# KAMEL: Knowledge Aware Medical Entity Linkage to Automate Health Insurance Claims Processing

**Sheng Jie Lui**[1,2], **Cheng Xiang**[1], **Shonali Krishnaswamy**[2]
[1]National University of Singapore
[2]AiDA Technologies
luishengjie@u.nus.edu / shengjie@aidatech.io, elexc@nus.edu.sg, shonali@aidatech.io

## Abstract

Automating the processing of health insurance claims to achieve "Straight-Through Processing" is one of the holy grails that all insurance companies aim to achieve. One of the major impediments to this automation is the difficulty in establishing the relationship between the underwriting exclusions that a policy has and the incoming claim's diagnosis information. Typically, policy underwriting exclusions are captured in free-text such as "Respiratory illnesses are excluded due to a pre-existing asthma condition". A medical claim coming from a hospital would have the diagnosis represented using the International Classification of Disease (ICD) codes from the World Health Organization. The complex and labour-intensive task of establishing the relationship between free-text underwriting exclusions in health insurance policies and medical diagnosis codes from health insurance claims is critical towards determining if a claim should be rejected due to underwriting exclusions. In this work, we present a novel framework that leverages both explicit and implicit domain knowledge present in medical ontologies and pre-trained language models respectively, to effectively establish the relationship between free-text describing medical conditions present in underwriting exclusions and the ICD-10CM diagnosis codes in health insurance claims. Termed KAMEL (Knowledge Aware Medical Entity Linkage), our proposed framework addresses the limitations faced by prior approaches when evaluated on real-world health insurance claims data. Our proposed framework have been deployed in several multi-national health insurance providers to automate their health insurance claims.

## Introduction

Health insurance claims adjudication is a complex and manual task that requires a claims assessor to verify the insured's policy details and historical claims against the current claim to determine if it should be approved or rejected. In the application area of automating health insurance claims processing, Straight-Through Processing (STP) is defined as the process where a health insurance claim is automatically processed without the need of manual intervention. A higher STP rate translates to faster claims reimbursement, which is ideal for health insurance companies as it improves the overall experience. One of the major impediments to this automation is the difficulty in establishing the relationship between the underwriting (UW) exclusions that a policy has and the incoming claim's diagnosis information.

The fundamental concept of health insurance is to pool and share the risks across a risk pool where the individual healthcare needs of the less healthy are subsidized by the relatively lower cost of the healthy (WHO 2000). To control and mitigate a specific risk of an insured risk pool, health insurance underwriting is performed to ascertain if a pre-existing medical condition should be covered and at what cost. Pre-existing medical conditions are often stored as unstructured free-text and are tied to their respective policyholders. Following medical treatments, when a claim is submitted, the claims adjudication process involves verifying the policyholder's UW exclusions against the medical diagnosis described in the claim. This ensures that the claim is not associated with any pre-existing medical conditions. Medical diagnosis are typically stored as ICD-10CM diagnosis codes. The International Classification of Disease (ICD) coding scheme (WHO 2004) is an industry standard used to identify a patient's medical diagnosis. With over 69,000 unique medical codes in the 10th version of the coding scheme, ICD-10CM provides an exhaustive categorization of various medical diagnosis. In addition, multiple codes are used to describe complex medical diagnosis. In the domain of health insurance, up to three unique ICD codes (primary, secondary, and tertiary) are typically used to describe the medical diagnoses associated to a claim. If an ICD code in the health insurance claim is related to the UW exclusion, the claim would be rejected as the corresponding diagnosis has been excluded from the policy. For instance, suppose a policyholder has the following UW exclusion, "Excluded from any disease or disorder of either eye". In this case, all subsequent claims with eye-related medical diagnosis such as cataract and glaucoma will be rejected. Therefore, to effectively determine if a claim should be rejected due to UW exclusion: (i) medical entities such as the diagnosis and other medical terms present in the UW exclusion must be extracted, and (ii) the extracted medical terms must be related to an ICD code present in the claim.

One of the major bottlenecks towards establishing the relationship between UW exclusions and ICD codes lies in the sparse, imbalanced, and complex nature of the ICD coding scheme. As ICD codes represent specific medical diagnosis,

(a)

| Underwriting Exclusion | Primary ICD Code |
|---|---|
| (i) Excluded from any disease or disorder of either **eye**. | H26 (Other Cataract) |
| (ii) Excluded from any disease or disorder of either **eye**. | H25 (Age-related cataract) |
| (iii) Excluded from any disease or disorder of either **eye**. | H40 (Glaucoma) |

(b)

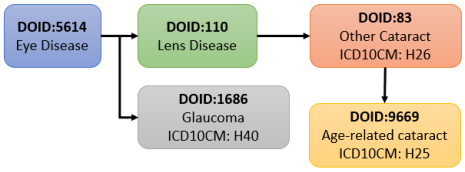


Figure 1: (a) Contains sample underwriting exclusions with their associated primary ICD code present in the health insurance claims. (b) Illustrates a snippet of the Disease Ontology (Schriml et al. 2021) describing the relationship between "eye disease" and the ICD codes present in (a).

a single medical term can correspond to multiple ICD codes. This can be observed in Figure 1, where the medical term 'eye' is associated with multiple ICD codes (H25, H26, and H40). To establish the relationship between UW exclusions and their respective ICD codes, a direct approach is to aggregate ICD codes based on their associated UW exclusions and model it as a multi-label classification problem. However, the sparse and imbalanced distribution of ICD codes may result in poor performance for rare and unseen ICD code combinations.

