level is more robust than applying keywords directly to the raw diagnosis strings extracted in Step 1.1. Individual diagnosis strings are often heterogeneous, contain variable phrasing, and may include negations or references to past medical history. Direct keyword matching at the string level would therefore introduce a higher rate of false positives and false negatives. By contrast, clustering first aggregates semantically similar diagnoses, allowing keyword matching to operate on stabilized and semantically coherent cluster representations. Similarly, manually reviewing all extracted diagnosis strings would be impractical due to their large number.

Classification of discharge letters using the weakly-labelled dataset (WLC)

The classification pipeline that exploits the weakly-labelled data produced by Step 1 corresponds to Step 2 (WLC) in Figure 1.a. Let $\tilde{\mathscr{D}} = \{(x_i, \tilde{y}_i) : i \in I_c\}$ denote the weakly-labelled dataset constructed in Step 1.3. Only documents with an extracted diagnosis string are included in this training set.

Each document undergoes a deterministic cleaning function $p : \mathscr{X} \to \mathscr{X}$ that removes headers and footers containing non-informative administrative content (see Supplementary Material C for details). The processed document is denoted $\bar{x}_i = p(x_i)$ (Step 2.1).

Since RoBERTa-based models are limited to 512 input tokens, we define a truncation operator $\tau_{512}(\cdot)$ that retains the first 512 tokens of the processed document. The final input to the classifier is therefore $x_i^* = \tau_{512}(\bar{x}_i)$. In our dataset, approximately 90% of discharge letters contain fewer than 512 tokens after preprocessing, so truncation rarely affects the input. For datasets with a larger proportion of long documents, alternative strategies could be considered, such as splitting the document into segments of at most 512 tokens and aggregating segment-level probabilities, or using architectures designed for longer contexts, as discussed in studies on long-document classification with BERT-based models[65].

The classification model is based on the domain-adapted transformer encoder $f_{\hat{\theta}}$ defined in Section Pre-training of a Transformer-based model for Italian medical documents (TPT). A linear classification layer with parameters $w \in \mathbb{R}^d$ and bias $b \in \mathbb{R}$ is added on top of the encoder.

For an input document x_i^*, the predicted probability of the disease is $\hat{p}_i = \sigma\left(w^\top f_{\hat{\theta}}(x_i^*) + b\right)$, where $\sigma(z) = 1/(1+e^{-z})$ is the logistic function.

Model parameters $\phi = (\hat{\theta}, w, b)$ are optimized by minimizing the binary cross-entropy loss over the weakly-labelled dataset:

$$\mathscr{L}_{\mathrm{WLC}}(\phi) = -\sum_{i \in I_c} [\tilde{y}_i \log \hat{p}_i + (1-\tilde{y}_i)\log(1-\hat{p}_i)]\,. \tag{5}$$

Experimental outline

The study uses anonymized discharge letters obtained from the Pedianet network[50]. Inclusion in the Pedianet database is voluntary and parents/legal guardians must provide informed consent for their children's anonymized data to be used for research purposes. The study was performed in accordance with the Declaration of Helsinki. Ethical approval of the study and access to the database were granted by the Internal Scientific Committee of So.Se.Te. Srl, the legal owner of Pedianet.

The main evaluation is conducted at the final classification stage of the pipeline, where predictions are compared against manually annotated gold labels y_i^{gold}. In addition, intermediate components of the pipeline—diagnosis extraction (Step 1.1) and clustering (Steps 1.2–1.3)—are evaluated on a randomly selected subset of 1,000 diagnosis strings for each dataset, manually annotated by domain experts. Each string was labeled as *correct*, *partially correct*, or *wrong* with respect to (i) completeness of the extracted diagnosis string and (ii) correctness of its assignment to first- and second-level clusters.

At the classification stage, we compare our pipeline with:

- Two unsupervised rule-based baselines:
  - *Rule-based full-text*: keyword matching applied to the entire discharge letter;
  - *Rule-based diagnosis*: keyword matching applied only to extracted diagnosis strings.

  The keywords correspond to the disease definitions introduced in Step 1.3.
- An *XPRESS-style weak supervision baseline* (*WS-XPRESS-UmBERTo-TPT*). Following the XPRESS paradigm[45], we generate weak labels directly applying the keyword definitions defined in Step 1.3 (positive and negative sets) to the discharge letters and train the UmBERTo-TPT model used in our pipeline on these weak labels.
- A zero-shot large language model (LLM) for Italian (*LLaMAntino-3-ANITA-8B-Inst-DPO-ITA*)[66], prompted to classify each discharge letter as positive or negative for bronchiolitis.

To assess the contribution of the transformer backbone and domain-adaptive pre-training (TPT), we replace the classifier in Step 2.2 with:

1. UmBERTo without TPT (*WS-UmBERTo*)[60];
2. XLM-RoBERTa-base[67], with (*WS-XLM-RoBERTa-TPT*) and without (*WS-XLM-RoBERTa*) TPT;
3. RemBERT[68], with (*WS-RemBERT-TPT*) and without (*WS-RemBERT-BPT*) TPT;
4. An LSTM-based recurrent neural network using Italian Word2Vec embeddings (*WS-LSTM*)[54].

The three transformer-based models with TPT are also evaluated in a fully supervised setting (*FS-UmBERTo-TPT*, *FS-XLM-RoBERTa-TPT*, *FS-RemBERT-TPT*) trained directly on gold labels, providing an upper bound on performance under full supervision.

Since weak labels are derived from clustered diagnosis strings, classification models may over-rely on diagnosis sections. To assess this potential bias, we evaluate models under two conditions:

1. Using the full discharge letter;

2. Removing the diagnosis strings used to construct weak labels (when present).

To better understand the effect of weak supervision, we perform additional analyses. First, we assess the robustness of the results with respect to the construction of weak labels by removing individual clusters from the set of positive weak labels and re-evaluating the models. Second, we analyze the impact of label noise by replacing varying proportions of gold labels with weak labels and measuring the resulting performance degradation. Finally, we evaluate classification performance against the weak labels themselves, allowing us to quantify the discrepancy between weak and gold supervision (in Supplementary Material F).

Classification performance is evaluated using precision (P), recall (R), F1-score (F1) for the positive class, area under the ROC curve (AUROC), and area under the precision–recall curve (AUPRC). For classifiers producing a continuous score, the classification threshold is selected within training folds using Youden's index[69], and all evaluation metrics are computed on held-out test folds reflecting the original class prevalence.

All metrics are computed using stratified cross-validation. For the bronchiolitis dataset, we use 10-fold cross-validation. For the smaller bronchitis validation dataset, which contains 97 positive cases, we use 5-fold cross-validation to ensure a sufficient number of positive examples in each test fold. For each metric, we report the mean and standard deviation across folds.

To assess statistical significance of performance differences, we perform paired permutation (sign-flip) tests on fold-wise differences between our method and each other model. All tests are two-sided. To control for multiple comparisons across model variants, p-values are adjusted using the Holm–Bonferroni procedure within each metric. Differences are considered statistically significant when the Holm-adjusted p-value is below 0.05.

AUROC and AUPRC are not reported for rule-based classifiers, as they do not produce probabilistic outputs. Fully supervised models are not evaluated on weak labels.

Since data originate from multiple hospitals and Local Health Units (LHUs), additional experiments assessing inter-site variability are reported in Supplementary Material F.

Training details

Additional pre-training of the transformer-based models on the E3C corpus was performed using the masked language modeling objective described in Section Pre-training of a Transformer-based model for Italian medical documents (TPT). The models were optimized using a learning rate of $2 \cdot 10^{-5}$ and weight decay of 0.01. A validation set consisting of 20% of the corpus was used for early stopping.

The LSTM baseline consists of a frozen Word2Vec embedding layer followed by a 32-unit LSTM layer, a dropout layer, a 16-unit fully connected layer, an additional dropout layer, and the final output layer. The model is trained using the Adam optimizer with a learning rate scheduler starting from 10^{-2}, decay rate 0.7, and decay step size of 100. Early stopping is applied based on validation performance.

The transformer-based classifiers are trained using the AdamW optimizer with a linear learning rate scheduler starting from 10^{-3} and weight decay of 0.01. These hyperparameters were selected on training folds of the bronchiolitis dataset using grid search. The same configuration was applied unchanged to the bronchitis dataset. Training proceeds for a number of epochs determined by early stopping (6 epochs for UmBERTo models, 4 for XLM-RoBERTa models, and 3 for RemBERT models).

To stabilize optimization, a gradual unfreezing strategy is applied. Initially, only the final classification layer is trained. After two epochs, the last transformer layer is unfrozen, and after four epochs (if reached), the last but one transformer layer is also unfrozen.

To mitigate class imbalance during training, negative examples are undersampled within each training fold, limiting their number to at most twice that of the positive class. No undersampling is applied to test folds. Additional results supporting this choice are reported in Supplementary Material F.

All experiments were implemented in Python. Training was performed on a single NVIDIA Tesla V100 GPU with 32 GB of VRAM.

