Neural machine translation of clinical procedure codes for medical diagnosis and uncertainty quantification

Pei-Hung Chung
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX 77843
chung95191@tamu.edu

Shuhan He*
Emergency Medicine
Massachusetts General Hospital
Boston, MA 02115
she@mgh.harvard.edu

Norawit Kijpaisalratana
Emergency Medicine
Massachusetts General Hospital
Boston, MA 02115
nkijpaisalratana1@mgh.harvard.edu

Abdel-badih el Ariss
Emergency Medicine
Massachusetts General Hospital
Boston, MA 02115
aariss@mgh.harvard.edu

Byung-Jun Yoon*
Computational Science Initiative
Brookhaven National Laboratory
Upton, NY 11973
Department of Electrical and Computer Engineering
Texas A&M University
College Station, TX 77843
bjyoon@tamu.edu

Abstract

A Clinical Decision Support System (CDSS) is designed to enhance clinician decision-making by combining system-generated recommendations with medical expertise. Given the high costs, intensive labor, and time-sensitive nature of medical treatments, there is a pressing need for efficient decision support, especially in complex emergency scenarios. In these scenarios, where information can be limited, an advanced CDSS framework that leverages AI (artificial intelligence) models to effectively reduce diagnostic uncertainty has utility. Such an AI-enabled CDSS framework with quantified uncertainty promises to be practical and beneficial in the demanding context of real-world medical care. In this study, we introduce the concept of Medical Entropy, quantifying uncertainties in patient outcomes predicted by neural machine translation based on the ICD-9 code of procedures. Our experimental results not only show strong correlations between procedure and diagnosis sequences based on the simple ICD-9 code but also demonstrate the promising capacity to model trends of uncertainties during hospitalizations through a data-driven approach.

1 Introduction

When a patient presents for clinical care hospital, clinicians sometimes face the challenge of initial uncertainty[1], necessitating more data from examinations and observations. This uncertainty is often highest at the point of first presentation and can be addressed through a series of appropriate

*Corresponding authors

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

procedures. As treatment progresses, effective therapies can progressively reduce the unknown aspects of the patient's condition. The goal is to minimize this uncertainty with effective treatments, as each illness presents several viable treatment options. However, the evolving nature of the patient's condition demands the identification of the most suitable treatment plan. In this context, a Clinical Decision Support System (CDSS) becomes crucial, assisting clinicians in decision-making by efficiently narrowing down uncertainties with limited information[2].

In this evolving landscape of clinical decision-making, the use of CDSS becomes pivotal. CDSSs help clinicians navigate through the complexities of medical care, much like Automatic Speech Recognition (ASR)[3] systems use phonemes as the minimal units to represent their semantic contents from the speech signals in the acoustic models, then predict the combinations of plausible words in sentences in the language models. In this study, we introduce a novel framework centered around reducing uncertainty in clinical decision-making. The initial validation of this framework is conducted through a retrospective review utilizing the International Classification of Diseases-Ninth Revision (ICD-9) and Current Procedural Terminology (CPT) procedure codes. We emphasize that the use of ICD-9 and CPT codes is instrumental for validation purposes, serving to corroborate our primary approach towards entropy and uncertainty quantification, and ultimately, the reduction thereof. While these codes are typically generated post patient stay, thus not providing prospective data, their utility in affirming the validity of our uncertainty quantification/reduction framework is invaluable. Our approach in this study harnesses the systematic structure of ICD-9 and CPT codes to encode the potential patients' conditions and further predict subsequent steps in a sequence, employing them like the elements of phoneme and subword in ASR system to guide decision-making. This methodology enables CDSSs to efficiently narrow down uncertainties with limited initial information, thereby assisting clinicians in making informed decisions[2]. Furthermore, the design of a CDSS integrates these encoded guidelines with the clinician's medical expertise, forming a synergy that is crucial in time-critical and resource-intensive medical scenarios. While standard operating procedures exist for inpatients, the dynamic and often critical nature of hospital environments demands quick yet accurate decision-making. Here, the CDSS, empowered by its encoded data akin to a linguistic system, plays a vital role in guiding clinicians through complex medical situations.

CDSS can be classified as knowledge-based and non-knowledge based. For knowledge-based systems, decisions are made based on predefined rules and medical guidelines[4]. In contrast, non-knowledge-based systems facilitate physicians making precise arrangements by data-driven approaches based on artificial intelligence (AI) / machine learning (ML) models [5]. Knowledge-based systems achieve great success in diagnostic support when giving medical expertise on symptoms and side effects. As for AI/ML models, despite their remarkable capacity to manipulate massive amounts of data in non-knowledge-based systems, they are a black box[6]. Data availability might also restrict feature extraction from provided data to represent the patient's health condition[7]. In other words, features obtained from vital signs, electrocardiograms, or laboratory results during a hospital stay might not be viable when a patient is newly admitted to the hospital or when medical facilities are not applicable for any reason.

Information entropy, also known as Shannon entropy[8], has demonstrated its utility in digital communication and data compression[9]. This idea of information entropy enables us to quantify the amount of uncertainty (or variability) of quantities of interest (QoI) based on their probability distributions and it has been also playing a central role in AI/ML.

Medical entropy has been previously applied to clinical calculators and has been proposed as a substitute for sensitivity and specificity[10]. In this study, we propose a framework to measure medical uncertainty at every stage during specific admissions by estimating the uncertainties caused by heterogeneous factors such as medical history, multifarious etiology, or uncaptured data in medical scenarios. This aims to optimize clinical decision-making by providing a more nuanced and comprehensive understanding of patient-specific variables, ultimately leading to tailored, efficient, and effective patient care through AI-enabled CDSS.

