# UTP: A Unified Term Presentation tool for clincial textual data using pattern-matching rules and dictionary-based ontologies

Monah Bou Hatoum[1], Jean Claude Charr[1], Alia Ghaddar[2], Christophe Guyeux[1], and David Laiymani[1]

[1] University of Franche-Comte, 90000 Belfort, France
[2] Department of Computer Science, International University of Beirut, Beirut P.O. Box 146404, Lebanon

**Abstract.** Clinical textual data such as discharge summaries and chief complaints summarize the patient's medical history and treatment plan. These unstructured complex data include ambiguous medical terms, abbreviations, diagnostic investigation values and dates which pose significant challenges for human and machine learning tasks to process them. This paper proposes a novel approach that transforms clinical text with different writing styles into a uniform and standard presentation using pattern-matching rules and JSON dictionary-based ontologies. The main goal of the proposed approach is to improve the communication between healthcare parties or professionals by improving the quality of the clinical textual data and reducing its heterogeneity and ambiguity. In addition, this data quality improvement enhances the performance of machine learning downstream tasks. Our approach identifies the abbreviations, medical terms, negations, dates, and investigation values from the unstructured textual data. Then, it replaces the detected entities with their corresponding unified and normalized presentation based on pattern-matching rules that relies on the linguistic features, pattern-matching rules, and JSON dictionaries. The inductive content analysis method was followed to generate the pattern-matching rules with the help of a medical team. Its role is to validate the accuracy of the detected entities. Finally, the proposed approach was applied to a massive real-world dataset in order to evaluate its impact on the performance of various machine learning models. The results show a significant improvement in performance after preprocessing the clinical textual data using our approach.

**Keywords:** Deep Learning · Natural Language Processing (NLP) · Computer-Aid Diagnosis · Chief Complaints · Text mining · Abbreviations · Medical Phrases · ChatGPT 4

## 1 Introduction

Clinical discharge summaries are crucial documents that comprehensively summarize the patient's medical history and treatment plan. These summaries serve as a vital source of information for healthcare providers, helping them make informed decisions about patient care. However, using different medical terms, abbreviations, and ways of presenting diagnostic investigation values and dates [19] in the discharge summary poses significant challenges for human and machine learning tasks [12,25,11].

One of the significant challenges in clinical discharge summaries is using different medical terms which refer to the same disease or health condition. For Example, a sudden coronary artery blockage that causes the heart muscle to stop beating can be called a *heart attack*, *myocardial infarction*, *coronary artery disease*, and *cardiac arrest*. Table 1 shows more examples of diseases with many synonyms. This variety of medical terms can be confusing for non-medical personnel and machine learning models because they sound like they describe different things. The medical staff should know the importance of using the standard vocabulary when communicating with non-medical personnel and patients to avoid misinterpretation [18,14]. Moreover, machine learning algorithms usually look for patterns in the data to

| **Medical Term** | **Synonyms** |
|---|---|
| Dyspnea | Shortness of breath, difficulty breathing |
| Edema | Swelling, Fluid retention |
| Hypercholesterolemia | High cholesterol, High blood cholesterol |
| Hypertension | High blood pressure, Elevated blood pressure high BP, HBP, HTN |
| Hypoglycemia | Low blood sugar, Low blood glucose |
| Hyperglycemia | High blood sugar, High blood glucose |
| Epistaxis | Nosebleed, Nasal bleeding |
| Conjunctivitis | Pink eye, Conjunctival inflammation |
| HyperPyrexia | Fever |
| Rhinorrhea | Runny nose |

Table 1. Example of disorders that have different synonyms.

classify them. Using different synonyms for the same diseases can make it more difficult for machine learning models to identify these patterns [12,1,16].

Another challenge in clinical discharge summaries is the use of abbreviations. Abbreviations are commonly used in clinical documentation to save time and space. However, many abbreviations could be ambiguous, especially when the same abbreviation has different meanings in different medical specialties [26,21]. For Example, *MS* refers to *Multiple Sclerosis* in the Neurology department and *Mitral Stenosis* in the Cardiology and Radiology departments [8]. In addition, some abbreviations use different short terms with the same meaning. For Example, *HTN* refers to *hypertension* and *HBP* refers to *high blood pressure*, which complicates the learning of patterns' identification for machine learning models.

Furthermore, presenting diagnostic investigation values and dates in discharge summaries also presents a significant challenge. Investigation values are numerical data that provide critical information about a patient's health. However, how these values are presented in the discharge summary can vary depending on the physician. For instance, some physicians may present the values in a numeric form while others may use descriptive terms such as *normal*, *elevated*, or *high*. For example, a *RBC: 4.3* for a female patient is considered as a *normal Red Blood Cell count*, whereas it is low for a male patient. These numerical values might confuse the medical staff and could lead to misinterpretations.

In addition, physicians might write the date formats in various ways, as shown in Table 2. Using dates inside the discharge summary is usually related to the episode of care or some past medical history. Machine learning tasks cannot understand the context of dates presented in the discharge summary. One way to address this problem is to convert the dates in the text into periods like *two days before* or *one week before* based on the detected date and the encounter date (visit date). This workaround will make it easier for machines to understand the dates' context and identify patterns in the data.

| Date Format | Example |
|---|---|
| dd MMM yyyy | 14 Feb 2023, 23 Mar 2023 |
| dd/MM/yyyy | 14/02/2023, 23/03/2023 |
| dd-MM-yyyy | 14-02-2023, 23-02-2023 |
| yyyy-MM-dd | 2023-02-14, 2023-02-23 |
| dd.MM.yyyy | 14.02.2023, 23.02.2023 |
| Period | before one week, after two days |

Table 2. An example of some different date formats physicians use in clinical textual data.

This paper proposes a novel "Unified Term Presentation" (UTP) approach for processing clinical textual data that transforms them into an easily readable, less complex, and ambiguous unified presentation. UTP is based on pattern-matching rules and dictionary-based ontologies for term recognition and uses domain-specific knowledge to transform clinical texts into a unified representation. UTP applies several changes to the input unstructured textual data by transforming all dates with their different formats into periods, replacing all the abbreviations with their complete expanded forms, replacing all the diagnostic test values with categorical values based on the normal ranges for the detected diagnostic terms and unifying the negation terms. UTP improves the quality of care by ensuring that all clinicians' data are standard and use the same terminology. In addition, Machine-learning algorithms can then use this representation to generate more accurate predictions. The main contributions of this paper on clinical textual data transformation are two-fold:

- Providing a novel approach (tool) that healthcare providers can use to unify the input data from physicians and apply data preprocessing for machine learning tasks.
- Providing an empirical study on a massive real-world dataset with over 9.57M records before and after applying our approach.