In this work, we propose a novel framework that can effectively establish the relationship between UW exclusions and ICD codes. We leverage both explicit and implicit domain knowledge present in medical ontologies and pre-trained language models respectively, to extract medical entities from underwriting exclusions and establish the relationship between the extracted medical entities and ICD codes.

In summary, the significance of this work lies in the way we incorporate domain knowledge into the machine learning process to improve its performance. We use health insurance claims adjudication as an application area to demonstrate through extensive experimental validation on real-world health insurance claims data that our proposed approach addresses the challenge of linking UW exclusion text to ICD-10CM codes.

The remainder of this paper is structured into several key sections: Initially, it reviews existing work on clinical text encoding and the establishment of relationships between free-text medical documents and ICD codes. This is followed by a presentation of our proposed methodology. Subsequently, we detail the results from our experiments on real-world health insurance claims data. Finally, the paper concludes with a summary of our findings.

# Related Work

This section provides an overview of previous research in (i) establishing the relationship between medical documents and ICD codes with Automated Clinic Coding as the application use case, (ii) representing medical documents using implicit domain knowledge present in domain-specific pre-trained language models, and (iii) extracting and representing medical entities present in free-text by leveraging on explicit domain knowledge present in medical ontologies.

## Automated Clinical Coding

Automated Clinical Coding involves assigning a set of medical codes to free-text medical documents (Sen et al. 2021). In this subsection, we review prior techniques used in Automated Clinical Coding and how they can be applicable for establishing the relationship between UW exclusions and ICD codes. A key limitation of prior work done in this domain is their narrow scope, which is confined to evaluating the approaches based on only the top-K most frequently occurring ICD codes (Mullenbach et al. 2018; Xu et al. 2019; Wang et al. 2020; Pascual, Luck, and Wattenhofer 2021). This is unacceptable in real-world healthcare environments as rare diseases have sequelae when neglected (Kaur, Ginige, and Obst 2021). To establish the relationship between free-text medical documents and ICD codes, most prior work modelled this problem as a multi-label classification challenge (Mullenbach et al. 2018; Xu et al. 2019; Wang et al. 2020; Pascual, Luck, and Wattenhofer 2021). Given a text input $X$, the objective was to predict a set of associated ICD codes $y \subseteq Y$ where $Y$ denotes the complete ICD code set.

On the other hand, Sen et al. (2021) reformulated the extreme multi-label classification challenge into a simple multi-class classification task. This was achieved by first training a binary text classifier to identify focus sentences (sentences that contain relevant medical entities). Next, each predicted focus sentence is fed into a multi-class text classifier trained to predict ICD codes. However, this approach assumes that each focus sentence corresponds to a unique ICD code, which is not the case for UW exclusions, as medical entities may be associated with multiple ICD codes, as demonstrated in Figure 1.

## Domain Specific Pre-trained Models

In recent years, the adoption of large-scale pre-trained models has gained attention. Notably, BERT (Devlin et al. 2019), showed that the inclusion of bidirectional context enables meaningful context-aware text representations to be generated for a variety of downstream tasks. To tackle domain-specific tasks, prior work fine-tuned BERT (Devlin et al. 2019) on biomedical datasets such as PubMed abstracts and MIMIC-III (Johnson et al. 2016). Trained on large domain-specific datasets, the embeddings generated by domain specific adaptations of BERT (Devlin et al. 2019), like BioBERT (Lee et al. 2019) and BioClinicalBERT (Alsentzer et al. 2019), contain implicit domain knowledge that can be leveraged for downstream tasks, such as assigning ICD codes to medical documents (Huang, Tsai, and Chen 2022).

One of the drawbacks of BERT (Devlin et al. 2019) is that it does not inherently compute independent sentence embeddings. Instead, sentence embeddings are generated by either averaging the output token embeddings (mean pooling) or by using the output embedding of the special classification token, CLS (Reimers and Gurevych 2019). SBERT (Reimers and Gurevych 2019), addressed this limitation by adopting a Siamese network architecture that generates fixed-sized, semantically meaningful sentence embeddings. The inclusion of the Siamese network architecture allows pairwise sentences to be trained using the Classification Objective Function: $o = softmax(Wt(u, v, |u - v|))$. In addition, this representation allows similarity measures such as Cosine Similarity and Euclidean Distance to be used to compute the semantically similarity of sentence pairs (Reimers and Gurevych 2019).

### Extracting Medical Entities using Ontologies

The extraction of medical entities from medical documents, such as UW exclusions, is critical for determining whether a claim should be rejected due to UW exclusion. However, annotating training data for Medical Entity Recognition (MER) is time-consuming and labour-intensive as it requires word-level annotations. In this subsection, we review prior approaches that leverage explicit domain knowledge present in medical ontologies for MER.

Trove (Fries et al. 2021), a weakly supervised entity classification framework, leveraged medical ontologies such as the Unified Medical Language Systems (UMLS) Metathesaurus (Bodenreider 2004), which contains extensive medical knowledge, to annotate medical entities in medical documents. Trove (Fries et al. 2021) demonstrated the effective use of explicit domain knowledge present in off-the-shelf medical ontologies to automate the label annotation process. Furthermore, the work done also highlighted the efficacy of combining domain-specific language models, such as BioBERT (Lee et al. 2019), with weak supervision to build low-cost and efficient classifiers for medical natural language processing.