2 Materials and Methods

2.1 Data Source and Experiment Compute Resource

In this study, we focused on the Medical Information Mart for Intensive Care (MIMIC)-IV database[11], particularly on the International Classification of Diseases (ICD) codes used for diagnosing and

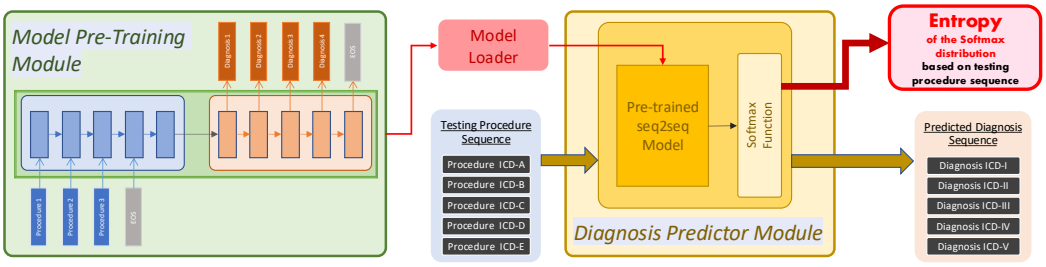

Figure 1: The proposed framework of entropy quantification includes Model Pre-Training Module (MPM) and Diagnosis Predictor Module (DPM) by using the procedure and diagnosis ICD-9 codes in the MIMIC-IV database.

documenting procedures during hospital admissions. Our experiments were conducted using cases recorded with ICD-9 codes[12], encompassing a total of 155,933 admissions.

ICD codes provide a standardized framework for diagnosis, while CPT codes document the procedures carried out. Although these codes may not directly mirror actual medical practices, they offer a clear and structured way to understand the activities and decisions during a patient's hospital stay[13]. This structured approach is not only beneficial for clarity in clinical understanding but also proves invaluable for AI/ML applications. The use of ICD codes for encoding diagnostic terms in clinical documentation has been a focal point of research, underscoring their crucial role in efficiently extracting and analyzing essential data from electronic health records[14]. This synergy between standardized medical coding and AI/ML is a key aspect of our study.

All experiments in this study were conducted on a CPU, using an Apple MacBook Pro equipped with the M1 Max chip. This setup not only provided sufficient computational power for processing the ICD-9 coded data and running the machine learning models, but also enabled us to carry out the experiments with relative ease, without requiring additional high-performance computing resources such as GPUs or cloud-based servers. This also ensured that the study remains reproducible on accessible hardware for users with similar computational setups.

2.2 Model Frameworks

Figure 1 illustrates the proposed framework of entropy quantification during hospitalization, which consists of two main modules: the Model Pre-training Module (MPM) and the Diagnosis Predictor Module (DPM). To specify, the upstream MPM is for pre-training predictive models, and the downstream DPM takes the pre-trained model to obtain the output distribution while predicting diagnosis. Given the testing procedure sequence, which can be arbitrary, the decoder in DPM can further predict diagnosis. Meanwhile, we can obtain the entropy of the output distribution, which is also the confidence of potential diagnosis.

In our framework's application to a clinical setting, the process begins when a physician inputs patient data accumulated since admission into the DPM. This model evaluates the data to estimate the diagnostic likelihood, represented as a probability distribution across potential diagnoses. It's important to note that these diagnoses represent a combination of different health conditions at a specific stage, rather than a single disease state.

Subsequently, the DPM calculates the information entropy derived from this diagnostic probability distribution, presenting it as "medical entropy" to the clinician. This entropy serves as a quantitative measure of uncertainty in the diagnosis. The physician, upon reviewing the medical entropy, can then propose potential interventions for the subsequent stage of patient care, aiming at reducing this uncertainty. This action allows the clinician to assess how these interventions might increase or decrease the medical entropy, thereby refining or expanding the differential diagnosis.

Throughout this interactive process, the physician gains insights not only into the immediate medical entropy linked to the current clinical query but also into a sequence of potential therapeutic options, each with their respective posterior medical entropies. Moreover, by inputting various stages of patient care into the system, the physician can analyze the trajectory of medical entropy, identifying key interventions that substantially impact patient outcomes. This aspect of the framework is particularly

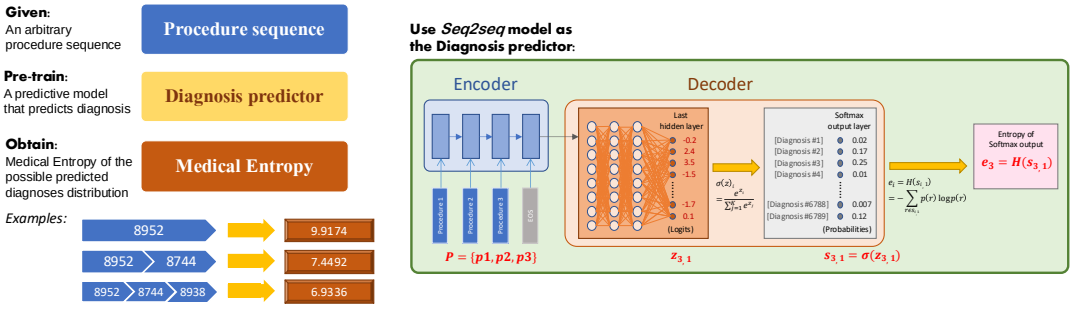

Figure 2: Illustration of an implementation of the proposed model by adopting the seq2seq model as the pre-trained model and an example of the entropy quantification of a procedure sequence.

beneficial for understanding the dynamic nature of medical decision-making over the course of a patient's hospital stay.