The paper is organized as follows: Section 2 presents a general overview of the data preprocessing paradigm. Section 3 provides in-depth information on the Unified Transformation Presentation model. Section 4 shows the experimental setup and results. Section 5 discusses our findings and the approach limitations. The paper ends with a conclusion and some perspectives.

## 2 Related Works

Research about clinical text mining has gained more significant interest recently. The development of machine learning and deep learning techniques helped address many complicated healthcare problems, such as extracting medical terms and valued information from clinical textual data. This section provides a general overview of some existing clinical textual data preprocessing methods.

### 2.1 Feature Extraction

Clinical Feature Extraction (CFE) and Clinical Named Entity Recognition (CNER) are NLP methods for extracting relevant and entity information from clinical textual data. Feature extraction methods are vital in improving the performance of machine learning training tasks. It eliminates the non-relevant data from the corpus and reduces the vocabulary size. Feature extraction is a widely explored topic, and there are many tools with impressive performance, such as *SciSpacy* [22], *Med-Flair* [7], and *CT-BERT* [15]. However, the amount of noise in the data and the data diversity significantly impact the performance of the CFE and CNER methods.

### 2.2 Deep Learning techniques

In recent years, NLP using Deep Learning techniques has achieved impressive results, especially with transformer-based technologies [24]. *BERT* (Bidirectional Encoder Representations from Transformers) [6] is a natural language processing model. It generates contextualized word embeddings and was trained on language modeling and the next sentence prediction tasks to generate a pre-trained model. It was published in 2018 and achieved state-of-the-art performance in NLP tasks. Currently, *BERT* pre-trained models are used for transfer learning by fine-tuning these models on specific domain datasets. *Clinical BERT* [2] and *BioBERT* [13] are two examples of transfer learning. However, Some studies showed that the non-contextualized techniques surprisingly performed better than *BERT* in industrial datasets [3].

### 2.3 Existing Solutions

**Natural Language Processing Tools** Several popular NLP libraries are used for different purposes. *spaCy* [10] is a widely used NLP library known for its speed and accuracy. It uses statistical and machine learning techniques to perform tasks such as named entity recognition, part-of-speech tagging, and dependency parsing. *scispaCy* [22] contains set pretrained models on medical datasets that were built on top of *spaCy*. On the other hand, *coreNLP* [17] is a Java-based NLP library that provides features like tokenization, POS tagging, and named entity recognition. *NLTK* [4] is another popular NLP library based on Python and provides features like tokenization, POS tagging, named entity recognition, and sentiment analysis. These NLP tools use linguistic features and pattern-matching rules to process text. However, they may have some limitations when handling complex and nuanced language used in clinical settings and keeping up with the rapid expansion of medical terminology.

**ChatGPT 4** *ChatGPT 4* is an AI chatbot developed by OpenAI. It can carry on open-ended conversations, summarize factual topics, and create stories. It is more accurate and informative than previous chatbots and has the potential to be used in a variety of applications, such as customer service, education, and entertainment. The main challenges to AI chatbots adoption are safety, trust, and cost. ChatGPT is a promising new development in AI chatbots, but these challenges must be addressed before it can be used to preprocess clinical textual data.

**Rule-Based approach** Authors in [8] provided a cleansing approach called *EMTE* that removes the irrelevant data from the clinical textual data using rule-based pattern rules. Their approach replaces the abbreviations with their expanded form and detects negations and medical terms. Then, they concatenate the tokens in the detected medical term using underscores. As a result, they were able to reduce the vocabulary size, and they achieved an F1-score of 69.68%. However, their approach have some limitations. For example, their approach did not tackle the different medical terms that have same meaning, which can be reduced more. Also, concatenating medical terms does not effectively reduce the vocabulary size. For example, as shown in Figure 1, there are nine different possible generated vocabularies for two medical terms (*high blood pressure*, *low blood pressure*), where it could be only four vocabularies (*high*, *low*,*pressure*, and *blood*) if the medical term concatenation was not used.

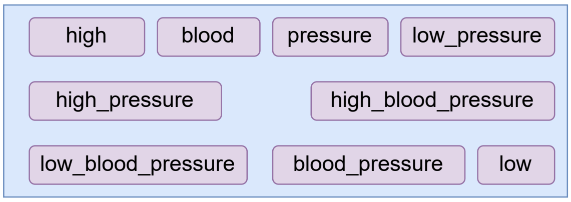

Figure 1: An example of the possible generated vocabularies from two medical terms (*high blood pressure* and *low blood pressure*) using *EMTE*

## 3 Model

We aim to provide a tool that removes confusion and prevents misinterpretation by nonspecialized physicians (such as medical staff, technicians, and patients) when reading clinical textual documents. This confusion can arise due to the extensive use of abbreviations by physicians from different specialties, complex medical terms that might be unfamiliar to other medical staff, and investigation values embedded with the clinical textual data. We propose a tool called *UTP* (Unified Term Presentation) that transforms complex clinical textual data into more readable and easy-to-understand medical textual data. *UTP* detects and converts the diagnostic values into categorical ranges based on the requested investigation and the patient demographical data. The categorical values contain (below range, within normal range, above range, negative, and positive). Also, *UTP* expands the detected abbreviations into their full form based on the physician's specialty. Moreover, *UTP* transforms complex medical terms into more general and readable terms by other nonspecialized individuals. In addition, *UTP* can transform the detected dates in the clinical textual data into a period presentation based on the visit date information, reducing the complexity of the machine-learning tasks.

Figure 2 shows the architecture of the *UTP* tool that consists of four processing components: *Date to Period*, *Investigation values*, *Abbreviations*, and *Medical Terms*. All these components are customized pipes added to the *spaCy* [10] tool. They override the existing *NER*b(Named Entity Recognition) of *spaCy*. The *UTP* tool employs the existing features in *spaCy* tool, like the linguistic features (*POS*,*TAG*, and *DEP*) and the pattern-matching rules engine in both detection and transformation processes.