The Neural Concept Recognizer (NCR) (Arbabi et al. 2019) introduced a novel approach to extract medical entities from documents, aligning them with medical concepts found in a reference ontology. NCR (Arbabi et al. 2019) utilized the hierarchical nature of medical ontologies as an implicit prior for embeddings, enabling the base model to more effectively generalize to unfamiliar synonyms. In addition, to address the challenge of processing documents with long input lengths, NCR (Arbabi et al. 2019) utilized a sliding window technique. This method segments lengthy documents into a list of key phrases, which are then fed into a multi-class classifier.

### Summary

In this section, we reviewed prior techniques used to establish the relationship between free-text medical documents and ICD codes. Prior work done in this domain placed emphasis on multi-label classification techniques and evaluated their approaches on only the top-K most common ICD codes. Most of the work done overlooked cases with rare and uncommon ICD code combinations, which is unacceptable for real-world applications especially in the context of automating health insurance claims processing. In addition, we also reviewed prior work that utilized both implicit and explicit domain knowledge, derived from pre-trained language models and medical ontologies respectively, for the identification of medical entities in free-text medical documents. This aspect is fundamental for our application scenario. The work introduced in the next section aims to overcome the challenges and limitations identified in these previous studies.

## Methodology

In this section, we introduce KAMEL (Knowledge Aware Medical Entity Linkage), a domain knowledge-driven approach that establishes the relationship between free-text UW exclusions and ICD codes in health insurance claims. This relationship is crucial to ensure that claims associated with UW exclusions are not paid out, a critical requirement for achieving STP in health insurance claims automation. KAMEL is composed of two primary components: Implicit Domain Knowledge Inference and Explicit Domain Knowledge Inference. Additionally, an essential aspect of KAMEL is the SBioBERT encoder, which plays a vital role in converting free-text UW exclusions into fixed-length vector representations. These vector representations are subsequently utilized for tasks in both Implicit and Explicit Domain Knowledge Inference. The architecture of KAMEL is depicted in Figure 2, with detailed discussions of each component in the subsequent subsections.

To the best of our knowledge, this is the first work that leverages both implicit and explicit domain knowledge from pre-trained language models and medical ontologies for learning and establishing the relationship between medical terms from UW exclusions and ICD-10CM codes.

### SBioBERT

To represent free-text UW exclusions, we leverage the SBERT (Reimers and Gurevych 2019) architecture, utilising the weights learnt from the pre-trained BioBERT (Lee et al. 2019) model. This encoder, termed SBioBERT, transforms free-text UW exclusions into fixed-length vectors for both Implicit and Explicit Domain Knowledge Inferences. Furthermore, SBioBERT can be fine-tuned on UW exclusions and their associated ICD code descriptions using the MultipleNegativesRankingLoss (Henderson et al. 2017).

### SSI: Semantic Similarity Inference

Embeddings generated by SBioBERT contain implicit domain knowledge learned from domain-specific datasets on which it has been trained. This allows meaningful sentence embeddings to be generated. The Semantic Similarity Inference (SSI) module determines the relationship between UW exclusions and ICD codes by computing the cosine similarity between the embeddings of UW exclusions and the ICD codes descriptions. A high cosine similarity score indicates semantic similarity, implying that the claim should be rejected due to UW exclusion. This process is formalized

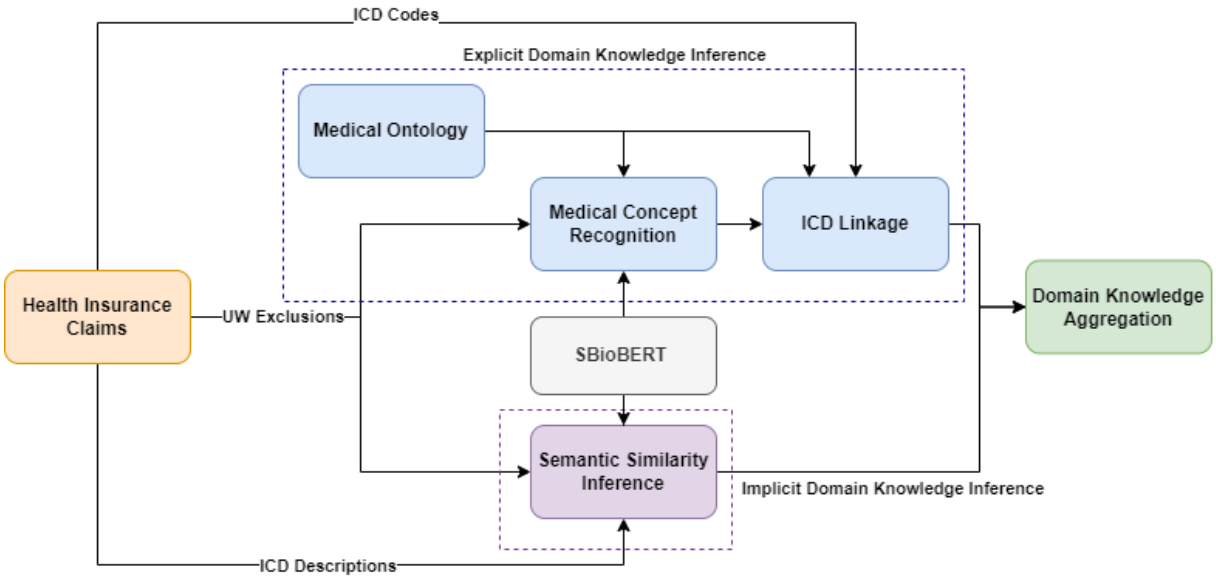


Figure 2: The overall architecture of the KAMEL framework consists of two key components: (i) Explicit Domain Knowledge Inference and (ii) Implicit Domain Knowledge Inference. Component (i) leverages explicit domain knowledge present in medical ontologies to extract and establish the relationship between medical entities and ICD codes. Component (ii) leverages implicit domain knowledge from learned sentence embeddings to compute the semantic similarity between UW exclusion text and ICD code descriptions.