To make this framework workable in general use, we utilize the procedure sequences and diagnosis sequences in the form of ICD codes as the observations in the hospital and the diagnoses of the current stage. In this model, the sequence of procedure-related ICD codes constitutes the input for the MPM, and the sequence of diagnostic ICD codes forms its output. Notably, this does not mean that the proposed framework cannot apply features such as vital signs or lab results to the MPM. Instead, the main reason we adopt the plain ICD codes as training data instead of utilizing the more complex feature embedded in more information is that we focus on keeping the framework flexible and workable in medical practice.

To address the issue of the capacity of the predictive model in MPM, we apply non-knowledge-based decision-making for the CDSS strategy. Still, this predictive model can be substituted with a different model architecture, should it be better suited for other settings. Since the source and target data are in sequence, in the current study, we exploit the seq2seq models[15] as the diagnosis predictor in MPM to duplicate a success for the tasks in the fields of natural language processing (NLP)[16, 17, 18, 19].

2.3 Model Overview

Figure 2 illustrates the models we apply to implement the proposed framework and the examples of obtaining the entropy trend from a procedure sequence. In this study, we utilize the seq2seq model to pre-train the model for diagnosis predictor. To construct the entropy trend of a procedure sequence we desired, we feed the procedures to the encoder in order, then the decoder will output a distribution of possible diagnoses. The distribution here is determined by the softmax function of the last hidden layer of the seq2seq diagnosis predictor. Eventually, we then obtain the entropy based on a specific procedure combination. Take the procedure sequence 8952-8744-8938 as an example, we first obtain an entropy of 9.9174 by taking 8952 as the input of the diagnosis predictor. Further, we get the entropies of 7.4492 and 6.9336 by feeding the procedure combinations 8952, 8744 and 8952, 8744, 8938, respectively.

We adopt the seq2seq architecture model to predict suggestive diagnoses based on the procedures received so far due to its similar characteristics to data in Natural Language Processing (NLP)-related tasks. The data in both fields are in sequence and have varied lengths. However, there are still major differences between the two. Each word in a source sentence could be crucial for predicting the target sequence in machine translation. In contrast, the importance of a specific diagnosis code in diagnosis sequences ranks by its order. Moreover, the procedure code can possess multiple significances for different orders, combinations, and repetition frequencies.

On the other hand, the source and the target sequences in NLP-related tasks have vital cause-and-effect relationships. On the contrary, it does not work for the procedure-diagnosis relationship. Namely, the diagnosis for an admission can derive from the given procedures that a patient had received at a specific moment. Yet, the bond between procedure and diagnosis becomes insignificant when reasoning backward since identical procedure sequences can lead to various diagnosis combinations. This adverse impact could be amplified if admissions with a single procedure are the vast majority in the dataset. Due to the insufficiently informative input as a single procedure, it is explicit that it would be unreasonable for the model to predict the diagnosis with ample information.

In the admissions with the ICD-9 code we adopt in the MIMIC-IV dataset, there are 57,322 and 36,040 cases for the admissions with one and two procedures, respectively. This includes 59.9% of the admissions in total. Take ICD code "9925" as an example, there are 1,947 admissions among admissions with a single procedure. The code "9925" entails "Injection or infusion of cancer chemotherapeutic substance," which means that we know that the patients come to the hospital for chemotherapy. It is reasonable that the patients receive only one procedure; however, the diagnoses for these cases are widely different.

2.4 Seq2seq diagnoses predictor

To predict diagnoses based on procedures during patient admission, we use a Seq2Seq model. Both diagnosis and procedure codes are sequential, with procedure codes following a chronological order, while the order of diagnosis codes reflects the severity of the patient's condition. The Seq2Seq model offers a flexible approach to capture this dynamic by processing variable-length sequences, making it ideal for linking procedures to diagnoses.

The Seq2Seq framework, based on the encoder-decoder architecture [15], takes a source sequence as input and generates a target sequence as output. The encoder processes the input sequence to create a context vector, an embedding of the entire source sequence, which is then passed to the decoder to generate the target sequence. In our case, the encoder processes the procedure sequence, and its final hidden state serves as the context vector. The decoder uses this context to predict diagnoses sequentially, with each prediction informing the next.

We also integrate an attention mechanism [20, 21] to allow the decoder to weigh and focus on the most relevant hidden states from the encoder, enhancing prediction accuracy. To further improve training efficiency and prevent error propagation, we apply the teacher-forcing technique [22], which inputs the ground truth directly during training.

We focus on predicting the distribution of diagnoses at different stages of admission using ICD-9 codes. The encoder is fed with procedure ICD-9 codes, while the decoder outputs diagnosis ICD-9 codes. During evaluation, the softmax output of the decoder provides a probabilistic distribution of diagnoses, aiding in entropy-based analyses.