This section presents in detail every feature of UTP and their importance for human and machine-learning readability and their impact on vocabulary size reduction while mitigating the term-learning confusion (e.g. using one unified term *high blood pressure* instead of multiple different terms like *hypertension*, *high blood pressure*, and *HTN*).

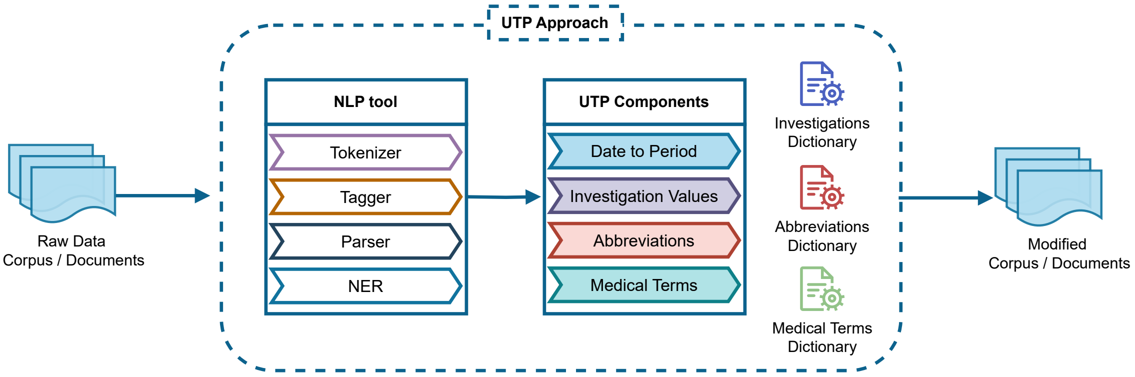

Figure 2: The preprocessing proposed approach *UTP*

### 3.1 Date to Period transformation

Physicians extensively use dates in their medical documents to explain the chronological development of the patient's care during admissions. It is also very important in outpatient departments, especially in the obstetrics and gynecology department where the "last menstrual period" (*LMP*) and "estimated due date" (*EDD*) play important roles during a patient's pregnancy and maternity. However, having dates without any reference in the clinical textual data is confusing for machine learning tasks, especially when no information is provided on the patient's visit date. Also, the patient's visit date is confidential and should not be shared. Therefore, transforming the dates found in clinical textual data into a period presentation is essential since it preserves patient confidentiality and helps reducing vocabulary size and confusion for the training tasks.

*UTP* detects different date styles found in the corpus and transforms the dates into periods compared to the document's effective date. For example, a patient visited the obstetrics and gynecology department on the $19^{th}$ of June 2022, and the physician wrote the following chief complaint (LMP: 01 Jun 2022, EDD: 08/03/2023). *UTP* converts this data into (LMP: before 2 weeks and 5 days, EDD: after 37 weeks and 3 days) since the obstetrics and gynecology department physicians prefer to track the pregnancy in weeks. Also, *UTP* can present the period in months like (EDD: 8 months, two weeks and 3 days).

### 3.2 Abbreviations and Investigation values transofrmation

The abbreviations used in clinical textual data can be either medical disorders (*ADD*, which stands for Attention deficit disorder), procedures (*SVD*, which stands for Spontaneous vaginal delivery), investigation results (*Hbg: 11.9*, *Hbg* stands for Hemoglobin), and general abbreviations (*hx*, which stands fro History of, *rx*, which stands for prescription) that have no restrictions for usage unlike other types.

**Investigation values** Investigation value is a combination of an abbreviation with a value. The abbreviation represents a laboratory or imaging test, and the value is the test result. Every test has a normal range that helps the physicians diagnose if the patient's test result is outside this normal range. The normal ranges are very critical since they strictly depend on the machines used in the given hospital. Every machine has its configuration, and every test has a different normal range based on the patient demographical data and the used standard units of measure. For example, normal range values of the "High-sensitive cardiac troponin (hs-cTn)" laboratory test should be less than 14 ng/ml or 4000 pg/ml. So, the value *hs-cTn: 150* is considered very high if the unit of measure is "ng/ml," and it is normal if the ussed unit of measure in the machine was "pg/ml".

*UTP* uses a JSON-based dictionary containing all the hospital's laboratory and imaging tests. Every entry contains the following information (abbreviation, fully expanded form, list of normal ranges based on the age restriction and gender restriction). *UTP* automatically builds pattern-matching rules using *SpaCy* to detect the investigation abbreviations and the values that come after these abbreviations. Then, based on the normal range of the detected investigation, *UTP* transforms the abbreviation into its expanded form and the value into a categorical value, leading to a unified and readable format. For example, *UTP* transforms *"Hbg: 11.9"* into *"hemoglobin below range"* for a male sample and into *"hemoglobin within normal range"* for a female sample.

**Abbreviations transformation** Working with the disorders and procedural abbreviations is ambiguous since the same abbreviation might have different meanings in different specialties. *UTP* uses a JSON-based dictionary that contains a list of abbreviations used in the considered hospital with the following information structure (abbreviation, fully expanded form, list of specialties). The list of specialties can be empty, which means the given abbreviation can be used in all departments without any confusion. *UTP* transforms the abbreviations into their expanded form by eliminating all abbreviations unrelated to the given sample's specialty, which mitigates the abbreviation ambiguity. For example, *UTP* transforms *"MS"* into *"multiple sclerosis"* for a sample from the neurology department and into *"mitral stenosis"* for a sample from the cardiology specialty.

### 3.3 Medical terms transofrmation

Physicians extensively use medical disorders' scientific names, making it difficult to understand by nonspecialized individuals like para-medical, nurses, and patients. *UTP* relies on both linguistic features and pattern-matching rules using the *SpaCy* tool to detect the medical terms loaded from a JSON-based dictionary that stores medical terms with their synonyms and the preferred term to be used. The JSON-based dictionary was built by extracting the medical terms with their synonyms from the *snomed-CT* database, then validated by a medical team to ensure its accuracy and properly identify the preferred term to be used.