in Equation 1, where $A$, $B$ represents the SBioBERT embeddings obtained from the UW exclusion and ICD description respectively, and $t$ represents a pre-defined threshold. A claim will be flagged as rejected due to UW exclusion if the cosine similarity score is greater than threshold $t$.

$$F(A, B, t) = \frac{AB}{\|A\|\|B\|} > t \tag{1}$$

## Medical Ontology

In this work we utilize medical ontologies to incorporate explicit domain knowledge into KAMEL. Specifically, we focus on two medical ontologies: (i) Disease Ontology (DO) (Schriml et al. 2021) and (ii) ICD Ontology. Medical ontologies, which are structured as directed acyclic graphs (DAGs), describe the relationships among various diagnoses. In the DO (Schriml et al. 2021), specific medical concepts include their associated ICD codes as node attributes. This allows relationships between medical concepts and corresponding ICD codes to be established. For example, as shown in Figure 1, the node DOID:83 (Other Cataract) contains the ICD-10CM node attribute H26. However, a significant limitation of DO (Schriml et al. 2021) is that it does not contain an exhaustive list of all ICD-10CM codes. To mitigate this, we complement DO (Schriml et al. 2021) by representing the ICD-10CM codes as an ontology termed ICD Ontology.

The ICD Ontology is generated based on the hierarchical properties of the ICD-10CM coding scheme (WHO 2004). We propose the use of a three-level ontology which is comprised of (i) ICD Chapter (e.g. H00-H59), (ii) ICD Section (e.g, H25-H28), and (iii) ICD Category (e.g. H26) nodes. Sub-category details (e.g. H26.0), that offer greater specificity, are stored as node attributes within their respective ICD Category nodes. This approach minimizes memory usage by requiring fewer nodes to represent the ontology, yet effectively preserves the hierarchical relationship among ICD codes.

## MCR: Medical Concept Recognition

This subsection details the Medical Concept Recognition (MCR) module, tasked with identifying medical concepts within free-text UW exclusions, based on a reference medical ontology. This is achieved by first embedding all medical concepts from the medical ontology into a reference embedding matrix $M_{ref}$ using the SBioBERT encoder. Next, the input UW exclusion is split into candidate key phrases using a varying sliding window approach described in Equation 2 where $w$ denotes the maximum window length and $S_i^l$ represents the sliding window function that generates a set of key phrases from sentence $S$ with a window length $l$. The varying sliding function, $VSW(S, w)$, generates key phrases from window length 1 to $w$. For instance, given the UW exclusion "Exclude eye related disorders", when $w = 2$, the following candidate key phrases will be generated: 'Exclude', 'eye', 'related', 'disorders', 'Exclude eye', 'eye related', 'related disorders'.

$$VSW(S, w) = \bigcup_{l=1}^{w} \{S_i^l\} \tag{2}$$

The candidate key phrases generated are processed with SBioBERT to obtain an embedding matrix $M_c$. To identify

Algorithm 1: MCR

**Input**: Free-text medical document ($M$); Reference Ontology ($O$)
**Parameter**: Max window length ($w$); Cosine-similarity threshold ($t$)
**Output**: a set of relevant medical concepts (nodes in $O$)

$M_{ref} = SBioBERT(nodes(O))$.
Using Equation 2, let $C = VSW(M, w)$.
$Z_n = \{\}$ # Initialize empty set
**for** $c \in C$ **do**
  $M_c = SBioBERT(c)$.
  **if** $F(M_{ref}, M_c, t)$ is true **then**
    $Z_n = Z_n \cup c$
  **end if**
**end for**
**return** $Z_n$

relevant medical terms, we compute the cosine similarity between $M_c$ and $M_{ref}$ as defined in Equation 1. The threshold $t$ is applied to ensure that only key phrases associated with medical concepts are extracted. Next, these extracted medical concepts are input into the ICD Linkage module. The complete MCR algorithm is detailed in Algorithm 1: MCR.

## ICD Linkage

This subsection focuses on establishing the relationship between the medical concepts identified in the MCR module and the ICD codes from the health insurance claim. The relationship between medical terms and ICD codes are typically classified into two categories: (i) direct relationship and (ii) complex relationship.

To determine if there exists a direct relationship between medical terms present in UW exclusions and ICD codes, the extracted terms and codes are mapped to their corresponding ontology nodes $N_{MC}$ and $N_{ICD}$. Next, a Depth-First Search (DFS) is performed to determine if a path exists between nodes in $N_{MC}$ and $N_{ICD}$. If a path exists between nodes in $N_{MC}$ and $N_{ICD}$, indicating a direct relationship between the medical term and ICD code, the claim should be flagged as rejected due to UW exclusion.