2.5 Entropy Quantification in Admissions

For the downstream task of the diagnoses predictor, we now focus on the input procedures and decoder's softmax output of the pre-trained model. Given a procedure sequence $P = \{p_1, p_2, ..., p_M\}$ with M procedures during an admission, we have the predicted diagnosis sequence $D_m = \{d_{m,1}, d_{m,2}, ..., d_{m,n}\}$ for every single step m in procedure sequence P, where $m = 1, \ldots, M$. For every diagnosis code d in the predicted diagnosis sequence D, the decoder outputs the diagnosis based on the softmax function $s_{m,k}$ of the last neural layer $z_{m,n}$ in the deep neural network of the decoder, where

$$s_{m,n} = \sigma(z_{m,n}), \quad and \tag{1}$$

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad for\ i = 1, ..., K \quad \text{and} \quad z = (z_1, ..., z_K) \in \mathbb{R}^K \tag{2}$$

$s_{m,k}$ is a vector that signifies the confidence of all the potential diagnoses candidates. The dimension of the $s_{m,n}$ is 6789 as there are 6789 codes in the dictionary of diagnoses ICD-9 codes in the MIMIC-IV database. Here, we choose the corresponding softmax output $s_{m,1}$ of the very first predicted diagnosis $d_{m,1}$, representing the probability distribution of all potential outcomes (i.e., diagnosis code). This means that the softmax output $s_{m,1}$ is directly determined by the context vector c that encodes all procedures in order, which is used by the decoder to predict the most important diagnosis code.

As a consequence, we obtain the sequence of softmax output $S = \{s_{1,1}, s_{2,1}, ..., s_{M,1}\}$ as the confidence of possible diagnoses by given cumulative input procedure sequence $CP_m = \{p_1, p_2, ..., p_m\}$. During the hospitalization of a given patient, from admission to discharge, we update the corresponding distribution after receiving every procedure. To further analyze the trends of the distributions, we quantify the uncertainty of the predicted diagnosis code by calculating the information entropy of this distribution. In this way, we have the entropy sequence $E = \{e_1, e_2, ..., e_m\}$, where

Table 1: Model performance of Diagnosis Predictor as F1-score, Jaccard Index, and First-N-Accuracy among different seq2seq model architectures.

Attention-based seq2seq Diagnosis Predictor		F1-score	Jaccard Index	First-N-Accuracy		
				N=1	N=2	N=3
no teacher forcing	1-layer	**1.08e-02**	**5.45e-03**	0.2717	0.2236	0.1951
	2-layer	9.94e-03	5.00e-03	0.2488	0.2019	0.1766
	3-layer	6.35e-03	3.18e-03	0.1871	0.1669	0.1512
with teacher forcing	1-layer	8.66e-03	4.35e-03	**0.2884**	**0.2281**	**0.2025**
	2-layer	8.02e-03	4.02e-03	0.2321	0.1986	0.1847
	3-layer	3.14e-03	1.57e-03	0.1505	0.1440	0.1341

$$e_i = H(s_{i,1}) = -\sum_{r \in s_{i,1}} p(r) \log p(r) = \mathbb{E}[-\log p(s_{i,1})] \tag{3}$$

Despite the semantic differences between the diagnosis-procedure data and NLP-related data, we assume that the seq2seq model still works as a diagnosis predictor since the diagnosis-procedure settings share similar sequential behavior with typical seq2seq scenarios in NLP tasks. To ameliorate the error propagation between the pre-trained model and the downstream entropy quantification, it is necessary to expose how satisfyingly the diagnosis predictor could go among different model architectures. Hence, defining the pertinent and reliable evaluation metrics for the diagnosis predictor is crucial and needed.

2.6 Model Performance Evaluation

Various evaluation metrics are defined in the field of NLP to assess diverse aspects. However, widely used evaluation metrics for the NLP area may not necessarily be suitable in our scenario. This is because, unlike typical NLP models, the diagnosis predictor cares less about the N-gram terms of the predicted diagnosis sequence. Many NLP evaluation metrics have a penalty term to ensure the model decodes longer sequences, which is not applicable in our case. Moreover, the order of the diagnosis sequence has imperative information about how important it is to the patient's condition, which is not the case in a natural language sequence.

For reasons stated above, instead of utilizing metrics such as the Word error rate (WER), Bilingual Evaluation Understudy (BLEU)[23], or their variants[24], we utilize the f1 score and Jaccard index[25] to examine the model performance of the seq2seq diagnosis predictor. Furthermore, we also propose First-N-accuracy to assess the model's capability of precisely predicting the most decisive diagnoses. That is to say, if N equals three, the First-N-accuracy will be counting the percentage of the first three predicted diagnoses appearing in the first three diagnoses in the ground truth.

To further assess the model's ability to handle uncertainty, we focus on the entropy trend throughout entire admissions from a global perspective. The dataset's procedure and diagnosis sequence pairs, reflecting real-world medical decisions, serve as ideal cases for examining entropy reduction. Every action by physicians is considered an attempt at reducing entropy. Despite potential fluctuations in individual entropy trends due to the unpredictable nature of hospital scenarios, the overall entropy trend should consistently decrease, regardless of the patient outcomes, be it discharge or passing away. This approach allows us to evaluate the model's performance in a dynamic, real-world medical setting.

3 Results

Table 1 summarizes the performance of the proposed framework for various model architectures. As can be seen in Table 1, the simplest 1-layer model surpasses other models in performance across all the metrics used. A noteworthy observation is that the simpler 2-layer model significantly outperforms the 3-layer model, while the performance gap between the 1-layer model and the 2-layer model is relatively modest. Additionally, to compare the performance of these three model architectures with and without teacher-forcing to investigate its impact on the overall performance. We trained the models with identical hyperparameters, where the only difference was whether the teacher-forcing feature was used or not[26]. The outcomes revealed that the models with teacher-forcing consistently mirrored the performance trends of models without teacher forcing. Furthermore, it was observed that the addition