Figure 3 shows how our approach *UTP* detects the the entities from a clinical textual data. The image shows the raw data, the detected entities, and the transformation results. There are different types of detected entities: *ABBR* stands for *Abbreviations*, *LAB_VALUE* stands for *Investigation values*, *MEDICAL* stands for *Medical Terms*, *PERIOD* stands for the *DATES*, and *NEG* stands for *NEGATIONS*. *UTP* was able to transform the detected entities into more readable clinical textual data. For example, the detected abbreviation *ga* was correctly transformed into *gestational age* depending on the specialty of the sample. the medical term *dysuria* was transformed into *painful urination*, which is more readable by nonspecialized individuals. Also, *UTP* was able to detect the investigation values *Hbg 12* and *wbc 14* correctly as *LAB_VALUE* entities. Then, *UTP* correctly transformed them based on the gender of the patient into *hemoglobin normal range* since the range for adult females is between 12 and 14. Moreover, the *wbc 14* was correctly transformed into *white blood cell count above range* since the range for adult females is between 4 and 11. Furthermore, *UTP* transformed the dates based on the encounter date into the period as depicted in Figure 3 .

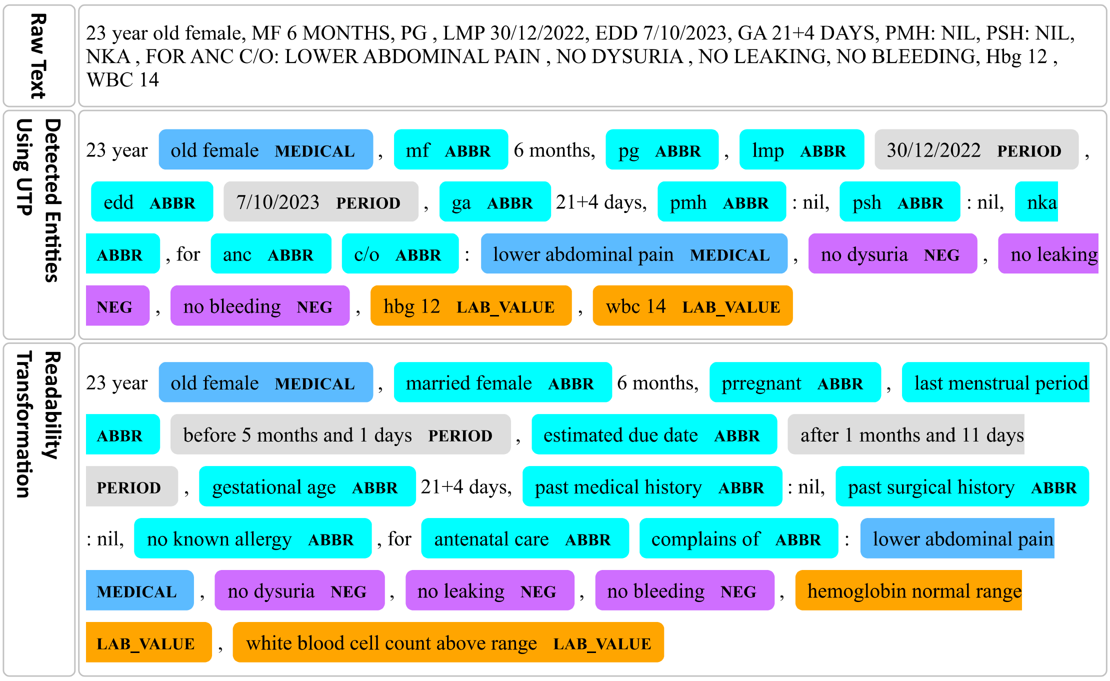
Raw Text
23 year old female, MF 6 MONTHS, PG , LMP 30/12/2022, EDD 7/10/2023, GA 21+4 DAYS, PMH: NIL, PSH: NIL, NKA , FOR ANC C/O: LOWER ABDOMINAL PAIN , NO DYSURIA , NO LEAKING, NO BLEEDING, Hbg 12 , WBC 14
Detected Entities Using UTP
23 year old female MEDICAL , mf ABBR 6 months, pg ABBR , lmp ABBR 30/12/2022 PERIOD , edd ABBR 7/10/2023 PERIOD , ga ABBR 21+4 days, pmh ABBR : nil, psh ABBR : nil, nka ABBR , for anc ABBR c/o ABBR : lower abdominal pain MEDICAL , no dysuria NEG , no leaking NEG , no bleeding NEG , hbg 12 LAB_VALUE , wbc 14 LAB_VALUE
Readability Transformation
23 year old female MEDICAL , married female ABBR 6 months, prregnant ABBR , last menstrual period ABBR before 5 months and 1 days PERIOD , estimated due date ABBR after 1 months and 11 days PERIOD , gestational age ABBR 21+4 days, past medical history ABBR : nil, past surgical history ABBR : nil, no known allergy ABBR , for antenatal care ABBR complains of ABBR : lower abdominal pain MEDICAL , no dysuria NEG , no leaking NEG , no bleeding NEG , hemoglobin normal range LAB_VALUE , white blood cell count above range LAB_VALUE

Figure 3: An example from Obstetrics & Gynecology department of how the *UTP* tool detects and transform a clinical textual data into more readable clinical textual data.

## 4 Experiments and Results

The proposed approach improves the readability of clinical textual data by transforming and converting the terms that are potentially causing conflicts either when communicating between professionals in healthcare institutes or when working on machine learning tasks. To evaluate the performance of this approach, several experiments were conducted to solve a multilabel classification problem for ICD-10 prediction on clinical textual data preprocessed by *UTP*. This section represents the experimental setup with their results.

### 4.1 Experiment setup

ICD-10 prediction is a well-known multilabel classification problem where the input data are the textual data (chief complaints, history of present illness, discharge summaries), and the ICD-10 codes are the labels. ICD-10 codes are hierarchical alphanumeric codes [20,5,23]. The number of ICD-10 codes used by hospitals differs depending on the available covered specialties. For example, in dental clinics, physicians mainly use codes in the range [K00 -

K14] that cover teeth and jaw problems. While in the Obstetrics & Gynecology department, the physicians mainly use the codes that start with *O* that cover delivery-related cases and *Z34*, Z35** that cover pregnancy-related cases. Also, there are some age and gender restrictions in the ICD-10 codes. For example, all codes that start with *P* are allowed only for newborn cases.