To establish more complex relationships between nodes with related medical concepts, we employ the concept of Valid Ancestors (VAs). VAs are defined as a set of Lowest Common Ancestors (LCAs) that can effectively establish the relationship between nodes in $N_{MC}$ and $N_{ICD}$. For instance, as illustrated in Figure 1, the term 'eye' serves as a VA to establish the relationship between diagnoses such as cataract and glaucoma. Obtaining the VA involves two key steps: (i) maximize the overall score of a set node pairs $NP_{pos} = (n_{MC}, n_{ICD})$ associated with claims that are rejected due to UW exclusion where $n_{MC} \in N_{MC}$ and $n_{ICD} \in N_{ICD}$ and (ii) minimize the overall score for node pairs $NP_{neg} = (n_{MC}, n_{ICD})$ associated with claims that are approved. The scoring function $F_{score}(n, NP)$ computes the number of node pairs, $NP$, present in the descendants of node $n$.

Algorithm 2: Enhanced ICD Linkage using VA

**Input**: Set of positive node pairs (pos); Set of negative node pairs (neg); Reference Ontology ($O$)
**Output**: list of valid ancestors

Let $N_{base} = flatten(NP_{pos})$ be a set of base nodes.
Let $n \in nodes(O)$.
Using Equation 3, let $R(n) = F_{score}(n, NP_{pos}) - F_{score}(n, NP_{neg})$ be the reward function.
**Bottom-up**: For each base node, recursively traverse to its parent node if $R(n_{parent}) > R(n_{current})$.
**Bottom-up**: Let $VA_{BU}$ be the terminal nodes obtained from step 4.
**Top-down**: For each node, $va_{BU} \in VA_{BU}$ perform a BFS.
**Top-down**: Store the children nodes in $VA_{TD}$ if $R(children(va_{BU})) > R(va_{BU})$.
**Aggregate**: Let $G_{VA}$ be a sub-graph of the ontology where its nodes $n_x \in (VA_{BU} \cup VA_{TD})$.
**Aggregate**: The set of valid ancestors $N_{VA}$ is the leaf nodes of $G_{VA}$
**return** $N_{VA}$

A naive approach to obtain a set of VAs is to compute the power set of all nodes in the ontology and identify the set that maximizes the reward function in Equation 3. However, since the cardinality of power set is $2^n$ where $n$ is the number of nodes in the ontology, computing the power set for all nodes in the ontology is impractical due to the large number of nodes involved.

$$R(n) = F_{score}(n, NP_{pos}) - F_{score}(n, NP_{neg}) \quad (3)$$

To address this challenge, we propose an enhanced ICD linkage algorithm, as highlighted in Algorithm 2. Initially, we extract all unique node pairs that are associated with claims rejected due to UW exclusion, $NP_{pos}$, and flatten them to obtain a set of base nodes, $N_{base}$. Next, we apply the Bottom-Up approach, which recursively traverse each base node $n_{base} \in N_{base}$ and stores the results in $VA_{BU}$. Subsequently, we implement the Top-Down approach, which recursively traverses each node in $VA_{BU}$ and stores the results $VA_{TD}$. Finally, we aggregate the results by generating a sub-graph $G_{VA}$, with its nodes represented as $n_{va} \in VA_{BU} \cup VA_{TD}$. The set of valid ancestors $N_{VA}$ is determined by identifying the leaf nodes of $G_{VA}$. During inference, if both the extracted medical concepts and ICD Codes share the same VA, the claim should be flagged as rejected due to UW exclusion.

## Domain Knowledge Aggregation

To obtain the final decision, we use the MAX operator to aggregate results from both the Implicit and Explicit Domain Knowledge Inference components, capturing global (sentence-level) and local (medical concept-level) representations respectively.

| | Health Insurance I | Health Insurance II | MIMIC-IV-ICD-10-N3 |
|---|---|---|---|
| # Total Claims/PR | 76197 | 255706 | 32318 |
| # Total Claims/PR (Train) | 36077 | 197101 | 22622 |
| # Approved Claims/PR (Train) | 28064 | 166863 | 5663 |
| # Rejected Claims/PR (Train) | 8013 | 30238 | 16959 |
| # Total Claims/PR (Test) | 40039 | 58312 | 9696 |
| # Approved Claims/PR (Test) | 30401 | 51180 | 2416 |
| # Rejected Claims/PR (Test) | 9638 | 7132 | 7280 |

Table 1: Summary of evaluation datasets with a detailed breakdown of claims / patient records (PR) in both training and test datasets for Health Insurance I, Health Insurance II, and MIMIC-IV-ICD-10-N3.

# Experiments

## Experimental Objectives

In this work, we evaluate the performance of KAMEL in establishing the relationship between UW exclusions and ICD codes, which is critical towards determining if a claim should be rejected due to UW exclusion. Two experiments were conducted to (i) evaluate the different components of KAMEL and (ii) evaluate the performance of KAMEL against conventional multi-label classification techniques. KAMEL leverages both implicit and explicit domain knowledge to infer the relationship between UW exclusions and ICD codes. To analyse the performance of both domain knowledge inference techniques, we first perform an ablation study on the different components of KAMEL. Next, to validate our domain knowledge-driven approach, we evaluate KAMEL against conventional data-driven multi-label classification techniques. In this experiment, KAMEL was evaluated against two benchmark multi-label classification techniques: (i) CAML (Mullenbach et al. 2018), and (ii) PLM-ICD (Huang, Tsai, and Chen 2022). The benchmarks were selected as both approaches were designed to assign a set of ICD codes to free-text medical documents with reproducible source codes.