Results

Datasets

The dataset for the main case study comprises 33,176 discharge letters related to 15,196 children who were admitted to Emergency Rooms and/or hospitalized in the Veneto Region between January 2017 and December 2020, with a family pediatrician adhering to the Pedianet network[50].

The dataset covers 44 different hospitals in the region, which can be grouped into 9 Local Health Units (LHUs) and 2 university hospitals. Each discharge letter includes the complete text of the letter, along with the date of the letter and an

anonymized identifier of the patient. A manual review of discharge letters was conducted to extract relevant diagnoses of bronchiolitis, resulting in 335 positive cases within this dataset. This review was performed by a data manager with a clinical background, with the support of clinicians for uncertain cases. These labels are the gold standard to validate our pipeline, and we do not make direct use of them in the pipeline training (except for the fully supervised experiments). The median length of these letters is 4,515 characters (1st-3rd quartiles 3,240-5,806).

As an additional validation setting, we analyze a smaller dataset of 3,188 discharge letters related to 2,364 children who were admitted to Emergency Rooms and/or hospitalized in the Veneto Region between January 2020 and May 2025, with a family pediatrician adhering to the Pedianet network. This dataset covers 31 hospitals and 11 LHUs. Manual review, following the same approach of the bronchiolitis dataset, identified 97 cases of bronchitis, which serve as gold labels for evaluation. The median length of these letters is 5,387 characters (1st-3rd quartiles 2,601-6,759).

Supplementary Material A.1 reports additional plots with data distributions among hospitals and LHUs for the two datasets. An example of a fictitious letter is reported in Supplementary Material A.2.

A secondary annotator independently reviewed a stratified sample of 30 positive and 30 negative discharge letters for each dataset, resulting in perfect inter-annotator agreement for bronchiolitis ($\kappa = 1.00$) and substantial agreement for bronchitis ($\kappa = 0.73$), with discrepancies primarily involving less certain cases.

Diagnosis extraction

Our diagnosis extraction module (Step 1.1) successfully extracted a diagnosis sentence from 89% of the discharge letters in the bronchiolitis dataset. We extracted 15,038 unique diagnosis sentences, of which 11,279 occurred only once, confirming the need for clustering them in wider groups that can be more easily related to known diseases and symptoms. ICD-9-CM codes are present only in 737 diagnosis strings (4.9%), covering 3,436 discharge letters (10.4%). The most common strings are the simplest, while more articulate ones have single occurrences. The median length of the extracted diagnoses strings is 49 characters, with first and third quartiles of 17 and 156 characters. Among the 335 bronchiolitis-labelled letters, we extracted 138 unique diagnosis strings from 306 letters.

The same extraction procedure was applied to the bronchitis validation dataset. Diagnosis strings were successfully extracted from 90% of the discharge letters, yielding 1,376 unique diagnosis strings, of which 1,092 occurred only once. Considering the 97 bronchitis-labelled letters, diagnosis strings were extracted from all of them, producing 88 unique diagnosis expressions.

The manual evaluation results of this step on a random sample of 1,000 strings for each dataset are summarized in Figure 2.a. The number of extracted strings that are not diagnosis strings or that are incomplete is very limited ($\leq 2\%$).

Diagnosis clustering

In the bronchiolitis dataset, the 15,038 extracted diagnosis strings were grouped into 590 clusters. Cluster sizes varied significantly, from a large cluster of 270 strings to several smaller clusters containing 5 strings. The subsequent second-level clustering reduced the number of clusters to 135. In the bronchitis dataset clusters were 77 and 37, respectively. A summary of the main clusters is reported in Supplementary Material E.

The quality of the clustering assignments was evaluated on a random sample of 1,000 diagnosis strings per dataset. Figure 2.b reports the proportion of assignments that were manually classified as correct, partially correct, or wrong at both clustering levels. Most of the *partially correct* cluster assignments are due to minor variations in disease specification, e.g., *fracture of right hand* assigned to a cluster for *fracture of left hand*, *severe influenza* assigned to a cluster for *mild influenza*. A smaller number of cases correspond to diagnoses mentioning multiple symptoms or diseases that only partially match the cluster definition, such as *vomiting and fever* assigned to a cluster describing *fever* alone.

Table 1 reports the main first-level clusters related to respiratory infections for both datasets, including the number of bronchiolitis or bronchitis cases covered by each cluster. Given space constraints, we report here only the 15 largest clusters mentioning bronchiolitis, bronchitis, bronchospasm, pneumonia, respiratory infections, otitis, or fever, since these are the clusters most closely related to the respiratory diseases considered in our case studies. For the bronchiolitis dataset, four clusters (#1, #2, #3, #5) contain the keyword *"bronchiolitis"*. Three of them include a large majority of bronchiolitis cases, whereas the fourth cluster (#5) has only half of its diagnoses labelled as bronchiolitis. These four clusters together cover 65.38% of all bronchiolitis cases. Additionally, 13 clusters include the keyword *"bronchospasm"*, a symptom frequently associated with bronchiolitis. According to physicians who annotated the data, bronchospasm is associated with bronchiolitis only if accompanied by fever. Only one of these 13 clusters (#4) has both *"bronchospasm"* and *"fever"* keywords, and indeed, it contains a high proportion (70%) of bronchiolitis cases, accounting for an additional 15% of the total bronchiolitis cases. Thus, these five clusters are those that, without using the gold labels, would be selected to define the weak labels for the training set, according to the keyword-based definitions *{positive={bronchiolitis},negative={}}* and *{positive={bronchospasm, fever},negative={}}*. Among the respiratory infection clusters, three additional clusters contain bronchiolitis cases (#6, #7, #8), although in a small number. For the bronchitis dataset, three clusters are selected at the first level (#1, #4, #6), according to the

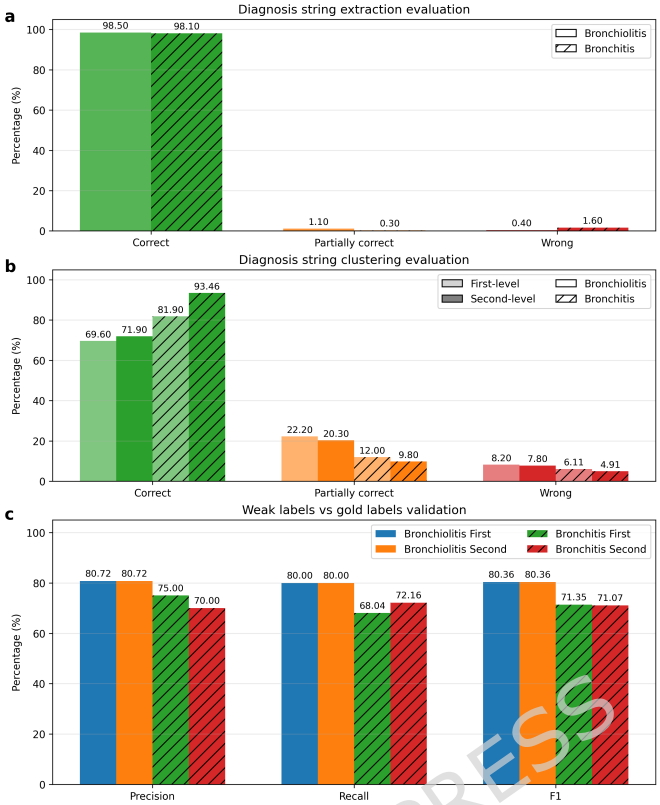

Figure 2. Evaluation of the intermediate steps of the pipeline on the bronchiolitis and bronchitis datasets. a. Evaluation of diagnosis string extraction on 1,000 manually annotated diagnosis strings per dataset, reporting the percentage of extracted strings classified as correct, partially correct, or wrong. b. Evaluation of diagnosis clustering after first-level and second-level clustering on the same set of annotated diagnosis strings. c. Validation of weak labels against gold labels in terms of precision, recall, and F1-score, reported separately for labels obtained from first-level and second-level clustering.

definitions *{positive={bronchitis},negative={}}* and *{positive={lower, tract, respiratory, infection},negative={upper}*. These cover 68.05% of bronchitis cases.

Table 2 reports the second-level clusters obtained from first-level clusters reported in Table 1. In particular, several clusters related to *bronchiolitis*, *bronchospasm*, *bronchitis*, and *respiratory infections* are aggregated into broader second-level clusters.

In the bronchiolitis case study the selection of weak labels for training remains unchanged, as second-level clusters #1, #2, #3, and #4 cover exactly the clusters that would have been selected at the first level. Consequently, a downstream comparison between first- and second-level clustering would yield identical weak labels and identical classification results in this specific setting. In the bronchitis dataset, the second level clustering merges two clusters that are partially related to *lower tract respiratory infections*, adding them to the set of positive weak labels.

Second-level clustering substantially reduces the number of clusters, to 20–50% of the original clusters in our datasets. In the specific case studies, Figure 2.c shows no effect on bronchiolitis weak labels and an increase in recall, at the cost of a decrease in precision, for bronchitis. Beyond these specific case studies, Figure 2.b provides an independent evaluation of clustering quality on a randomly selected sample of 1,000 diagnosis strings. In this broader assessment, second-level clustering increases the proportion of correct assignments and reduces partially correct or incorrect mappings in both datasets. This indicates a systematic refinement of cluster coherence, supporting the role of second-level clustering in improving weak-label quality and interpretability, although this does not necessarily translate into substantial gains in downstream classification performance.