### 4.2 Medical Dataset

The dataset, used in the experiments, was retrieved from a private Saudi hospital and consists of over 9.6 M records with over 3,100 ICD-10 codes. The dataset contained data from 24 different specialties. Figure 4 shows the different specialties with their relative proportions. Data from the Internal medicine department makes up to 23.7% of the total data samples while data from Obstetrics & Gynecology department makes up to 13.2% which reflects how much the data is imbalanced. It is worth mentioning that the data collected from the hospital was anonymized.

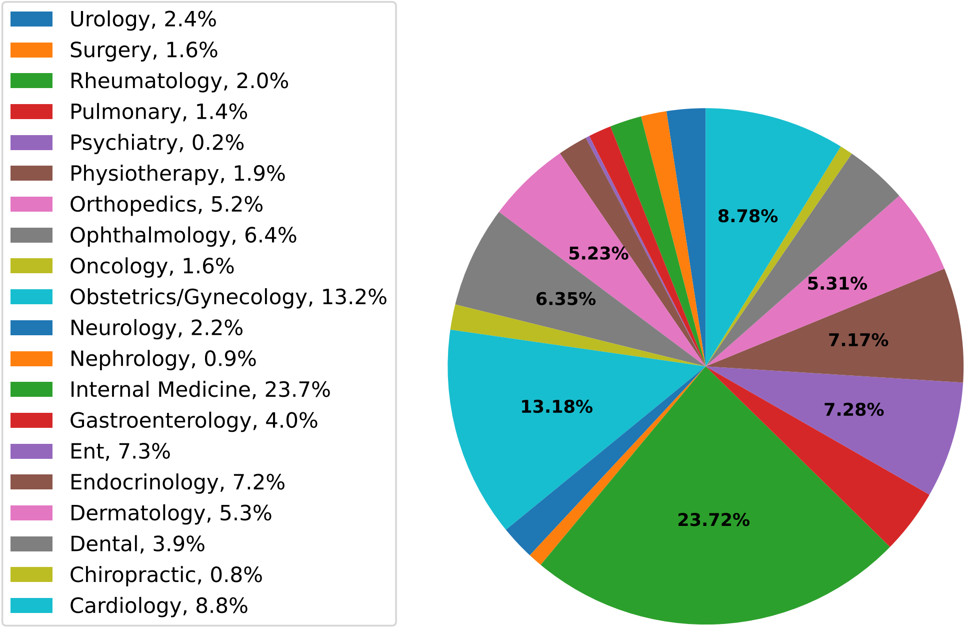

Figure 4: The chart shows the relative proportion of the 24 specialties.

### 4.3 JSON dictionaries

**Abbreviation dictionary** In most hospitals, the medical staff is required to follow standards and regulations when documenting patients' information. One such standard is to avoid using ambiguous abbreviations, such as *LFT* that can stands for either *Lung Function Test* or *Liver Function Test* [9]. To avoid confustion, hospitals issue an internal standard list of allowed common and specialty specific abbreviations. It is worth noting that some abbreviations might have different meanings in different hospitals or specialties and departments depending on the defined list of standard abbreviations and conventions. Therefore, it is important to transform the abbreviations into their expanded form when communicating with other institutes, such as Healthcare Isurance Companies. For the sake of the experiment, we built the abbreviation JSON dictionary while taking into consideration the internal policy of the hospital regarding the allowed abbreviation usage. The structure of the dictionary is similar to the one used in [8].

**Diagnostic Values** Diagnostic tests help physicians to diagnose a patient's condition or monitor the progression of the disease. The test results help physicians to manage the patient's medical problems, if any. The test results are either numerical values (measurements) or categorical (findings). For example, blood tests can measure the amount, size, and concentration of different cells and substances in the blood. Most of these laboratory tests are reported with numerical values compared to standard range values (also known as normal range values or reference values). Moreover, these standard values setups directly depends of the used machine in the laboratory or radiology department.

On the other hand, pregnancy tests detect the presence of the hormone (human chorionic gonadotropin (hCG)) in women's urine or blood. The result of such a test is either positive (indicating pregnancy) or negative (indicating no pregnancy). We focused on tests that have numerical values since these values might vary from one patient to another depending on the gender and age of the patient and other conditions. The standard normal values for the diagnostic tests were retrieved from the hospital's database and were stored in a JSON-based dictionary.

**Medical terms synonyms** A custom *medical_term_detector* was built using the spaCy python library and the linguistic features (DEP, POS, TAG) that detect and annotate the possible medical terms. Then, the unique discovered medical terms were extracted and added to a new corpus that contains only medical terms. Afterwards, the available *snomedct* database from the UMLS website was used to extract all the available synonyms for each term in the new corpus. Then, the inductive content analysis method was followed to remove all found words from the generated corpus of unique medical terms. The resulting medical terms are those that contain typos, those that have no synonyms, and those that have synonyms not listed in the *snomedct* database. With the help of a medical team, a total of 23,671 medical terms were checked and validated.

### 4.4 Machine Learning Task

The *Clinical_BERT* and *BioBERT* pretrained models were used as a baseline to compare the performance of the ICD-10 prediction classification task before preprocessing the clinical textual data using our approach *UTP* and after. The experiments setup was using *Clinical_BERT* and *BioBERT* as an embedding layer, in addition to a classifier layer. Then both models were fine-tuned using two corpora (Raw Data corpus, Data processed corpus using *UTP*).

In assessing the efficacy of the Unified Term Presentation (*UTP*) method, Table 3 offers crucial insights by showcasing a substantial reduction in vocabulary complexity. Before *UTP* implementation, the corpus comprised 553,712 unique terms, which, after applying *UTP*'s normalization techniques, including date conversions, categorization of diagnostic values, abbreviation expansions, and medical term standardization, was significantly reduced to 310,514 terms. This reduction underscores *UTP*'s role in enhancing data readability and simplification, which is essential for improving machine learning models' interpretability and performance in clinical applications.