## Experiment Set-up

In the experiments conducted, the F1 and Accuracy score were used to evaluate the proposed approach. As the objective is to increase STP, an approach that has the right trade-off between precision and recall is critical. For instance, an approach with high precision but low recall will result in a high number of claims that should be rejected due to UW exclusions not being identified. On the other hand, an approach with high recall but low precision will result in numerous claims to be incorrectly flagged as rejected due to UW exclusions. The former results in significant claims leakage, with incorrect payouts that will adversely affect the company's profitability. The latter will decrease the STP rate, increasing the time taken for claims to be processed and negatively impacting the customer experience. The F1-score, which harmonizes precision and recall, is an effective metric for gauging this balance.

In this work, we present three variations of KAMEL: (i) KAMEL-ICD: trained on the ICD Ontology, (ii) KAMEL-DO: trained on the DO (Schriml et al. 2021), and (iii) KAMEL-DOICD: trained on both ICD Ontology and DO (Schriml et al. 2021). The different variations of KAMEL were evaluated to provide a better understanding of learning from different ontologies. To simplify the training process, we fine-tuned both the SBioBERT model and the VA component (detailed in the ICD Linkage subsection) of KAMEL on claims with a single distinct ICD code. To determine the optimal SSI cosine similarity threshold $t$ (Equation 1), we validated KAMEL on a range of cosine similarity thresholds, specifically from 0.8 to 0.9. Among the thresholds tested, a cosine similarity threshold of 0.9 yielded the highest F1 score. Therefore, we selected the threshold $t = 0.9$ to ensure that inference is made on semantically similar descriptions.

When evaluating KAMEL against its multi-label classification benchmarks, CAML (Mullenbach et al. 2018), and PLM-ICD (Huang, Tsai, and Chen 2022), we preprocess the data by aggregating ICD codes based on their associated UW Exclusions, enabling its use in a multi-label classification scenario. Similar to KAMEL, we used BioBERT (Lee et al. 2019) as the base model for PLM-ICD (Huang, Tsai, and Chen 2022). Both benchmark models were trained for up to 10 epochs. During evaluation, we set the classification threshold of the benchmark approaches, $t_{bench} = 0.5$. It is worth mentioning that multiple experiments (with varying threshold values) were conducted and $t_{bench} = 0.5$ achieved the highest F1 score. A claim is rejected due to UW exclusion when the UW exclusion is associated to the ICD-10CM code. We determine if a claim should be rejected using the following equation: $f_{eval}(U, P) = ||U \cap P|| \geq 1$ where $P$ and $U$ denote the set of predicted and labelled ICD codes respectively.

## Datasets

We evaluate our approach on three real-world datasets: (i) Health Insurance I, (ii) Health Insurance II and (iii) MIMIC-IV-ICD-10-N3. Dataset (i) and (ii) are real-world private health insurance datasets and dataset (iii) is a publicly available dataset used to simulate the detection of UW exclusion. The datasets used comprises of seven key fields: (i) Free-text UW Exclusion / discharge summary, (ii) Primary ICD Code, (iii) Secondary ICD Code, (iv) Tertiary ICD Code, (v) Primary ICD Description, (vi) Secondary ICD Description, and (vii) Tertiary ICD Description. Table I provides a detailed breakdown of the datasets used.

The MIMICIV-ICD-10-N3 dataset originates from the MIMICIV dataset (Johnson et al. 2023). In MIMIC-IV-ICD-

10-N3 the discharge summary is used to simulate UW exclusions. Given that health insurance claims typically comprises of up to three ICD codes, only records with a maximum of three distinct ICD codes were included. To replicate approved claims (those that were not rejected due to UW exclusion), we introduce patient records (PRs) containing unrelated ICD codes into the dataset.

## Experiment Results

Two experiments were conducted to (i) validate the effectiveness of the different components of KAMEL and (ii) assess KAMEL against the benchmark multi-label classification approaches. Table 2 and 3 present the results obtained for Experiments (i) and (ii), respectively.

To verify the effectiveness of the different components of KAMEL, we perform an ablation study on all three datasets and presented the results in Table 2. It is evident from the improved performance of KAMEL when compared to its SSI component that Implicit Domain Knowledge Inference can be further enhanced through the inclusion of Explicit Domain Knowledge Inference. In addition, it can be observed that SSI underperformed when evaluated on MIMIC-IV-ICD-10-N3. This can be attributed to the complex structure of the discharge summaries, which encompasses numerous medical terms, including medical history, discharge medications, and allergies. As SSI computes the cosine similarity of the entire discharge summary against the corresponding ICD code descriptions, the presence of irrelevant medical terms in the discharge summary can introduce noise, leading to poor performance. Additionally, it is apparent that KAMEL-DO generally exhibits a lower recall score than KAMEL-ICD. This difference may be due to the limitations of DO (Schriml et al. 2021), which does not encompass a comprehensive list of all ICD-10CM codes. Nevertheless, the challenge arising from learning with incomplete ontologies can be addressed by incorporating multiple ontologies, as demonstrated by the improved recall and F1-score of KAMEL-DOICD compared to both KAMEL-DO and KAMEL-ICD.