The selected clusters cover 80% of the original bronchiolitis cases and 72% of the bronchitis cases, as illustrated in Figure 2.c, which compares weak labels derived from clustering with the manually annotated gold labels.

To better understand the false negatives, we reviewed 20 discharge letters labelled as bronchiolitis in the gold standard but not covered by our weak labels. We found 10 cases where the diagnosis refers to other symptoms or diseases (e.g., otitis, dyspnea, fever) while bronchiolitis is mentioned elsewhere in the letter; 4 cases where bronchiolitis appears in the diagnosis

Table 1. Keyword-based representation of the first 15 first-level clusters related to respiratory infections in each dataset, with the % of cases in each cluster, the % to the total number of cases in the dataset and the indication of the clusters that would be selected as weak labels at this stage.

Cluster #	Cluster keywords [ITA]	Cluster keywords [ENG]	Within cluster cases %	% of total cases covered	Positive weak label
		Bronchiolitis			
1	bronchiolite lieve	mild bronchiolitis	85.19	6.87	y
2	acuta bronchiolite iniziale lieve	acute mild initial bronchiolitis	85.14	44.48	y
3	acuta bronchiolite corso insufficienza respiratoria	acute bronchiolitis undergoing respiratory failure	74.05	11.94	y
4	acuto broncospasmo corso febbre paziente	acute bronchospasm ongoing fever patient	70.00	15.02	y
5	alimentazione bronchiolite difficolta'	feeding difficulties bronchiolitis	50.00	2.09	y
6	acuta infezione insufficienza respiratoria transitoria virosi	acute respiratory infection transient virosis	12.70	2.39	n
7	broncospasmo infezione respiratoria	bronchospasm respiratory infection	6.67	0.30	n
8	broncospasmo disidratazione lieve	bronchospasm mild dehydration	1.17	0.60	n
9	flogosi alte vie respiratorie febbre	high airways phlogosis fever	0.00	0.00	n
10	broncospasmo infezione vie respiratorie	bronchospasm airways infection	0.00	0.00	n
11	acuto broncospasmo	acute bronchospasm	0.00	0.00	n
12	broncospasmo flogosi vie	bronchospasm airways phlogosis	0.00	0.00	n
13	broncospasmo polmonite	bronchospasm pneumonia	0.00	0.00	n
14	broncospasmo corso	ongoing bronchospasm	0.00	0.00	n
15	broncopolmonite sinistra	left bronchopneumonia	0.00	0.00	n
		Bronchitis			
1	bronchite acuta	acute bronchitis	90.48	39.18	y
2	broncospasmo acuto parainfettivo	acute parainfectious bronchospasm	22.73	5.15	n
3	broncospasmo acuto	acute bronchospasm	16.67	5.15	n
4	bronchite	bronchitis	84.62	22.68	y
5	nausea vomito febbre	nausea vomit fever	5.00	1.03	n
6	insufficienza respiratoria infezione basse	respiratory insufficiency lower infection	30.00	6.19	y
7	rinite polmonite basale destra	right basal pneumonia rhinitis	10.00	3.09	n
8	broncospasmo	bronchospasm	9.09	2.06	n
9	polmonite sinistra insufficienza respiratoria	left pneumonia respiratory insufficiency	9.09	1.03	n
10	otite media acuta	acute otitis media	0.00	0.00	n
11	infezione basse alte vie respiratorie	respiratory infection lower upper	33.33	4.12	n
12	laringite ipoglottica	hypoglottic laryngitis	0.00	0.00	n
13	bronchiolite lieve	mild bronchiolitis	4.76	1.03	n
14	infezione	infection	3.57	1.03	n
15	febbre flogosi alte vie	fever upper tract phlogosis	3.23	2.06	n

alongside other diseases, leading to assignment to a cluster related to the other condition; 3 cases of respiratory issues without explicit mention of bronchiolitis, respiratory syncytial virus, or bronchospasm with fever; 2 cases with no diagnosis string present; and 1 case referring to bronchiolitis in a family member rather than the patient. These results suggest that most false negatives arise from the absence of explicit bronchiolitis indications in the diagnosis strings, rather than from clustering errors.

To assess the effect of the TPT step on clustering results, Supplementary Material E reports the clusters obtained using the original UmBERTo model in our clustering pipeline, on the bronchiolitis dataset. The number of clusters related to respiratory infections is higher without TPT (34 vs 27), and their keywords appear more heterogeneous. Selecting weak labels by applying the same criterion we used on our clusters, we would cover only 47% of bronchiolitis cases with the clusters from the original UmBERTo model. Moreover, these clusters are more heterogeneous since they introduce more false negatives in the weak labels (precision decreases to 72 %).

Additional results on the effect of clustering algorithms and hyperparameters are reported in Supplementary Material E.

Disease classification

Table 3 reports the cross-validation results for the classification task, including precision, recall, F1-score, AUROC, and AUPRC. The table compares our weakly supervised pipeline with different transformer-based classifiers, fully unsupervised rule-based and zero-shot LLM approaches, and fully supervised versions of the transformer-based classifiers. Results are shown for both datasets and for two experimental settings: keeping the diagnosis string in the discharge letter or removing it.

On the bronchiolitis dataset, our weakly supervised pipeline outperforms fully unsupervised approaches and maintains only a limited gap with the fully supervised models. The TPT step consistently improves performance across models: the differences are limited, but there are no cases in which TPT hurts performance. Among the transformer-based classifiers, UmBERTo generally achieves the best results, outperforming RemBERT and XLM-RoBERTa across most metrics. The LSTM classifier shows a clear performance gap with respect to the transformer-based models.

Rule-based full-text, which verifies the presence of keywords in the entire letter, shows lower precision and recall than the majority of the weakly supervised models, emphasizing the need for a more sophisticated approach. The results further decline when the diagnosis string is excluded. *Rule-based diagnosis*, using only diagnosis strings, yields better precision than the rule applied to the complete text, though recall remains similar.

Table 2. Keyword-based representation of the second-level clusters that are obtained as merge of first level clusters reported in Table 1, with the % of cases in each cluster, the % to the total number of cases in the dataset and the indication of those selected as positive weak labels.

1st lvl cluster	2nd level cluster keywords [ITA]	2nd level cluster keywords [ENG]	1st lvl cluster	1st level cluster keywords [ITA]	1st level cluster keywords [ENG]	Within cluster bronch. %	% of total bronch. covered	Positive weak label
Bronchiolitis								
1	bronchiolite lieve	mild bronchiolitis	1	acuta bronchiolite iniziale lieve	acute mild initial bronchiolitis	85.16	51.34	y
			2	bronchiolite lieve	mild bronchiolitis			
2	broncospasmo infezione respiratoria	bronchospasm respiratory infection	7	broncospasmo infezione vie respiratorie	bronchospasm infection airways	2.90	0.30	n
			10	broncospasmo infezione respiratoria	bronchospasm respiratory infection			
3	broncospasmo	bronchospasm	11	acuto broncospasmo	acute bronchospasm	0.00	0.00	n
			12	broncospasmo flogosi vie	bronchospasm airways flogosis			
			13	broncospasmo polmonite	bronchospasm pneumonia			
			14	broncospasmo corso	ongoing bronchospasm			
			22	broncospasmo otite	bronchospasm otitis			
			23	broncospasmo parainfettivo	bronchospasm parainfective			
			24	broncospasmo virosi	bronchospasm virosis			
			17	broncospasmo episodio	bronchospasm episode			
4	infezione vie respiratorie	airways infection	18	infezione alte vie respiratorie	high airways infection	0.00	0.00	n
			19	flogosi infezione alte vie respiratorie	flogosis high airways infection	0.00	0.00	
			27	altre infezioni acute vie respiratorie	other acute respiratory infection airways	0.00	0.00	
			26	crisi epilessia infezione paziente polmonare respiratoria	seizure epilepsy respiratory infection lung patient	0.00	0.00	
			16	febbricola febbrile infezione respiratoria virosi	febrile fever respiratory infection virosis	0.00	0.00	
Bronchitis								
1	bronchite	bronchitis	1	bronchite acute	acute bronchitis	88.24	61.86	y
			4	bronchite	bronchitis			
2	broncospasmo	bronchospasm	2	broncospasmo acuto parainfettivo	acute parainfectious bronchospasm	16.22	12.37	n
			3	broncospasmo acuto	acute bronchospasm			
			8	broncospasmo	bronchospasm			
3	infezione basse vie respiratorie	lower tract respiratory infection	6	insufficienza respiratoria infezione basse	respiratory insufficiency lower infection	31.25	10.31	y
			11	infezione alte basse vie respiratorie	respiratory infection lower upper tract			

The zero-shot approach with the Italian LLaMAntino-3 LLM does not outperform our best weakly supervised results but shows competitive performance.