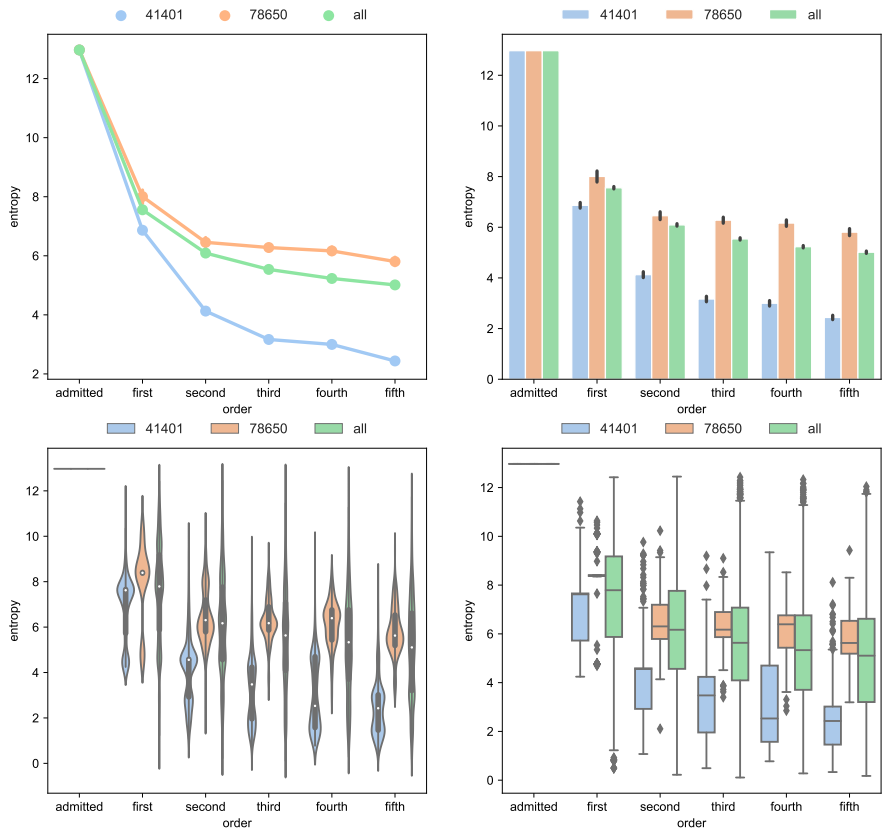

Figure 3: Trends of average entropy for admission cases with five procedures. The three colors show the entropy trends clustered by their primal diagnosis, which are diagnosis ICD-9 code 41401 (Coronary atherosclerosis of native coronary artery), 78650 (Chest pain), and all admissions with 5 procedures.

of teacher forcing did not enhance the performance compared to the original models in the tested scenarios.

Figure 3 shows the average entropy trend of 3 different clusters, grouped by their primal diagnosis ICD codes. To better grasp the trends of uncertainty drops, we focus on the cases with the same length. Specifically, we selected the cases of patients who received a total of five procedures during their admissions. In the figure, "All" indicates the average entropy trends for all diseases, whereas "41401" and "78650" are groups clustered by their primary diagnosis, "Coronary atherosclerosis of native coronary artery" and "Chest pain," respectively. The total cases among the three clusters, All, 41401, and 78650, are 8,705, 679, and 220 admissions, respectively.

As shown in Figure 3, all the admissions have the same initial entropies at the very beginning of the admissions. That is the condition when the patients arrive at the hospital without receiving any procedure. To provide a reference for comparisons, we assume that the 6,984 potential procedures are equally distributed. In this case, the initial entropy is set to 12.76. In another case, this initial entropy will be 9.43 if we assume the potential procedures are distributed by their frequencies (estimated based on the MIMIC-IV dataset). In all the figures in this paper, we set the initial entropy to 12.76, taking that the initial distribution of procedures is uniform.

Note that the "entropy trend" mentioned in this paper reveals the entropy of the distributions of the potential diagnoses when receiving a new procedure made by the physician during the hospitalization. The entropy trend reflects how the medical entropy of patients changes from admission to discharge, as a

Table 2: Entropy of most frequent first N=1 procedure.

#	ICD-9 code	Cases	Frequency	Entropy After Receiving 1st procedure
1	66	4528	2.90%	9.8625
2	8938	4036	2.59%	7.0969
3	741	3443	2.21%	7.6941
4	8952	3426	2.20%	9.9164
5	7569	3399	2.18%	6.3773
6	3893	3242	2.08%	9.1968
7	9925	3111	1.99%	7.2922
8	3995	3019	1.94%	8.4310
9	9671	2790	1.79%	10.2779
10	5491	2703	1.73%	8.4703

result of receiving procedures. If the patient received three procedures during the entire hospitalization, for example, procedures A, B, and C, we would have four entropies in their entropy trend. The entropy trend includes the initial entropy without receiving any procedure and the respective entropies of the predicted diagnosis distributions by given procedure sequence $\{A\}$, $\{A, B\}$, and $\{A, B, C\}$. Overall, Figure 3 clearly shows that the average entropy tends to decrease as the number of received procedures increases for all three clusters.

Next, we investigated the entropy drop for the most frequent first N procedures (for N=1, 2, and 3) since the very start of admission. These results are summarized in tables 2 to 4. Information theory states that the more information we know about the world (or the patient's state, in this study), the smaller entropy we get as a result (i.e., less uncertainty regarding the patient's diagnosis in the current study). The drops of the entropies reveal the reduction of the uncertainties. As Figure 3 shows, the entropy tends to decrease when every upcoming procedure arrives, it is imperative to examine how entropy reduction acts among various procedures.