To validate KAMEL's performance against multi-label classification approaches, we evaluated the performance of KAMEL-DOICD against CAML (Mullenbach et al. 2018) and PLM-ICD (Huang, Tsai, and Chen 2022), and presented the findings in Table 3. Based on the results obtained, KAMEL-DOICD consistently outperformed the benchmark multi-label classification techniques, underscoring its efficacy. Although PLM-ICD (Huang, Tsai, and Chen 2022) achieved a higher recall for both health insurance datasets, its low precision score resulted in a low F1 score. This suggests that PLM-ICD (Huang, Tsai, and Chen 2022) struggles to accurately establish the relationship between free-text UW exclusions and ICD Codes. This underperformance can be attributed to the sparse representation of ICD codes, leading to a large number of labels. Additionally, the naive approach of aggregating ICD codes based on their UW exclusions overlooks textual variations that represent similar medical concepts.

Furthermore, as demonstrated in Table 3, due to the complex structure of the discharge summary and sparse ICD code representation in MIMIC-IV-ICD-10-N3, techniques that represent the entire discharge summary as a single vector failed to accurately capture the relationship between free-text descriptions and ICD codes. Consequently, this resulted in a low accuracy and F1-score for CAML (Mullenbach et al. 2018) and PLM-ICD (Huang, Tsai, and Chen 2022). Conversely, KAMEL's approach of first extracting medical entities from the free-text medical document yielded significantly higher accuracy and F1-scores. This highlights the versatility of KAMEL when applied to different types of medical documents.

In summary, the results underscore the effectiveness of the KAMEL framework as it consistently outperformed all benchmark approaches in both accuracy and F1-score. Additionally, KAMEL's explicit domain knowledge inference capabilities enable the establishment of clear and explainable relationships between UW exclusions and ICD codes. This enhanced explainability and traceability are particularly valuable in real-world production settings. For instance, consider a policy with the exclusion clause "Exclusion of the heart or the cardiovascular system related disease" and the ICD code I47.1 (Supraventricular tachycardia). KAMEL can effectively extract the medical entity 'cardiovascular' and clearly establish the relationship between 'cardiovascular' and I47.1 by utilising the structural properties of the medical ontology: cardiovascular $\rightarrow$ heart $\rightarrow$ Other forms of heart disease (I30-I5A) $\rightarrow$ Supraventricular tachycardia (I47.1).

# Conclusion and Future Work

## Conclusion

Based on the results obtained, the KAMEL framework highlights the efficacy of incorporating domain knowledge into the learning process for real-world machine learning applications. We applied this framework to the domain of real-world health insurance claims adjudication and presented our methodology for establishing the relationship between free-text UW exclusions and ICD codes. Through extensive experimental validation using real-world health insurance claims data, we demonstrated that our approach effectively addresses the challenge of establishing this relationship. The KAMEL framework, introduced in this work, focus primarily on determining whether a claim should be rejected due to UW exclusions. It plays a critical role within the Smart Claims system (AIDA 2023), an automated solution for health insurance claims adjudication. This work represents a significant contribution to the development of practical and deployable solutions for automated health insurance claims adjudication in real-world scenarios.

## Limitations and Future Work

The proposed approach focuses on two key aspects: (i) extracting medical entities present in the UW exclusions and (ii) establishing the relationship between the extracted medical entities and ICD codes present in the claim. Notably, it does not currently consider inclusion and exclusion terms. For instance, given the UW exclusion "Exclude all eye related conditions except Cataract", KAMEL does not cor-

| | DOMAIN KNOWLEDGE INFERENCE | METRIC | Health Insurance I | Health Insurance II | MIMIC-IV-ICD-10-N3 |
|---|---|---|---|---|---|
| SSI | Implicit | accuracy | 0.8508 | 0.8244 | 0.4974 |
| | | f1 | 0.1641 | 0.5339 | 0.0000 |
| | | precision | 0.9816 | 0.7992 | 0.0000 |
| | | recall | 0.0895 | 0.4008 | 0.0000 |
| KAMEL-ICD | Implicit + Explicit | accuracy | 0.8641 | 0.8326 | 0.7685 |
| | | f1 | 0.2940 | 0.5685 | 0.7384 |
| | | precision | 0.9767 | 0.8042 | **0.8550** |
| | | recall | 0.1731 | 0.4397 | 0.6497 |
| KAMEL-DO | Implicit + Explicit | accuracy | 0.8636 | 0.7968 | 0.7143 |
| | | f1 | 0.3200 | 0.3240 | 0.6690 |
| | | precision | 0.9236 | **0.9778** | 0.8012 |
| | | recall | 0.1808 | 0.1941 | 0.5742 |
| KAMEL-DOICD | Implicit + Explicit | accuracy | **0.9743** | **0.8534** | **0.7750** |
| | | f1 | **0.3924** | **0.6423** | **0.7775** |
| | | precision | **0.9361** | 0.8278 | 0.7728 |
| | | recall | **0.2482** | **0.5247** | **0.7823** |

Table 2: Ablation experimental results of various components of KAMEL validated on Health Insurance I, Health Insurance II, and MIMIC-IV-ICD-10-N3. The best scores are highlighted in bold.