The model trained using weak labels derived from the XPRESS framework performs substantially worse than all other approaches in this setting. This is expected, since this implementation of the XPRESS framework applies the same keyword rules defined for our cluster selection across the entire clinical text, generating a larger number of noisy weak labels compared with the more targeted extraction of diagnosis strings.

Results on the bronchitis validation dataset show a broadly similar ranking of models, suggesting consistent trends with the main case study. Despite the substantially smaller dataset size, transformer-based classifiers again outperform rule-based and zero-shot approaches, and UmBERTo-based models achieve the best overall performance among the weakly supervised methods. These findings, however, should be interpreted with caution given the limited number of positive bronchitis cases.

We also repeated the bronchitis experiments without applying the second-level clustering step when generating weak labels. This led to a small decrease in performance (AUPRC reduction of 1-2%), indicating that the impact of second-level clustering on downstream performance is limited in this setting. Detailed results are reported in Supplementary Material F.

Precision–recall curves for the different models are reported in Figure 3, providing a complementary view of model ranking performance across recall levels.

In Figure 4, we observe the decrease in performance when replacing increasing percentages of gold labels with weak labels. This scenario is representative of many real-world settings where only a fraction of the data can be manually annotated. The results show a generally smooth and monotonic decrease in AUPRC as the proportion of weak labels increases, with similar trends observed for both bronchiolitis and bronchitis.

Table 3. Classification results on gold-labels reported as mean (standard deviation) across cross-validation folds. Bold indicates best overall performance per metric, per dataset. Underline indicates best performance among non-fully-supervised models. Statistical significance was assessed using two-sided paired permutation tests across folds, comparing each model to WS-UmBERTo-TPT. P-values were adjusted using the Holm correction within each metric ($^{*}p_{adj} < 0.05$, $^{**}p_{adj} < 0.01$).

Learning	Classifier	Bronchiolitis					Bronchitis				
		P [%]	R [%]	F1 [%]	AUROC [%]	AUPRC [%]	P [%]	R [%]	F1 [%]	AUROC [%]	AUPRC [%]
With diagnosis string											
Weakly-supervised	WS-UmBERTo-TPT	78.34 (6.77)	77.57 (6.89)	78.14 (4.89)	77.68 (4.30)	73.13 (4.93)	73.06 (5.95)	82.11 (5.82)	78.02 (5.25)	82.30 (5.10)	76.72 (5.02)
	WS-XLM-RoBERTa-TPT	69.54* (5.45)	71.33 (5.12)	70.85* (4.52)	69.24* (3.89)	62.69* (3.67)	71.57 (5.94)	75.26* (5.46)	73.37* (5.16)	76.41* (5.32)	71.01 (5.45)
	WS-RemBERT-TPT	67.41* (7.01)	70.11 (6.82)	69.23* (5.24)	68.00* (5.03)	60.51* (3.47)	69.23 (5.28)	74.23* (6.12)	71.64* (5.87)	76.15* (5.17)	68.39* (5.56)
	WS-UmBERTo	68.29 (6.53)	75.99 (6.41)	72.15 (4.93)	71.21 (4.21)	63.29 (5.66)	73.27 (5.73)	76.29* (6.19)	74.75 (5.45)	80.31 (5.36)	75.29 (5.98)
	WS-XLM-RoBERTa	67.49* (5.28)	68.67* (5.09)	68.02* (4.32)	67.65* (3.65)	61.11* (3.23)	70.14 (6.14)	73.96* (5.65)	72.00* (5.49)	78.84 (5.91)	72.24* (6.11)
	WS-RemBERT	65.24* (6.81)	63.82* (6.84)	64.96* (5.13)	65.74* (4.92)	59.02* (3.16)	68.41* (5.81)	73.21* (5.47)	70.73* (5.66)	76.52* (6.06)	69.20* (5.41)
	WS-LSTM	57.39** (8.31)	60.86** (11.52)	59.87** (9.69)	63.29** (7.36)	49.94** (5.13)	60.78** (6.25)	63.92** (5.66)	62.31** (6.28)	70.12** (6.24)	60.55** (5.54)
	WS-XPRESS-UmBERTo-TPT	16.57** (2.21)	60.52** (9.18)	29.02** (9.04)	54.18** (10.24)	19.91** (1.74)	38.46** (5.96)	51.55** (5.81)	44.05** (5.80)	56.89** (5.77)	42.84** (5.45)
Unsupervised	Rule-based full-text	65.57* (6.56)	68.76* (4.00)	67.22* (6.19)	-	-	15.79** (5.51)	32.88** (5.67)	21.33** (5.81)	-	-
	Rule-based diagnosis	74.61 (4.23)	67.42* (3.15)	70.74* (3.41)	-	-	21.74** (6.15)	20.55** (5.68)	21.13** (5.87)	-	-
	LLaMAntino-3	78.26 (5.72)	74.64* (7.48)	75.12 (4.29)	71.67 (3.48)	67.92* (4.47)	65.21* (5.51)	72.61* (5.89)	68.71* (5.67)	75.29* (5.68)	68.13* (4.74)
Supervised	FS-UmBERTo-TPT	80.30 (6.24)	**81.52*** (6.59)	80.83 (4.12)	**83.45*** (4.02)	**77.57*** (2.02)	**74.49** (4.45)	**82.47** (4.68)	**78.28** (4.51)	**85.93** (4.47)	**79.48** (3.52)
	FS-XLM-RoBERTa-TPT	75.29* (5.21)	78.31 (4.36)	76.54 (4.14)	78.54 (3.42)	72.23 (2.70)	74.29 (5.55)	80.41 (5.01)	77.23 (4.74)	83.82 (5.54)	76.10 (4.61)
	FS-RemBERT-TPT	73.24 (6.58)	75.28 (5.93)	74.21* (4.68)	76.39 (4.31)	68.33 (5.76)	72.66 (5.52)	76.70 (5.54)	74.62 (6.12)	83.14 (5.73)	75.26 (5.61)
Without diagnosis string											
Weakly-supervised	WS-UmBERTo-TPT	74.25 (6.54)	73.52 (6.12)	73.89 (5.64)	74.27 (4.83)	69.27 (4.36)	72.09 (5.86)	**79.90** (5.81)	75.79 (5.80)	79.26 (5.74)	74.71 (5.18)
	WS-XLM-RoBERTa-TPT	66.83* (5.69)	67.52* (5.88)	67.19* (4.57)	67.25* (3.65)	59.29* (4.52)	69.39 (5.86)	70.10* (5.51)	69.74* (5.84)	75.13 (5.64)	69.33* (5.54)
	WS-RemBERT-TPT	64.53* (6.58)	63.66* (6.21)	64.21* (5.84)	65.47* (4.97)	56.75* (3.43)	69.31 (5.56)	72.16* (5.97)	70.71* (5.63)	74.26* (6.38)	69.91* (5.61)
	WS-UmBERTo	67.55* (6.38)	76.40 (6.25)	71.74 (5.47)	69.98 (4.59)	63.27 (2.66)	72.00 (6.01)	74.23* (5.67)	73.10 (5.73)	73.45* (5.39)	69.09 (5.84)
	WS-XLM-RoBERTa	63.98* (5.72)	65.04* (5.63)	64.69* (4.98)	64.87* (4.12)	54.30* (3.31)	68.26 (5.64)	69.97* (5.11)	69.10* (5.88)	70.87* (5.42)	65.39* (5.15)
	WS-RemBERT	62.25* (6.58)	60.48** (6.96)	61.65* (5.24)	62.21* (4.87)	55.87* (4.71)	68.10 (5.26)	70.16* (5.69)	69.11* (5.54)	73.14* (5.91)	69.34* (5.82)
	WS-LSTM	50.14** (7.14)	63.97** (10.12)	56.34** (8.65)	59.02** (6.51)	43.30** (4.74)	58.45** (5.51)	61.87** (5.25)	60.11** (5.31)	66.52** (5.37)	60.24** (5.36)
	WS-XPRESS-UmBERTo-TPT	16.03** (6.52)	64.50* (12.62)	27.36** (8.48)	52.23** (10.12)	17.89** (2.33)	35.47** (5.54)	50.51** (6.24)	41.67** (5.94)	52.28** (5.64)	45.43** (6.02)
Unsupervised	Rule-based full-text	62.33** (6.78)	60.05** (8.42)	61.15** (6.83)	-	-	14.09** (5.61)	30.00** (5.89)	19.18** (5.76)	-	-
	LLaMAntino-3	70.42* (5.26)	72.98 (5.84)	71.16 (3.98)	68.52* (4.55)	62.24* (4.22)	60.12* (5.01)	67.25* (5.15)	63.49* (5.23)	68.21* (5.02)	59.31* (2.74)
Supervised	FS-UmBERTo-TPT	78.25* (5.20)	76.21 (4.87)	77.15 (4.01)	**78.51*** (3.84)	**75.44*** (3.55)	**72.99** (5.36)	79.59 (5.96)	**76.36** (5.84)	**82.63** (5.64)	**76.82** (6.37)
	FS-XLM-RoBERTa-TPT	72.54 (4.26)	73.65 (4.16)	72.96 (3.19)	74.08 (3.34)	70.61 (3.39)	73.33 (5.51)	79.38 (6.11)	76.24 (5.43)	81.15 (5.30)	74.34 (5.38)
	FS-RemBERT-TPT	71.49 (5.46)	70.87 (4.12)	71.21 (3.95)	72.68 (3.87)	64.06 (4.59)	71.57 (5.63)	75.26 (5.84)	73.37 (5.69)	78.26 (5.88)	71.28 (5.67)