Table 2 demonstrates the entropy drops of the ten most frequent first procedures at the beginning of the admission. The ICD-9 procedure code "7569"(ranked #5) has the lowest entropy of 6.37, describing "Repair of other current obstetric laceration." In contrast, the ICD-9 procedure code "9671"(ranked #9) has the highest entropy of 10.27 for "Continuous invasive mechanical ventilation for less than 96 consecutive hours."

To make the proposed framework applicable to real-world medical use, the entropy quantification of a specific procedure sequence in any situation has to be more explainable and informative to physicians instead of being a "black box". In other words, it is indispensable to analyze the entropy trends individually. For better demonstration, we choose examples of actual cases in the MIMIC-IV dataset with the same number of total received procedures and with similar final diagnoses.

Figure 4 presents the entropy trends of three cases with sepsis diagnosis, each undergoing six procedures. Two cases were discharged from the hospital, and one passed away. These cases were selected from the MIMIC-IV dataset due to their similar final diagnoses and received procedures. The entropy trends are depicted for each admission: Admission #1 (blue dotted line), Admission #2 (orange line), and Admission #3 (green line).

In Admission #1, a significant drop in entropy is observed after the first procedure, "4562 - Other partial resection of small intestine." However, subsequent procedures do not further reduce the entropy significantly. In Admission #2, the entropy initially drops dramatically after the first procedure, "5459 - Other lysis of peritoneal adhesions," but rises after procedures 9604 and 9671, before decreasing again following procedure 9915. In contrast, Admission #3 shows a gradual decrease in entropy, indicating a more consistent reduction in uncertainty with each procedure.

4 Discussion

In this study, the focus was on assessing a framework designed to measure and manage medical entropy during various stages of hospital admission. The findings suggest that the framework has potential in quantifying and managing uncertainty in clinical decision-making, adapting to different stages of a patient's hospital stay.

The analysis of trends in Figure 3 demonstrates the effectiveness of the proposed entropy quantification method in modeling changes in patient conditions. Among the three entropy trends, the trend with all cases is the smoothest. Comparing the differences between the two diagnosis ICD-9 codes, "78650", representing "chest pain" doesn't decline as much as "41401" which denotes "Coronary atherosclerosis

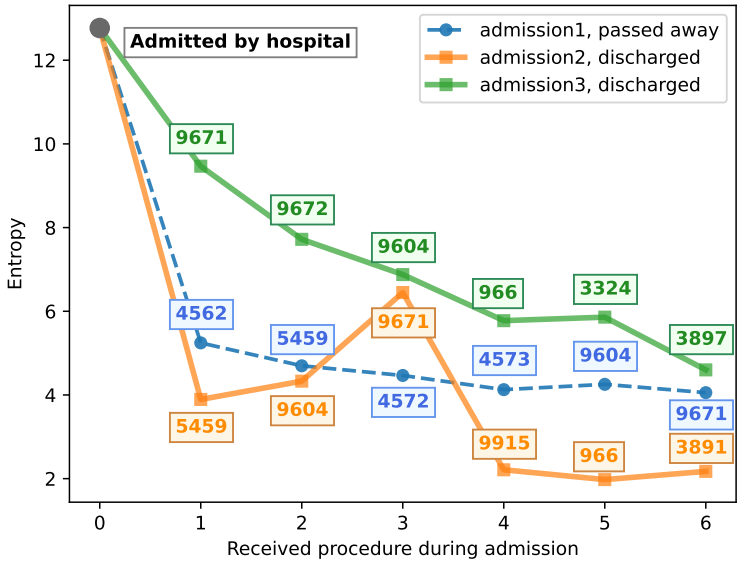

Figure 4: Entropy trends of three admissions with sepsis diagnosis.

Table 3: Entropy of most frequent first N=2 procedures

#	ICD-9 code	Cases	Frequency	Entropy After Receiving	
				1st procedure	2nd procedure
1	0066 3607	2920	18.72‰	9.8504	7.3538
2	8952 8938	1507	9.66‰	9.9137	7.3577
3	8952 8744	1323	8.48‰	9.9196	7.4492
4	3722 8856	1316	8.44‰	10.3795	4.5525
5	0066 3606	930	5.96‰	9.8362	8.8169
6	9671 9604	925	5.93‰	10.2844	9.3869
7	8938 8952	867	5.56‰	7.0969	7.1663
8	3950 3990	696	4.46‰	9.8251	9.3711
9	7569 734	631	4.05‰	6.3829	5.8259
10	9604 9671	598	3.83‰	10.4688	9.5779

of native coronary artery" does. Moreover, "78650" has a more significant standard deviation. This could be attributed to the fact that the conditions related to chest pain are sometimes more uncertain than coronary artery diseases. In addition, the boxplot showing the entropies of the first procedure of "78650" seems to be very concentrated since more than 99% of the physicians' order an electrocardiogram, which is known as the fastest and most straightforward test to evaluate heart conditions.

The findings from Tables 2, 3, and 4 provide insights into how medical procedures influence diagnostic uncertainty. For instance, Table2 highlights differences in uncertainty between procedures like "7569" (related to delivery) and "9671" (used for respiratory failure). The former typically leads to more straightforward diagnoses, while the latter involves more uncertainty due to various unknown factors. Table 3 examines entropy drops for the first two procedures after admission. Interestingly, procedures "0066" (PTCA) paired with either "3607" (drug-eluting stent) or "3606" (non-drug-eluting stent) show that the framework can differentiate between these combinations. Additionally, the "3722-8856" pair