| DATA | METRIC | CAML | PLM-ICD | KAMEL-DOID |
|---|---|---|---|---|
| Health Insurance I | accuracy | 0.5310 | 0.4105 | **0.9743** |
| | f1 | 0.2352 | 0.2593 | **0.3924** |
| | precision | 0.1604 | 0.1632 | **0.9361** |
| | recall | 0.4410 | **0.6310** | 0.2482 |
| Health Insurance II | accuracy | 0.5415 | 0.4191 | **0.8534** |
| | f1 | 0.3610 | 0.3346 | **0.6423** |
| | precision | 0.2775 | 0.2348 | **0.8278** |
| | recall | 0.5163 | **0.5824** | 0.5247 |
| MIMIC-IV-ICD-10-N3 | accuracy | 0.2822 | 0.2881 | **0.7750** |
| | f1 | 0.0000 | 0.1225 | **0.7775** |
| | precision | 0.0000 | **0.8211** | 0.7728 |
| | recall | 0.0000 | 0.0662 | **0.7823** |

Table 3: Experiment results evaluating KAMEL against multi-label classification techniques validated on Health Insurance I, Health Insurance II, and MIMIC-IV-ICD-10-N3. The best scores are highlighted in bold.

rectly interpret that cataract-related diagnosis should not be excluded. Our future work will involve enhancing KAMEL's medical concept recognition capabilities to incorporate inclusion and exclusion terms. In addition, since KAMEL relies on medical ontologies to establish the relationship between medical concepts and ICD codes, our future work also involves enhancing these ontologies to establish more complex disease-to-disease and disease-to-ICD relationships.

## References

AIDA. 2023. AiDA Technologies: Smart Claims. https://www.aidatech.io/smart_claims. [Online; accessed 2023-11-06].

Alsentzer, E.; Murphy, J.; Boag, W.; Weng, W.-H.; Jindi, D.; Naumann, T.; and McDermott, M. 2019. Publicly Available Clinical BERT Embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, 72–78. Minneapolis, Minnesota, USA: Association for Computational Linguistics.

Arbabi, A.; Adams, D. R.; Fidler, S.; and Brudno, M. 2019. Identifying Clinical Terms in Medical Text Using Ontology-Guided Machine Learning. *JMIR Med Inform*, 7(2): e12596.

Bodenreider, O. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32(Database-Issue): 267–270.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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. Minneapolis, Minnesota: Association for Computational Linguistics.

Fries, J. A.; Steinberg, E.; Khattar, S.; Fleming, S. L.; Posada, J.; Callahan, A.; and Shah, N. H. 2021. Ontology-

driven weak supervision for clinical entity classification in electronic health records. *Nature communications*, 12(1): 2017.

Henderson, M.; Al-Rfou, R.; Strope, B.; Sung, Y.-H.; Lukács, L.; Guo, R.; Kumar, S.; Miklos, B.; and Kurzweil, R. 2017. Efficient Natural Language Response Suggestion for Smart Reply. *ArXiv*, abs/1705.00652.

Huang, C.-W.; Tsai, S.-C.; and Chen, Y.-N. 2022. PLM-ICD: Automatic ICD Coding with Pretrained Language Models. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, 5, 10–20. Seattle, WA: Association for Computational Linguistics.

Johnson, A.; Bulgarelli, L.; Shen, L.; Gayles, A.; Shammout, A.; Horng, S.; Pollard, T.; Hao, S.; Moody, B.; Gow, B.; Lehman, L.-w.; Celi, L.; and Mark, R. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10: 1.

Johnson, A.; Pollard, T.; Shen, L.; Lehman, L.-w.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Celi, L.; and Mark, R. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3: 160035.

Kaur, R.; Ginige, J. A.; and Obst, O. 2021. A Systematic Literature Review of Automated ICD Coding and Classification Systems using Discharge Summaries. *ArXiv*, abs/2107.10652.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; and Eisenstein, J. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 4, 1101–1111. New Orleans, Louisiana: Association for Computational Linguistics.

Pascual, D.; Luck, S.; and Wattenhofer, R. 2021. Towards BERT-based Automatic ICD Coding: Limitations and Opportunities. In *Workshop on Biomedical Natural Language Processing*.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Conference on Empirical Methods in Natural Language Processing*.

Schriml, L.; Munro, J.; Schor, M.; Olley, D.; McCracken, C.; Felix, V.; Baron, J. A.; Jackson, R.; Bello, S.; Bearer, C.; Lichenstein, R.; Bisordi, K.; Dialo, N.; Giglio, M.; and Greene, C. 2021. The Human Disease Ontology 2022 update. *Nucleic Acids Research*, 50(3).

Sen, C.; Ye, B.; Aslam, J.; and Tahmasebi, A. 2021. From Extreme Multi-label to Multi-class: A Hierarchical Approach for Automated ICD-10 Coding Using Phrase-level Attention.

Wang, S.-M.; hsuan Chang, Y.; Kuo, L.-C.; Lai, F.; Chen, Y.-N. V.; yun Yu, F.; Chen, C.-W.; wei Li, Z.; and Chung, Y.-F. 2020. Using Deep Learning for Automatic Icd-10 Classification from Free-Text Data.

WHO, W. H. O. 2000. *The World health report : 2000 : health systems : improving performance*. 1. World Health Organization.

WHO, W. H. O. 2004. ICD-10 : international statistical classification of diseases and related health problems : tenth revision.

Xu, K.; Lam, M.; Pang, J.; Gao, X.; Band, C.; Mathur, P.; Papay, F.; Khanna, A. K.; Cywinski, J. B.; Maheshwari, K.; Xie, P.; and Xing, E. P. 2019. Multimodal Machine Learning for Automated ICD Coding. In Doshi-Velez, F.; Fackler, J.; Jung, K.; Kale, D.; Ranganath, R.; Wallace, B.; and Wiens, J., eds., *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, 197–215. PMLR.