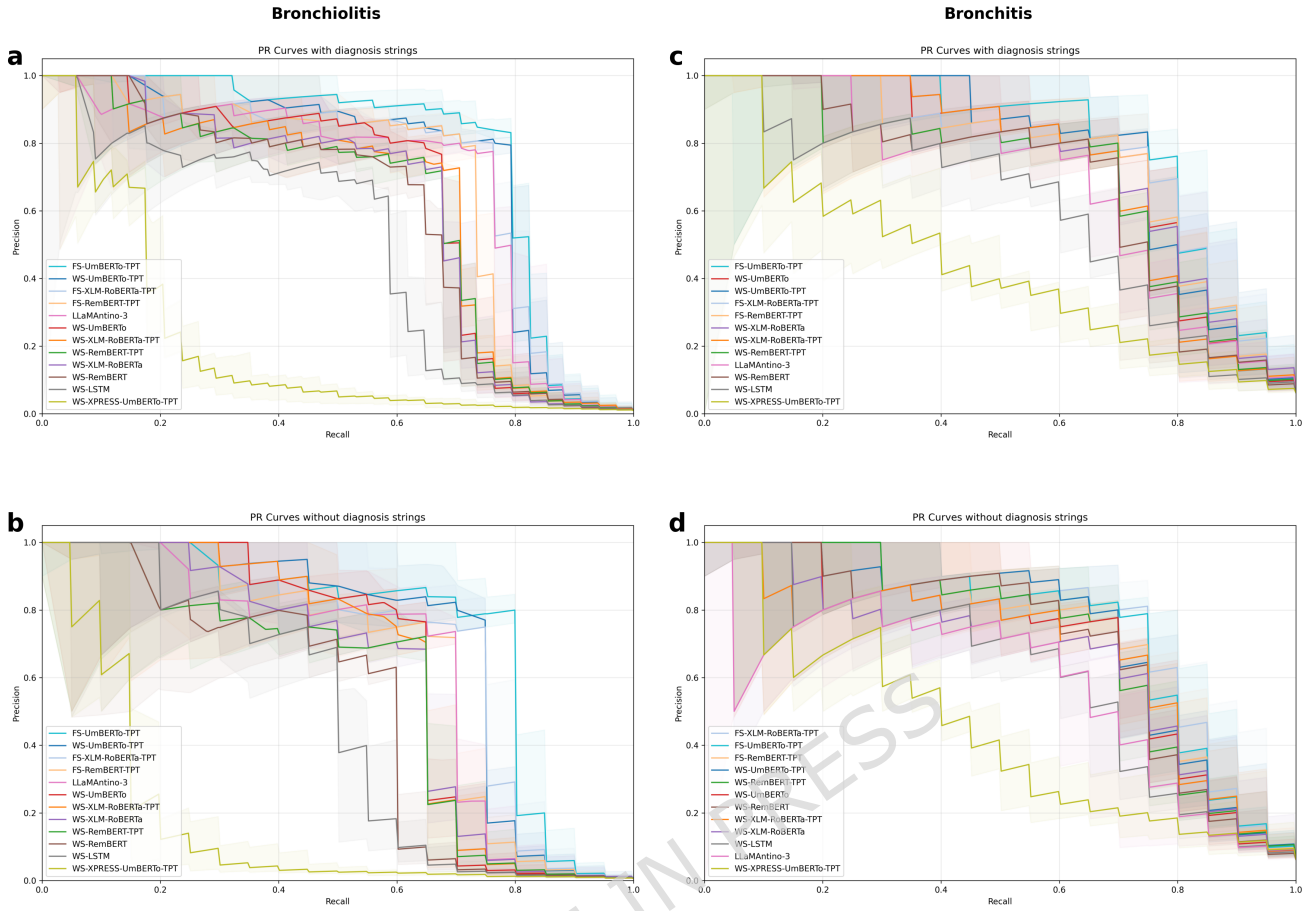

Figure 3. Precision-Recall curves on the bronchiolitis (a,b) and bronchitis (c,d) gold labels with (a,c) and without (b,d) diagnosis strings. Shaded areas represent 95% C.I.

In Table 4, we assess the sensitivity of our best weakly supervised model (*WS-UmBERTo-TPT*) to the selection of clusters used to generate weak labels. In particular, we remove one first-level cluster at a time from the set of positive clusters and evaluate the resulting classification performance.

For the bronchiolitis dataset, excluding clusters #5 or #1 has little effect on performance, which is reasonable considering their smaller size. Removing clusters #3 or #4 leads to a moderate decrease in performance. In particular, cluster #4 corresponds to the second bronchiolitis definition ("*bronchospasm*" and "*fever*"); excluding this cluster is therefore equivalent to omitting that definition in the keyword specification step. Only excluding cluster #2 produces a substantial drop in performance and an increase in variance, which is expected since it accounts for almost half of the bronchiolitis cases.

A similar trend is observed for the bronchitis dataset. Removing cluster #11 or #6 produces limited changes in performance, whereas excluding clusters #1 or #4 leads to a larger decrease in classification performance. Overall, these results confirm the robustness of the model with respect to moderate changes in the selection of clusters used to generate weak labels and, more generally, to variations in the disease definitions used in the weak-label construction step.

Additional analyses evaluating model performance when training and testing on weak labels, together with further stratified results across hospitals, LHUs, and pediatric vs non-pediatric emergency departments, are reported in Supplementary Material F. Some hospitals are more represented in the dataset, leading to superior performance, but also hospitals with a similar number of records show considerable performance variability, suggesting that reporting practices in discharge letters might influence disease detection.

Discussion

This study introduces a novel pipeline demonstrating how weakly supervised learning can be effectively applied to clinical NLP. To our knowledge, this is the first work addressing diagnosis identification from Italian discharge letters. Although we primarily

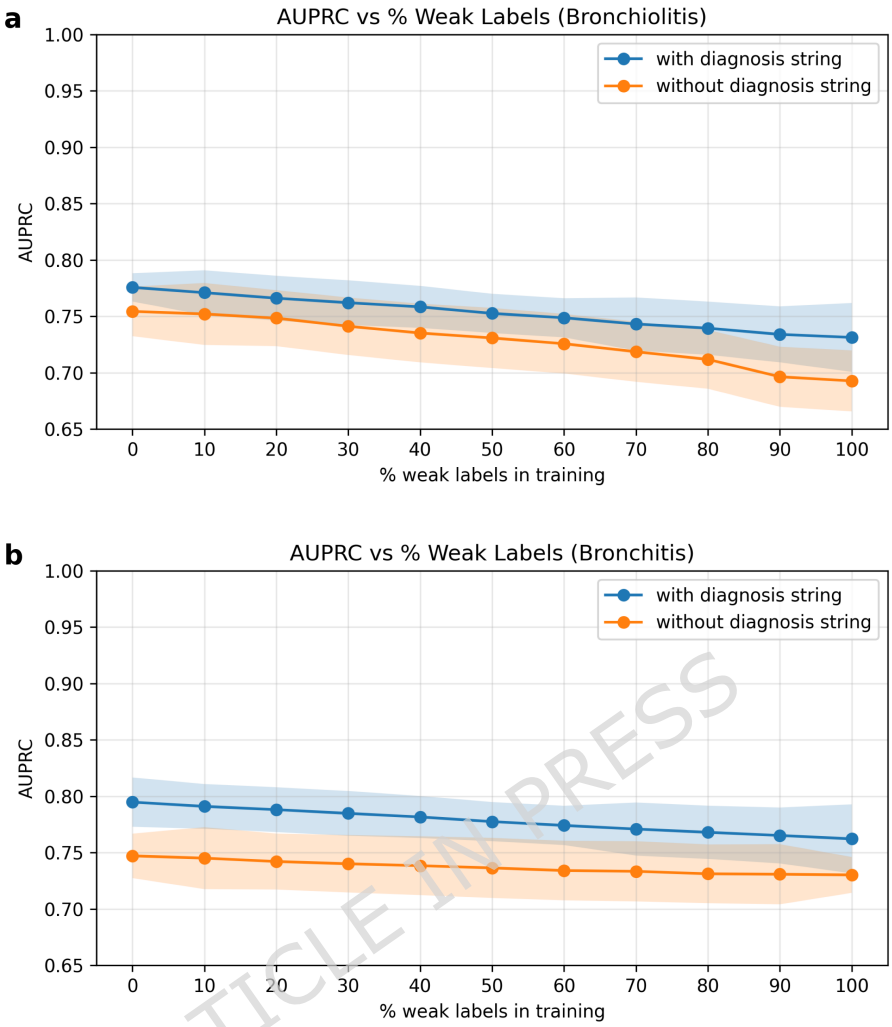

Figure 4. AUPRC with varying percentages of weak labels on the bronchiolitis (a) and bronchitis (b) datasets. Mean AUPRC (±95% CI) on gold labels, with and without diagnosis strings, as different proportions of gold labels are replaced by weak labels. Results are averaged over 5 repetitions of cross-validation. The y-axis is truncated to improve visibility of the trend.

validated it using a case study on bronchiolitis, we also conducted a secondary validation on a smaller dataset for bronchitis.