Table 4: Entropy of most frequent first N=3 procedures

#	ICD-9 code	Cases	Frequency	Entropy After Receiving		
				1st procedure	2nd procedure	3rd procedure
1	8952 8744 8938	989	6.3‰	9.9174	7.4492	6.9336
2	0066 3607 3722	887	5.7‰	9.7971	7.3538	4.3214
3	8952 8938 8744	769	4.9‰	9.9189	7.3577	7.1235
4	0066 3607 0045	705	4.5‰	9.8413	7.3538	6.0209
5	8938 8952 8744	489	3.1‰	7.1029	7.1663	6.6327
6	3612 3615 3961	372	2.4‰	9.0666	3.4497	1.4552
7	3613 3615 3961	359	2.3‰	9.9924	4.3351	2.0094
8	0066 3607 0040	332	2.1‰	9.8252	7.3538	6.2223
9	0066 3606 3722	306	2.0‰	9.8325	8.8169	5.6517
10	0066 3607 0046	286	1.8‰	9.8265	7.3538	5.6552

("Left heart catheterization" and "Coronary arteriography") shows the largest entropy drop, reflecting the detailed information provided by coronary arteriography. Table 4 extends this analysis to the first three procedures. The entropy for triplet "3612-3615-3961" (critical coronary bypass surgeries) drops significantly, while triplet "8952-8744-8938" (preliminary heart-related exams) shows only a modest decline. This contrast underscores the importance of life-saving procedures in reducing uncertainty. The entropy trends in Figure 4 demonstrate varied impacts of procedures on patient outcomes. In Admission #1, initial entropy reduction reflects the effectiveness of the first procedure, but later procedures contribute little, suggesting the severity of the condition. Admission #2 shows fluctuating entropy, indicating varying effectiveness of the procedures, with a final drop suggesting more informative interventions. Admission #3 displays a steady decrease in entropy, indicating consistent reduction in uncertainty with each procedure.

These findings highlight the importance of procedure sequencing and selection in managing complex cases like sepsis, where uncertainty is high. Understanding entropy trends can help physicians make more informed decisions, offering a more transparent and explainable approach to medical practice. The proposed framework, tested in both simulated and real-world scenarios, demonstrates its potential usefulness across diverse medical cases. The use of ICD codes addresses issues like missing values and outliers, promoting standardization and reliability. These insights have implications for enhancing Clinical Decision Support Systems (CDSS) by integrating data-driven AI/ML models. By quantifying medical entropy, the framework may assist clinicians in making robust, timely decisions, particularly in high-uncertainty situations. However, this study has some limitations. The seq2seq model, while effective here, may not be optimal in all clinical scenarios, and further exploration of alternative models is necessary. Additionally, relying on ICD codes may limit the depth of analysis, as they are generated after patient stays. Their use here is primarily for validation.

Future research could leverage Electronic Health Records (EHR) audit logs to provide a more granular measure of entropy throughout hospital visits, capturing real-time decision-making processes. EHR audit logs offer rich, sequential data that could enhance the predictive power and applicability of our framework in clinical environments [27, 28, 29].

In summary, this study introduces a novel framework for addressing medical uncertainty in clinical settings. While the approach shows promise in various scenarios, further research is needed to explore its broader impact on healthcare, particularly in advancing Clinical Decision Support Systems.

5 Conclusion

In conclusion, this study presents an innovative framework for quantifying and managing medical entropy during hospital admissions. The framework effectively quantifies and adapts to the dynamic nature of clinical decision-making, providing a nuanced understanding of patient-specific variables through the use of entropy quantification. Utilizing the MIMIC-IV dataset and focusing on ICD codes, the study demonstrates how the proposed framework can aid clinicians in reducing diagnostic uncertainty, particularly in complex and time-sensitive medical scenarios. While the seq2seq model used has shown promise, the study acknowledges its potential limitations in certain clinical settings and the need for further exploration of alternative predictive models. Overall, this research contributes significantly to the enhancement of Clinical Decision Support Systems, offering a novel approach to handling medical uncertainty and improving patient care outcomes.

References

[1] Rahul Alam, Sudeh Cheraghi-Sohi, Maria Panagioti, Aneez Esmail, Stephen Campbell, and Efharis Panagopoulou. Managing diagnostic uncertainty in primary care: a systematic critical review. *BMC Family Practice*, 18(1):1–13, 2017.

[2] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine*, 3(1):17, 2020.

[3] Mohamed Benzeghiba, Renato De Mori, Olivier Deroo, Stephane Dupont, Teodora Erbes, Denis Jouvet, Luciano Fissore, Pietro Laface, Alfred Mertins, Christophe Ris, et al. Automatic speech recognition and speech variability: A review. *Speech communication*, 49(10-11):763–786, 2007.

[4] Jeff Kabachinski. A look at clinical decision support systems. *Biomedical Instrumentation & Technology*, 47(5):432–434, 2013.

[5] Anna Markella Antoniadi, Yuhan Du, Yasmine Guendouz, Lan Wei, Claudia Mazo, Brett A Becker, and Catherine Mooney. Current challenges and future opportunities for xai in machine learning-based clinical decision support systems: a systematic review. *Applied Sciences*, 11(11):5088, 2021.

[6] Qian Xu, Wenzhao Xie, Bolin Liao, Chao Hu, Lu Qin, Zhengzijin Yang, Huan Xiong, Yi Lyu, Yue Zhou, Aijing Luo, et al. Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: A systematic review. *Journal of Healthcare Engineering*, 2023, 2023.

[7] Winnie Chen, Claire Maree O'Bryan, Gillian Gorham, Kirsten Howard, Bhavya Balasubramanya, Patrick Coffey, Asanga Abeyaratne, and Alan Cass. Barriers and enablers to implementing and using clinical decision support systems for chronic diseases: a qualitative systematic review and meta-aggregation. *Implementation Science Communications*, 3(1):1–20, 2022.