Table 4 shows the hyperparameters used in the experiments. Moreover, the data was split into a training dataset with 70% of the data, a validation dataset with 10%, and a testing dataset with 20%. All experiments were run on *Colab Pro+* from Google. The data was processed and tokenized locally using the hospital servers, and the training tasks were done on *Colab Pro+* using the numerical data and labels. The data was converted from text to numerical values using *BERTTokenizer*. Moreover, *Colab Pro+* has a maximum of 24 hours execution time allowed. This limitation was overcomed using a 12-folds cross-validation approach to avoid repeating the whole experiment.

| Corpus | Unique Words | Change | Total Change |
|---|---|---|---|
| Raw Data | 553,712 | - | - |
| Dates to period | 482,124 | -12.93% | -12.93% |
| Diagnostic values to categorical | 434,002 | -9.98% | -21.62% |
| Abbreviations to expanded form | 359,891 | -17.08% | -35.00% |
| Medical terms normalization | 310,514 | -13.72% | -43.92% |

Table 3. The steps of the UTP approach on a corpus which contains over 9.57M samples with 553,712 unique words.

| | |
|---|---|
| Learning Rate | 2e-5 |
| Batch Size | 32 |
| Epsilon | 1e-8 |
| K-fold | 12 |
| Optimizer | Adam |
| Loss function | Binary cross entropy |

Table 4. *Clinical_BERT* and *BioBERT* models hyperparameters.

| F1-Micro scores | | | | | | |
|---|---|---|---|---|---|---|
| Department | Raw Data | Date | Abbreviation | Investigation Values | Medical Terms | UTP |
| Obstetrics & Gyneocology | 85.36% | 89.10% | 85.98% | 86.15% | 87.01% | 91.05% |
| Pediatrics | 81.78% | 83.39% | 82.24% | 82.44% | 83.33% | 84.74% |
| Emergency Room | 39.74% | 39.85% | 40.49% | 40.84% | 40.98% | 41.12% |
| Internal Medicine | 47.21% | 47.22% | 48.13% | 48.21% | 48.57% | 48.68% |
| F1-Micro gain score using gain% = (After - Before) / Before | | | | | | |
| Department | Raw Data | Date | Abbreviation | Investigation Values | Medical Terms | UTP |
| Obstetrics & Gyneocology | – | 4.38% | 0.73% | 0.93% | 1.93% | 6.67% |
| Pediatrics | – | 1.97% | 0.56% | 0.81% | 1.90% | 3.62% |
| Emergency Room | – | 0.28% | 1.89% | 2.77% | 3.12% | 3.47% |
| Internal Medicine | – | 0.02% | 1.95% | 2.12% | 2.88% | 3.11% |

Table 5. The f1-micro scores when applying the various individual transformations on data from four different departments.

In order to know the contribution of every individual transformation (dates, investigations, abbreviations, and medical terms) of *UTP*, several experiments were conducted to measure the f1-micro of the raw data compared to an individual transformation.

Table 5 shows the contributions made by individual transformation on four different specialties datasets using f1-micro score performance value and the gain value. The results show that the *Date* transformation improved the f1-micro score by *4.38%* and *1.79%* on data from the *"Obstetrics & Gynecology"* and *"Pediatrics"* departments, respectively. In contrast, the contribution of the *Date* transformation was low on data from the *"Emergency Room"* and *"Internal Medicine"* departments. The reason behind the low contribution is due to the fact that dates are not expensively used in these departments compared to the *"Obstetrics & Gynecology"* and *"Pediatrics"* departments.

Table 6 presents the results of the experiments of supervised learning on the multi-label classification problem using a large real-world dataset. The conducted experiments used two different BERT models *Clinical BERT* and *BIO BERT*. The results of the ICD-10 predictions from clinical textual data results are shown for the *Training*, *Validation*, and *Testing* datasets. The columns represents the recall, and F1-score (micro and weighted) evaluation metrics. The experiments that used the corpus modified by *UTP* outperformed the results of the raw data corpus for all the considered metrics. For example, the *UTP+Bio BERT* experiment gave *micro-F1* score around *74.64 ± 2.28 e-03* for the for from the test dataset compared to *68.11 ± 2.17 e-03* achieved by *Bio BERT*. Similarly, the *UTP+ Clinical BERT* achieved a *micro-F1* score around *73.25 ± 2.11 e-03* compared to the *67.16 ± 2.18 e-03* score achieved by *Clinical BERT*.

| Training Results | | | |
|---|---|---|---|
| Experiments | Recall | micro-F1 | weighted-F1 |
| Clinical BERT | 65.51 ± 8.66 e-03 | 76.89 ± 7.92 e-03 | 73.14 ± 9.18 e-03 |
| Bio BERT | 65.21 ± 8.31 e-03 | 76.47 ± 8.06 e-03 | 73.05 ± 8.03 e-03 |
| UTP + Clinical BERT | 72.88 ± 6.57 e-03 | 84.03 ± 6.18 e-03 | 84.94 ± 7.13 e-03 |
| UTP + Bio BERT | 72.19 ± 6.12 e-03 | 83.75 ± 5.81 e-03 | 84.31 ± 6.52 e-03 |
| Validation Results | | | |
| Experiments | Recall | micro-F1 | weighted-F1 |
| Clinical BERT | 58.64 ± 6.81 e-03 | 67.92 ± 7.32 e-03 | 65.20 ± 6.19 e-03 |
| Bio BERT | 59.13 ± 6.07 e-03 | 68.38 ± 6.82 e-03 | 65.71 ± 6.64 e-03 |
| UTP + Clinical BERT | 66.87 ± 5.78 e-03 | 73.77 ± 5.98 e-03 | 71.01 ± 5.01 e-03 |
| UTP + Bio BERT | 67.33 ± 5.06 e-03 | 74.32 ± 5.79 e-03 | 71.33 ± 5.77 e-03 |
| Testing Results | | | |
| Experiments | Recall | micro-F1 | weighted-F1 |
| Clinical BERT | 58.07 ± 3.43 e-03 | 67.16 ± 2.18 e-03 | 65.95 ± 2.81 e-03 |
| Bio BERT | 59.97 ± 3.31 e-03 | 68.11 ± 2.17 e-03 | 65.43 ± 2.80 e-03 |
| UTP+ Clinical BERT | 66.08 ± 2.21 e-03 | 73.25 ± 2.11 e-03 | 71.09 ± 2.09 e-03 |
| UTP+ Bio BERT | 67.83 ± 2.88 e-03 | 74.64 ± 2.28 e-03 | 72.01 ± 2.20 e-03 |

Table 6. The result of two well-known specific domain BERT models on the large real-world dataset before and after applying *UTP*.