We observed a good mapping between diagnosis strings and their clusters' descriptions in the manually evaluated subset of strings, and indeed a good match between weak and gold labels, confirming that the SAL step can effectively identify relevant cases. Despite the inevitable noise introduced by weak labels, its magnitude remains limited, allowing the model to perform well when evaluated against gold labels. Broadly similar trends are observed in both datasets; however, results on the bronchitis dataset should be interpreted with caution given the limited number of positive cases. Additionally, the differences between settings where we retained the diagnosis strings and those where we removed them are limited. This indicates that the model is effectively learning from the entire text of the discharge letters, making it particularly advantageous when the automatic extraction of diagnosis strings fails. Our manual exploration of some weak labels further revealed that diagnosis strings do not always mention the disease responsible for hospitalization. In some cases, they only refer to symptoms, and in cases of multiple diseases, not all may be reported. These findings highlight the necessity for a classification model based on the complete text of discharge letters, as the one we developed.

In contrast, rule-based approaches underperformed because diseases like bronchiolitis and bronchitis are not always explicitly mentioned or may appear in negated or hypothetical contexts. While more sophisticated rule-based systems could address some of these issues, such approaches lack the scalability and generalizability of our pipeline. Conversely, our pipeline can be adapted to other diseases, making it a potentially versatile tool for clinical NLP. A similar issue arises when weak

Table 4. Classification results after removing one 1st level cluster for weak labels at a time, on the best weakly-supervised classification models. Cross validation results with std. dev. in parenthesis.

Clusters for weak labels	P [%]	R [%]	F1 [%]	AUROC [%]	AUPRC [%]
Bronchiolitis					
#1. #2. #3. #4.	70.75 (6.48)	81.27 (6.99)	75.65 (4.29)	75.49 (5.57)	67.95 (4.32)
#1. #2. #4. #5.	68.42 (8.70)	75.09 (8.58)	71.60 (7.26)	72.67 (6.87)	61.44 (4.75)
#2. #3. #4. #5.	68.54 (6.30)	82.35 (4.94)	74.81 (3.82)	75.83 (5.22)	66.07 (3.64)
#1. #3. #4. #5.	57.29 (12.97)	59.03 (15.94)	58.15 (12.45)	68.90 (10.04)	47.79 (3.39)
#1. #2. #3. #5.	67.86 (6.36)	73.42 (8.64)	70.53 (5.33)	70.22 (3.10)	62.81 (3.19)
Bronchitis					
#1. #4. #6.	73.41 (5.84)	79.12 (5.71)	76.26 (5.19)	79.66 (5.22)	75.03 (5.87)
#1. #4. #11.	75.67 (5.96)	77.23 (6.05)	76.64 (5.85)	77.22 (5.48)	73.78 (6.09)
#1. #6. #11.	66.52 (6.41)	70.12 (6.84)	68.35 (6.35)	70.52 (6.07)	61.91 (5.68)
#4. #6. #11.	62.15 (6.96)	60.23 (6.84)	61.15 (6.40)	61.22 (5.45)	56.68 (6.25)

labels are generated using the XPRESS framework applied across the entire clinical text, which produces substantially noisier labels compared with the diagnosis-string-based weak labeling used in our pipeline. The zero-shot LLM performed better than rule-based and some weakly supervised settings, but it did not surpass the best weakly supervised configuration. Despite the reasoning capacities of modern LLMs, their training data rarely include documents and tasks of this type, limiting performance.

Domain pretraining (TPT) improved results across all models. The best performance was obtained with UmBERTo, trained entirely on Italian data, confirming the advantage of language-specific representations even over larger multilingual models.

Compared with fully supervised classifiers, our weakly supervised pipeline shows a reduction in performance of 4–6% AUPRC. While this difference is statistically significant, its practical relevance depends on the intended application. In the context of cohort selection from large collections of discharge letters, the goal is to identify groups of potentially relevant patients at scale, rather than to achieve perfect classification at the individual level. In this case, such a reduction may be considered acceptable given the substantial decrease in annotation effort, which would otherwise require more than 1,500 hours of expert time in our case study. This highlights the practical trade-off between performance and scalability offered by weakly supervised approaches.

The experiments in which we progressively replaced gold labels with weak labels (Figure 4) further support this trade-off in a realistic scenario where only a small portion of the data can be manually annotated. Even in this mixed-label setting, performance degraded smoothly and remained close to the fully supervised baseline as long as a modest fraction of gold labels was available, indicating strong resilience to label noise.

The pipeline has also shown robustness to the cluster selection used to generate weak labels across both case studies. Performance degradation was proportional to the reduction in weak-label coverage, indicating stable behavior under moderate variations in disease definitions. Although second-level clustering provided no or only modest improvements in downstream performance in our case studies, it improved cluster coherence and weak label correctness in the independent evaluation of clustering quality (Figure 2.b), thereby contributing to more interpretable cluster structures.

It is important to clarify that, while the pipeline does not require document-level manual annotation, it relies on disease-specific keyword definitions provided as minimal domain input. This step requires collaboration with clinicians to reflect the terminology commonly used in discharge letters, but it is substantially less demanding than full annotation of individual documents. As shown in the robustness analysis, moderate variations in cluster selection lead to proportionally limited changes in performance, suggesting that the approach is not overly sensitive to minor imperfections in keyword specification.

While our results are promising, several limitations should be acknowledged. First, the gold labels for both datasets were assigned primarily by a single annotator, introducing potential uncertainty in the reference standard. We performed an additional inter-annotator agreement analysis on a stratified subset of discharge letters, but this evaluation was limited in size and does not fully replace a systematic double-annotation process on the complete datasets. Although the task consisted of binary disease identification, which is less complex than fine-grained clinical annotation, some degree of subjectivity cannot be excluded. Second, a limitation concerns the additional validation on the bronchitis dataset. Due to the relatively small number of positive cases (97 in total), each test fold in the 5-fold cross-validation contains approximately 19 positive examples. This results in limited statistical power to detect differences between models and increases the variability of performance estimates. Consequently, results on the bronchitis dataset should be primarily viewed as indicative of generalization trends rather than as a definitive comparative evaluation. Third, the multi-centric nature of our dataset, though a strength in its diversity, also presents an imbalance in the number of records from different hospitals and LHUs, which are also all located within a single

Italian region. This is not only due to the geographical distribution of the population, but might also reflect some biases in the distribution of the family pediatricians adhering to the Pedianet network. Testing the pipeline on more balanced datasets and externally validating it on data from other regions would strengthen its generalizability and allow a more precise assessment of performance differences among hospitals. Finally, the evaluation of intermediate steps relied on the manual review of a relatively small subset of diagnoses strings. Extending this annotation to a larger subset and assessing entire clusters would require substantial effort but could enable a more detailed analysis of the performance of each component in the pipeline.

Future research can extend this work along multiple directions. Although we evaluated the pipeline on two different disease-specific datasets, further validation on additional clinical conditions and datasets is required to fully assess its generalizability. Future studies should also include complete double annotation to formally assess inter-annotator agreement and better quantify the reliability of reference labels. Incorporating hospital or LHU identifiers in hierarchical classification models could further improve performance by accounting for site-specific documentation practices. Moreover, enhancing the interpretability of model predictions will be crucial for widespread clinical adoption. Identifying which portions of each discharge letter most influence the classifier's decision would help clinicians assess reliability and build trust. Techniques from explainable artificial intelligence could be applied to reveal the reasoning process behind model outputs and possibly recognize potential biases present in the training data.

These future developments would further strengthen both the methodological robustness and clinical utility of weakly supervised NLP systems for healthcare text, supporting scalable and trustworthy solutions for real-world applications.

Conclusions

We presented a weakly supervised pipeline for disease identification from clinical discharge letters, designed to minimize the need for manual annotation while maintaining competitive classification performance. The proposed approach combines automatic extraction of diagnosis strings, clustering-based weak-label generation, and transformer-based text classification.

In a case study on bronchiolitis using Italian discharge letters, the pipeline outperformed unsupervised rule-based approaches and achieved performance close to fully supervised models. Similar model rankings were observed in a secondary validation on a smaller bronchitis dataset. The results also highlight the importance of targeted weak-label generation: simpler strategies, such as in the XPRESS-based baseline, produced substantially noisier labels and markedly worse performance.