[8] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[9] Robert B Ash. *Information theory*. Courier Corporation, 2012.

[10] Shuhan He, Paul Chong, Byung-Jun Yoon, Pei-Hung Chung, David Chen, Sammer Marzouk, Kameron C Black, Wilson Sharp, Pedram Safari, Joshua N Goldstein, et al. Entropy removal of medical diagnostics. *Scientific Reports*, 14(1):1181, 2024.

[11] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

[12] World Health Organization et al. International classification of diseases—ninth revision (icd-9). *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, 63(45):343–344, 1988.

[13] Toni Henderson, Jennie Shepheard, and Vijaya Sundararajan. Quality of diagnosis and procedure coding in icd-10 administrative data. *Medical care*, pages 1011–1019, 2006.

[14] Aitziber Atutxa, Alicia Pérez, and Arantza Casillas. Machine learning approaches on diagnostic term encoding with the icd for clinical documentation. *IEEE journal of biomedical and health informatics*, 22(4):1323–1329, 2017.

[15] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014.

[16] KR1442 Chowdhary and KR Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020.

[17] Diksha Khurana, Aditya Koli, Kiran Khatter, and Sukhdev Singh. Natural language processing: State of the art, current trends and challenges. *Multimedia tools and applications*, 82(3):3713–3744, 2023.

[18] Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

[19] Hana Yousuf, Michael Lahzi, Said A Salloum, and Khaled Shaalan. A systematic review on sequence-to-sequence learning with neural network and its models. *International Journal of Electrical & Computer Engineering (2088-8708)*, 11(3), 2021.

[20] Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*, 2015.

[21] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[22] Boris Belousov and Jan Peters. Entropic regularization of markov decision processes. *Entropy*, 21(7):674, 2019.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[24] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[25] Luciano da F Costa. Further generalizations of the jaccard index. *arXiv preprint arXiv:2110.09619*, 2021.

[26] Alex M Lamb, Anirudh Goyal ALIAS PARTH GOYAL, Ying Zhang, Saizheng Zhang, Aaron C Courville, and Yoshua Bengio. Professor forcing: A new algorithm for training recurrent networks. *Advances in neural information processing systems*, 29, 2016.

[27] Julia Adler-Milstein, Jason S Adelman, Ming Tai-Seale, Vimla L Patel, and Chris Dymek. Ehr audit logs: a new goldmine for health services research? *Journal of biomedical informatics*, 101:103343, 2020.

[28] Seunghwan Kim, Sunny S Lou, Laura R Baratta, and Thomas Kannampallil. Classifying clinical work settings using ehr audit logs: A machine learning approach. *American Journal of Managed Care*, 29(1), 2023.

[29] Nandita Bhaskhar, Wui Ip, Jonathan H Chen, and Daniel L Rubin. Clinical outcome prediction using observational supervision with electronic health records and audit logs. *Journal of Biomedical Informatics*, 147:104522, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made about developing a neural machine translation framework to predict medical diagnoses and quantify uncertainty. The claims align well with the methods and results sections, reflecting the contributions made.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations, particularly regarding the reliance on ICD-9 codes and the seq2seq model's suitability for clinical scenarios. It also notes that further exploration of alternative predictive models is necessary (Discussion and Conclusion sections).

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not present formal theoretical results or proofs, as it focuses on empirical and applied machine learning techniques.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper provides detailed descriptions of the data sources (MIMIC-IV), the model architecture (seq2seq), and entropy quantification methods. This information is sufficient for reproducing the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The MIMIC-IV dataset used in this study is a restricted-access resource. To access the data, users must apply for credentialed access by completing specific training and obtaining approval. Due to these restrictions, we are unable to provide open access to the dataset or code used for the experiments. However, the instructions for accessing the data and reproducing the experiments are detailed, and any credentialed user can follow the process described to replicate the results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper describes the data (ICD-9 codes), the seq2seq model used, and the evaluation metrics (F1 score, Jaccard Index, and First-N accuracy). It specifies the number of admissions and procedures in the dataset.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: This paper focuses on proposing the framework of medical entropy rather than emphasizing model performance. Instead of reporting traditional experiment statistical significance, we provide an in-depth discussion of the medical entropy trends, with error bars and confidence intervals included to represent variability. Specifically, Figure 3 illustrates the average entropy trends across different patient admissions, accompanied by error bars

that reflect the standard deviation of entropy values at each stage. These visualizations help quantify the uncertainty reduction as procedures are performed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper specifies that all experiments were conducted on a CPU using an Apple MacBook Pro with the M1 Max chip. This setup provided sufficient computational power for processing ICD-9 data and running machine learning models. The information ensures reproducibility on accessible hardware and indicates that no high-performance computing resources, such as GPUs or cloud servers, were necessary.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research aligns with NeurIPS' ethical guidelines, as the paper focuses on improving clinical decision-making through AI while acknowledging potential limitations and transparency concerns.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses the potential positive societal impacts, particularly the enhancement of clinical decision support systems and the reduction of uncertainty in diagnosis. The discussion section addresses both positive and potential negative consequences.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The models presented in the paper are unlikely to pose high risks for misuse, as they are tailored toward medical diagnosis and uncertainty quantification.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: The paper uses the MIMIC-IV database, which is a public dataset. The original source of the data is credited, and the relevant references are provided.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification:The paper does not introduce new datasets or models that require documentation.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing or direct research with human subjects. It uses de-identified, publicly available clinical data.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: As the paper relies on de-identified data from the MIMIC-IV dataset, IRB approval is not required.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.