### 4.5 Conmparison with ChatGPT 4 tool

| Experiments | Accuracy | Recall | Micro-F1 | Weighted-F1 |
|---|---|---|---|---|
| Raw | 91.43% | 82.84% | 90.56% | 88.95% |
| UTP | 92.16% | 83.13% | 91.05% | 89.27% |
| ChatGPT | 92.29% | 83.75% | 91.32% | 89.70% |

Table 7. Comparison between *UTP* and *ChatGPT 4* while preprocessing 5,000 chief complaints form the Obstetrics & Gynecology department.

In this subsection, the *UTP* approach, built using pattern-matching rules and JSON-based dictionaries, is compared to *ChatGPT 4* that can do the same work using machine learning techniques. The comparison with *ChatGPT 4* was conducted on a dataset consisting of 5000 samples with 7 ICD-10 codes labels from the *"Obstetrics & Gynecology"* department.

This department was chosen because its physicians frequently use dates, investigation values, abbreviations, and medical terms in their clinical data. In order to preserve the patient's confidential medical data while using *ChatGPT 4*, all the sensitive information was removed from the dataset. The updated visit date was added to the textual data with a tag called *"reference_date"*. Then, the *ChatGPT 4* API was used to prompt the following questions : *"Change the dates found in the text into periods, change the investigation values into categorical values as normal, below range, and above range. Finally, convert the complex medical terms into a more readable format that patients can easily read"*. *ChatGPT 4* results were very impressive. It converted most of the samples on the first try. Also, it included additional sentences to better explain the chief complaints. However, some issues were detected during the experiment that are discussed in the next section.

Table 7 compares the results obtained while using *ClinicalBERT* on *Raw Data* (without any modification on the clinical textual data), preprocessed data using *UTP* and preprocessed data using *ChatGPT 4*. The results show that *ChatGPT 4* gives slightly better results than *UTP*. *ChatGPT 4* did not only replace the terms in the clinical textual data, but also rephrased the sentences with different words. Sentence rephrasing is an impressive capabilities that *ChatGPT 4* can accomplish. However, in some samples, the *ChatGPT 4* included bias and irrelevant information to the patient's case, which gives nonrealistic diagnosis that might affect the patient's care if this tool was used in production.

## 5 Discussion

The *UTP* tool transforms medical textual data into more readable data by humans and machine learning algorithms. Moreover, unlike the pre-trained models, *UTP* does not require retraining or additional resources (Memory and GPU/TPU). Retraining the deep learning models is a major challenge in real-world healthcare applications since it requires a large number of samples and additional resources. This section discusses the results of the experiments using BERT pre-trained models on a huge dataset. The experiments were run on raw data before any transformation and on preprocessed data using *UTP*.

### 5.1 Advantages

In general speaking, *UTP* transformed the information from a complex presentation into more easy-to-understand and less confusing for human and machine learning models. Moreover, *UTP* succeeded in reducing the vocabulary size of the corpus and unified the medical terms, which is very helpful for machine learning, as shown in Table 3. For example, replacing all the synonyms of *"high blood pressure"* with *"hypertension"* reduces the complexity of the training of the deep neural network by reducing the vocabulary size. Furthermore, *UTP* is flexible and maintainable because it is easy to add and update the JSON files to include new terms or abbreviations definitions. However, other machine learning techniques require obtaining samples and retraining the model if machine learning approaches were implemented.

### 5.2 Limitations

During the experiments, we mainly focused on the clinical textual data (discharge summaries and chief complaints) to predict the ICD-10 codes. However, in some specialties, physicians rely on the ICD-10 codes more than the documentation. For example, traumatic cases in the emergency room require including information on when, where, and how to describe the cause of trauma, the activity, and the location. Physicians include all this information by selecting the related ICD-10 codes without having a detailed medical documentation, leading to low prediction performance as shown in Table 5. On the other hand, when a department uses a wide range of ICD-10 codes, it decreases the prediction performance, such as in the Internal Medicine department. Also, diseases like diabetes and hypertension can be confusing since they might appear together. The first as the primary diagnosis and the second as the secondary one and vice versa, leading to confusion in training and hence a low performance.

### 5.3 ChatGPT 4 limitations

Although *ChatGPT 4*, which uses a machine-learning approach, gives impressive results. However, an additional and time-consuming work on the dataset was required to be able to send it to *ChatGPT 4*. Moreover, *ChatGPT 4* was trained on public datasets. Therefore, it is exposed to a wider range of biases and inaccuracies than a private pre-trained model has. During the experiment, it was detected that in some samples, nonsensical and nonrelevant sentences were generated by *ChatGPT 4*. Furthermore, the data had to be manually cleaned and the dates obfuscated before using the APIs, which is time-consuming and not applicable to real-world applications in healthcare.

## 6 Conclusion

In this paper, we proposed a new tool called *UTP* that transforms the clinical textual data into more readable and easy-to-understand for human and machine learning tasks. *UTP* uses JSON-based dictionaries and linguistic features to build pattern-matching rules for their flexibility and maintainability. *UTP* unifies medical terms, replaces investigation values with categorical values, converts abbreviations into expanded forms, and transforms dates into periods. Experiments showed the effectiveness of *UTP* on vocabulary reduction without losing information. Also, we demonstrated its positive impact on the performance of machine learning models (micro-F1 reached 74.64% in testing results). In addition, *UTP* was compared to the most recent machine learning tools that can do the same tasks *ChatGPT 4*. As a result of this study, the *UTP* tool is now integrated with the Saudi *"Specialized Medical Center"* hospital systems. Several recommendations were raised to improve the medical documentations, especially the traumatic cases in the emergency service. In future work, we aim to work on improving the accuracy of the ICD-10 prediction by exploring more vital input features. We would also like to investigate the relationships among the ICD-10 labels themselves to improve the ICD-10 predictions. We believe that using a graph-based classifiers which takes into consideration the labels' relationships can improve furthermore the accuracy of the training.