Experiments with mixed gold and weak labels further showed that the proposed approach is resilient to label noise and can maintain strong performance even when only a limited portion of the dataset is manually annotated. In addition, the classification models were able to effectively learn from the entire discharge letter text, reducing dependence on perfectly extracted diagnosis strings.

By substantially reducing the annotation burden while maintaining robust performance, this approach offers a practical and scalable framework for applying NLP to large collections of clinical documents.

References

1. Tayefi, M. *et al.* Challenges and opportunities beyond structured data in analysis of electronic health records. *WIREs Comput. Stat.* **13**, e1549 (2021).
2. Cowie, M. R. *et al.* Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.* **106**, 1–9 (2017).
3. Nadkarni, P. M., Ohno-Machado, L. & Chapman, W. W. Natural language processing: an introduction. *J. Am. Med. Informatics Assoc.* **18**, 544–551 (2011).
4. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* **82**, 3713–3744 (2023).
5. Hossain, E. *et al.* Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Comput. Biol. Medicine* **155**, 106649 (2023).
6. Friedman, C. & Elhadad, N. Natural language processing in health care and biomedicine. *Biomed. Informatics: Comput. Appl. Heal. Care Biomed.* 255–284 (2014).
7. Iroju, O. G. & Olaleke, J. O. A systematic review of natural language processing in healthcare. *Int. J. Inf. Technol. Comput. Sci.* **8**, 44–50 (2015).
8. Névéol, A., Dalianis, H., Velupillai, S., Savova, G. & Zweigenbaum, P. Clinical natural language processing in languages other than english: opportunities and challenges. *J. Biomed. Semant.* **9**, 1–13 (2018).
9. World Health Organization. *International Statistical Classification of Diseases and Related Health Problems – 10th Revision – 5th Edition*, vol. 3 (World Health Organization, 2016).

10. Walsh, S. H. The clinician's perspective on electronic health records and how they can affect patient care. *BMJ* **328**, 1184–1187 (2004).

11. Millares Martin, P. Consultation analysis: use of free text versus coded text. *Heal. Technol.* **11**, 349–357 (2021).

12. Walsh, P. & Rothenberg, S. J. Which ICD-9-CM codes should be used for bronchiolitis research? *BMC Med. Res. Methodol.* **18**, 1–11 (2018).

13. Cozzolino, F. *et al.* Accuracy of colorectal cancer ICD-9-CM codes in italian administrative healthcare databases: a cross-sectional diagnostic study. *BMJ Open* **8**, e020630 (2018).

14. Cipparone, C. W. *et al.* Inaccuracy of ICD-9 codes for chronic kidney disease: a study from two practice-based research networks (PBRNs). *The J. Am. Board Fam. Medicine* **28**, 678–682 (2015).

15. Dong, H. *et al.* Automated clinical coding: what, why, and where we are? *npj Digit. Medicine* **5**, 159 (2022).

16. Mehrafarin, H., Rajaee, S. & Pilehvar, M. T. On the importance of data size in probing fine-tuned models. In *Findings of the Association for Computational Linguistics: ACL 2022*, 228–238 (2022).

17. Lanera, C. *et al.* Use of machine learning techniques for case-detection of varicella zoster using routinely collected textual ambulatory records: pilot observational study. *JMIR Med. Informatics* **8**, e14330 (2020).

18. Lanera, C. *et al.* A deep learning approach to estimate the incidence of infectious disease cases for routinely collected ambulatory records: the example of varicella-zoster. *Int. J. Environ. Res. Public Heal.* **19**, 5959 (2022).

19. Sciannameo, V. *et al.* Deep learning for predicting urgent hospitalizations in elderly population using healthcare administrative databases. *SSRN pre-print 4022016* (2022).

20. Hammami, L. *et al.* Automated classification of cancer morphology from italian pathology reports using natural language processing techniques: a rule-based approach. *J. Biomed. Informatics* **116**, 103712 (2021).

21. Viani, N. *et al.* Information extraction from italian medical reports: an ontology-driven approach. *Int. J. Med. Informatics* **111**, 140–148 (2018).

22. Viani, N. *et al.* Supervised methods to extract clinical events from cardiology reports in italian. *J. Biomed. Informatics* **95**, 103219 (2019).

23. Johnson, A. E. *et al.* MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).

24. Johnson, A. E. *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).

25. Edin, J. *et al.* Automated medical coding on MIMIC-III and MIMIC-IV: a critical review and replicability study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2572–2582 (2023).

26. Li, F. & Yu, H. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 8180–8187 (2020).

27. Li, M. *et al.* Automated ICD-9 coding via a deep learning approach. *IEEE/ACM Transactions on Comput. Biol. Bioinforma.* **16**, 1193–1202 (2018).

28. Falter, M. *et al.* Using natural language processing for automated classification of disease and to identify misclassified ICD codes in cardiac disease. *Eur. Hear. Journal-Digital Heal.* **5**, 229–234 (2024).

29. Boytcheva, S. Automatic matching of ICD-10 codes to diagnoses in discharge letters. In *Proceedings of the Second Workshop on Biomedical Natural Language Processing*, 11–18 (2011).

30. Velichkov, B. *et al.* Automatic ICD-10 codes association to diagnosis: Bulgarian case. In *CSBio'20: Proceedings of the Eleventh International Conference on Computational Systems-Biology and Bioinformatics*, 46–53 (2020).

31. Bagheri, A. *et al.* Automatic ICD-10 classification of diseases from dutch discharge letters. In *In conjunction with the 13th International Joint Conference on Biomedical Engineering Systems and Technologies-BIOSTEC 2020* (2020).

32. van Leeuwen, J. R., Penne, E. L., Rabelink, T., Knevel, R. & Teng, Y. O. Using an artificial intelligence tool incorporating natural language processing to identify patients with a diagnosis of ANCA-associated vasculitis in electronic health records. *Comput. Biol. Medicine* **168**, 107757 (2024).

33. Lenc, L. *et al.* Czech medical coding assistant based on transformer networks. *Comput. Biol. Medicine* 108672 (2024).

34. Remmer, S., Lamproudis, A. & Dalianis, H. Multi-label diagnosis classification of swedish discharge summaries–ICD-10 code assignment using KB-BERT. In *International Conference Recent Advances in Natural Language Processing (RANLP'21), online, September 1-3, 2021*, 1158–1166 (INCOMA Ltd., 2021).

35. Sanyal, J., Rubin, D. & Banerjee, I. A weakly supervised model for the automated detection of adverse events using clinical notes. *J. Biomed. Informatics* **126**, 103969 (2022).

36. Wang, Y. *et al.* A clinical text classification paradigm using weak supervision and deep representation. *BMC Med. Informatics Decis. Mak.* **19**, 1 (2019).

37. Cusick, M. *et al.* Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *J. Psychiatr. Res.* **136**, 95–102 (2021).

38. Wang, S. Y. *et al.* Leveraging weak supervision to perform named entity recognition in electronic health records progress notes to identify the ophthalmology exam. *Int. J. Med. Informatics* **167**, 104864 (2022).

39. Humbert-Droz, M., Mukherjee, P., Gevaert, O. *et al.* Strategies to address the lack of labeled data for supervised machine learning training with electronic health records: case study for the extraction of symptoms from clinical notes. *JMIR Med. Informatics* **10**, e32903 (2022).

40. Greco, K. F. *et al.* A weakly supervised transformer for rare disease diagnosis and subphenotyping from EHRs with pulmonary case studies. *npj Digit. Medicine* (2026).

41. Yu, S. *et al.* Enabling phenotypic big data with PheNorm. *J. Am. Med. Informatics Assoc.* **25**, 54–60 (2018).

42. Ahuja, Y. *et al.* sureLDA: a multidisease automated phenotyping method for the electronic health record. *J. Am. Med. Informatics Assoc.* **27**, 1235–1243 (2020).

43. Halpern, Y., Horng, S., Choi, Y. & Sontag, D. Electronic medical record phenotyping using the anchor and learn framework. *J. Am. Med. Informatics Assoc.* **23**, 731–740 (2016).

44. Wagholikar, K. B., Estiri, H., Murphy, M. & Murphy, S. N. Polar labeling: silver standard algorithm for training disease classifiers. *Bioinformatics* **36**, 3200–3206 (2020).

45. Agarwal, V. *et al.* Learning statistical models of phenotypes using noisy labeled training data. *J. Am. Med. Informatics Assoc.* **23**, 1166–1173 (2016).

46. Smith, R., Fries, J. A., Hancock, B. & Bach, S. H. Language models in the loop: incorporating prompting into weak supervision. *ACM/JMS J. Data Sci.* **1**, 1–30 (2024).

47. Das, A., Talati, I. A., Chaves, J. M. Z., Rubin, D. & Banerjee, I. Weakly supervised language models for automated extraction of critical findings from radiology reports. *npj Digit. Medicine* **8**, 257 (2025).

48. De Ferrari, J., Ñanculef, R., Benoit, D., Araya, M. & Solar, M. Assessing GPT as a weak oracle for annotating radiological studies. In *International Conference on Artificial Intelligence in Medicine*, 98–109 (Springer, 2025).

49. Hall, C. B. *et al.* The burden of respiratory syncytial virus infection in young children. *New Engl. J. Medicine* **360**, 588–598 (2009).

50. Cantarutti, A. & Giaquinto, C. Pedianet database. In *Databases for Pharmacoepidemiological Research*, 159–164 (Springer, 2021).

51. Durani, Y., Friedman, M. J. & Attia, M. W. Clinical predictors of respiratory syncytial virus infection in children. *Pediatr. Int.* **50**, 352–355 (2008).

52. Gebremedhin, A. T., Hogan, A. B., Blyth, C. C., Glass, K. & Moore, H. C. Developing a prediction model to estimate the true burden of respiratory syncytial virus (RSV) in hospitalised children in western australia. *Sci. Reports* **12**, 332 (2022).

53. Vartiainen, P. *et al.* Risk factors for severe respiratory syncytial virus infection during the first year of life: development and validation of a clinical prediction model. *The Lancet Digit. Heal.* **5**, e821–e830 (2023).

54. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. In *ICLR Workshop*, 1–12 (2013).

55. Vaswani, A. *et al.* Attention is all you need. *Adv. Neural Inf. Process. Syst.* **30** (2017).

56. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186 (2019).

57. Liu, Y. *et al.* RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

58. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

59. Alsentzer, E. *et al.* Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78 (2019).

60. Tamburini, F. How "BERTology" changed the state-of-the-art also for italian NLP. *Comput. Linguist. CLiC-it 2020* 415 (2020).

61. Magnini, B., Altuna, B., Lavelli, A., Speranza, M. & Zanoli, R. The E3C project: collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics* (2020).

62. Campello, R. J., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172 (Springer, 2013).

63. Asyaky, M. S. & Mandala, R. Improving the performance of HDBSCAN on short text clustering by using word embedding and UMAP. In *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 1–6 (IEEE, 2021).

64. Yeasmin, S., Afrin, N. & Huq, M. R. Transformer-based text clustering for newspaper articles. In *International Conference on Machine Intelligence and Emerging Technologies*, 443–457 (Springer, 2022).

65. Limsopatham, N. Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, 210–216 (2021).

66. Polignano, M., Basile, P. & Semeraro, G. Advanced natural-based interaction for the italian language: LLaMAntino-3-ANITA. *arXiv preprint arXiv:2405.07101* (2024).

67. Conneau, A. *et al.* Unsupervised cross-lingual representation learning at scale. In Jurafsky, D., Chai, J., Schluter, N. & Tetreault, J. (eds.) *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451 (Association for Computational Linguistics, Online, 2020).

68. Chung, H. W., Fevry, T., Tsai, H., Johnson, M. & Ruder, S. Rethinking embedding coupling in pre-trained language models. *arXiv preprint arXiv:2010.12821* (2020).

69. Youden, W. J. Index for rating diagnostic tests. *Cancer* **3**, 32–35 (1950).

Acknowledgements

The present research has been supported by MUR, grant "Dipartimento di Eccellenza 2023-2027".

Francesca Ieva acknowledges the National Plan for NRRP Complementary Investments "Advanced Technologies for Human-centred Medicine" (PNC0000003).

Elisa Barbieri declares that her position is funded within the project “INF-ACT - One Health Basic and translational Research Actions addressing Unmet Needs on Emerging Infectious Diseases”, ID MUR PE_00000007, under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2 – CUP C93C22005170007.

The authors thank all the family paediatricians collaborating in Pedianet: Eva Alfieri, Michela Alfiero Bordigato, Angelo Alongi, Biagio Amoroso, Rosaria Ancarola, Barbara Andreola, Giampaolo Anese, Roberta Angelini, Maria Grazia Apostolo, Giovanna Argo, Giovanni Avarello, Lucia Azzoni, Maria Carolina Barbazza, Patrizia Barbieri, Gabriele Belluzzi, Eleonora Benetti, Filippo Biasci, Franca Boe, Stefano Bollettini, Francesco Bonaiuto, Anna Maria Bontempelli, Matteo Bonza, Sara Bozzetto, Andrea Bruna, Ivana Brusaterra, Massimo Caccini, Laura Calì, Sonia Camposilvan, Laura Cantalupi, Luigi Cantarutti, Chiara Cardarelli, Giovanna Carli, Sylvia Carnazza, Massimo Castaldo, Stefano Castelli, Monica Cavedagni, Giuseppe Egidio Cera, Chiara Chillemi, Francesca Cichello, Giuseppe Cicione, Carla Ciscato, Mariangela Clerici Schoeller, Samuele Cocchiola, Giuseppe Collacciani, Valeria Conte, Roberta Corro’, Rosaria Costagliola, Nicola Costanzo, Sandra Cozzani, Giancarlo Cuboni, Giorgia Curia, Caterina D’alia, Vito Francesco D’Amanti, Antonio D’Avino, Roberto De Clara, Lorenzo De Giovanni, Annamaria De Marchi, Gigliola Del Ponte, Tiziana Di Giampietro, Giuseppe Di Mauro, Giuseppe Di Santo, Piero Di Saverio, Mattea Dieli, Marco Dolci, Mattia Doria, Dania El Mazloum, Maria Carmen Fadda, Pietro Falco, Mario Fama, Marco Faraci, Maria Immacolata Farina, Tania Favilli, Mariagrazia Federico, Michele Felice, Maurizio Ferraiuolo, Michele Ferretti, Mauro Gabriele Ferretti, Paolo Forcina, Patrizia Foti, Luisa Freo, Ezio Frison, Fabrizio Fusco, Giovanni Gallo, Roberto Gallo, Andrea Galvagno, Alberta Gentili, Pierfrancesco Gentilucci, Giuliana Giampaolo, Francesco Gianfredi, Isabella Giuseppin, Laura Gnesi, Costantino Gobbi, Renza Granzon, Mauro Grelloni, Mirco Grugnetti, Antonina Isca, Urania Elisabetta Lagrasta, Maria Rosaria Letta, Giuseppe Lietti, Cinzia Lista, Ricciardo Lucantonio, Francesco Luise, Enrico Marano, Francesca Marine, Lorenzo Mariniello, Gabriella Marostica, Sergio Masotti, Stefano Meneghetti, Massimo Milani, Stella Vittoria Milone, Donatella Moggia, Angela Maria Monteleone, Pierangela Mussinu, Anna Naccari, Immacolata Naso, Flavia Nicoloso, Cristina Novarini, Laura Maria Olimpi, Riccardo Ongaro, Maria Maddalena Palma, Angela Pasinato, Andrea Passarella, Pasquale Pazzola, Monica Perin, Danilo Perri, Silvana Rosa Pescosolido, Giovanni Petrazzuoli, Giuseppe Petrotto, Patrizia Picco, Ambrogina Pirola, Lorena Pisanello, Daniele Pittarello, Elena Porro, Antonino Puma, Maria Paola Puocci, Andrea Righetti,

Rosaria Rizzari, Cristiano Rosafio, Paolo Rosas, Bruno Ruffato, Lucia Ruggieri, Annamaria Ruscitti, Annarita Russo, Pietro Salamone, Daniela Sambugaro, Luigi Saretta, Vittoria Sarno, Valentina Savio, Nico Maria Sciolla, Rossella Semenzato, Paolo Senesi, Carla Silvan, Giorgia Soldà, Valter Spanevello, Sabrina Spedale, Francesco Speranza, Sara Stefani, Francesco Storelli, Paolo Tambaro, Giacomo Toffol, Gabriele Tonelli, Silvia Tulone, Angelo Giuseppe Tummarello, Cristina Vallongo, Sergio Venditti, Maria Grazia Vitale, Concetta Volpe, Francescopaolo Volpe, Aldo Vozzi, Giulia Zanon, Maria Luisa Zuccolo.

Author contributions

V.T.: Methodology, Software, Investigation, Writing - Original Draft E.B.: Conceptualization, Data Curation, Writing - Review & Editing A.C.: Conceptualization, Data Curation, Writing - Review & Editing C.G.: Conceptualization, Supervision F.I.: Conceptualization, Methodology, Writing - Review & Editing, Supervision. All authors reviewed and approved the manuscript.

Data availability

Data used in this study are not publicly available due to their confidential nature. Data are however available upon reasonable request to the corresponding author, subject to approval by the Internal Scientific Committee of Società Servizi Telematici Srl, the legal owner of Pedianet.
The code developed for the analysis is available at
https://github.com/vittot/weakly-supervised-classification-italian-discharge-letters.

Funding

The present research has been supported by MUR, grant "Dipartimento di Eccellenza 2023-2027".

Francesca Ieva acknowledges the National Plan for NRRP Complementary Investments "Advanced Technologies for Human-centred Medicine" (PNC0000003).

Elisa Barbieri declares that her position is funded within the project “INF-ACT - One Health Basic and translational Research Actions addressing Unmet Needs on Emerging Infectious Diseases”, ID MUR PE_00000007, under the National Recovery and Resilience Plan (NRRP), Mission 4, Component 2 – CUP C93C22005170007.

Additional information

Competing interests The authors declare no competing interests.