## References

1. Abrahamsson, E., Forni, T., Skeppstedt, M., Kvist, M.: Medical text simplification using synonym replacement: Adapting assessment of word difficulty to a compounding language (2014). https://doi.org/10.3115/v1/w14-1207, http://dx.doi.org/10.3115/v1/W14-1207
2. Alsentzer, E., Murphy, J.R., Boag, W., Weng, W.H., Jin, D., Naumann, T., McDermott, M.B.A.: Publicly available clinical bert embeddings (2019). https://doi.org/10.48550/ARXIV.1904.03323, https://arxiv.org/abs/1904.03323
3. Arora, S., May, A., Zhang, J., Ré, C.: Contextual embeddings: When are they worth it? (2020). https://doi.org/10.18653/v1/2020.acl-main.236, http://dx.doi.org/10.18653/v1/2020.acl-main.236
4. Bird, S., Klein, E., Loper, E.: Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc." (2009)
5. Chen, P.F., Wang, S.M., Liao, W.C., Kuo, L.C., Chen, K.C., Lin, Y.C., Yang, C.Y., Chiu, C.H., Chang, S.C., Lai, F.: Automatic icd-10 coding and training system: Deep neural network based on supervised learning (Aug 2021). https://doi.org/10.2196/23230, http://dx.doi.org/10.2196/23230
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2018). https://doi.org/10.48550/ARXIV.1810.04805, https://arxiv.org/abs/1810.04805
7. ElDin, H.G., AbdulRazek, M., Abdelshafi, M., Sahlol, A.T.: Med-flair: medical named entity recognition for diseases and medications based on flair embedding. Procedia Computer Science **189**, 67–75 (2021). https://doi.org/https://doi.org/10.1016/j.procs.2021.05.078, https://www.sciencedirect.com/science/article/pii/S1877050921011753, aI in Computational Linguistics
8. Hatoum, M., Charr, J.C., Guyeux, C., Laiymani, D., Ghaddar, A.: Emte: An enhanced medical terms extractor using pattern matching rules (2023). https://doi.org/10.5220/0011717300003393, http://dx.doi.org/10.5220/0011717300003393

9. Holper, S., Barmanray, R., Colman, B., Yates, C.J., Liew, D., Smallwood, D.: Ambiguous medical abbreviation study: challenges and opportunities (Sep 2020). https://doi.org/10.1111/imj.14442, http://dx.doi.org/10.1111/imj.14442
10. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020)
11. Jolobe, O.M.P.: Medical abbreviations generate potentially dangerous ambiguity (04 2018). https://doi.org/10.1093/qjmed/hcy074, http://dx.doi.org/10.1093/qjmed/hcy074
12. Leaman, R., Khare, R., Lu, Z.: Challenges in clinical natural language processing for automated disorder normalization (Oct 2015). https://doi.org/10.1016/j.jbi.2015.07.010, http://dx.doi.org/10.1016/j.jbi.2015.07.010
13. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining (2019). https://doi.org/10.48550/ARXIV.1901.08746
14. Li, I., Yasunaga, M., Nuzumlali, M.Y., Caraballo, C., Mahajan, S., Krumholz, H., Radev, D.: A neural topic-attention model for medical term abbreviation disambiguation (2019). https://doi.org/10.48550/ARXIV.1910.14076, https://arxiv.org/abs/1910.14076
15. Liu, X., Hersch, G.L., Khalil, I., Devarakonda, M.: Clinical trial information extraction with bert (2021). https://doi.org/10.48550/ARXIV.2110.10027
16. Maciejewski, M.L., Weaver, E.M., Hebert, P.L.: Synonyms in health services research methodology (Jul 2010). https://doi.org/10.1177/1077558710372809, http://dx.doi.org/10.1177/1077558710372809
17. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Association for Computational Linguistics (ACL) System Demonstrations. pp. 55–60 (2014)
18. Martin, A.K., Green, T.L., McCarthy, A.L., Sowa, P.M., Laakso, E.L.: Healthcare teams: Terminology, confusion, and ramifications (Apr 2022). https://doi.org/10.2147/jmdh.s342197, http://dx.doi.org/10.2147/JMDH.S342197
19. Moharasan, G., Ho, T.B.: Extraction of temporal events from clinical text using semi-supervised conditional random fields (2017). https://doi.org/10.1007/978-3-319-61845-6_41, http://dx.doi.org/10.1007/978-3-319-61845-6_41
20. Moons, E., Khanna, A., Akkasi, A., Moens, M.F.: A comparison of deep learning methods for icd coding of clinical records (Jul 2020). https://doi.org/10.3390/app10155262, http://dx.doi.org/10.3390/app10155262
21. Mustafa, A., Rahimi Azghadi, M.: Automated machine learning for healthcare and clinical notes analysis (2021). https://doi.org/10.3390/computers10020024, https://www.mdpi.com/2073-431X/10/2/24
22. Neumann, M., King, D., Beltagy, I., Ammar, W.: Scispacy: Fast and robust models for biomedical natural language processing (2019). https://doi.org/10.48550/ARXIV.1902.07669, https://arxiv.org/abs/1902.07669
23. Sammani, A., Bagheri, A., van der Heijden, P.G.M., te Riele, A.S.J.M., Baas, A.F., Oosters, C.A.J., Oberski, D., Asselbergs, F.W.: Automatic multilabel detection of icd10 codes in dutch cardiology discharge letters using neural networks (Feb 2021). https://doi.org/10.1038/s41746-021-00404-9, http://dx.doi.org/10.1038/s41746-021-00404-9
24. Singh, S., Mahmood, A.: The nlp cookbook: Modern recipes for transformer based deep learning architectures (2021). https://doi.org/10.1109/access.2021.3077350, http://dx.doi.org/10.1109/ACCESS.2021.3077350
25. Skreta, M., Arbabi, A., Wang, J., Brudno, M.: Training without training data: Improving the generalizability of automated medical abbreviation disambiguation (2019). https://doi.org/10.48550/ARXIV.1912.06174, https://arxiv.org/abs/1912.06174
26. Vermeir, P., Vandijck, D., Degroote, S., Peleman, R., Verhaeghe, R., Mortier, E., Hallaert, G., Van Daele, S., Buylaert, W., Vogelaers, D.: Communication in healthcare: a narrative review of the literature and practical recommendations (Jul 2015). https://doi.org/10.1111/ijcp.12686, http://dx.doi.org/10.1111/ijcp.12686