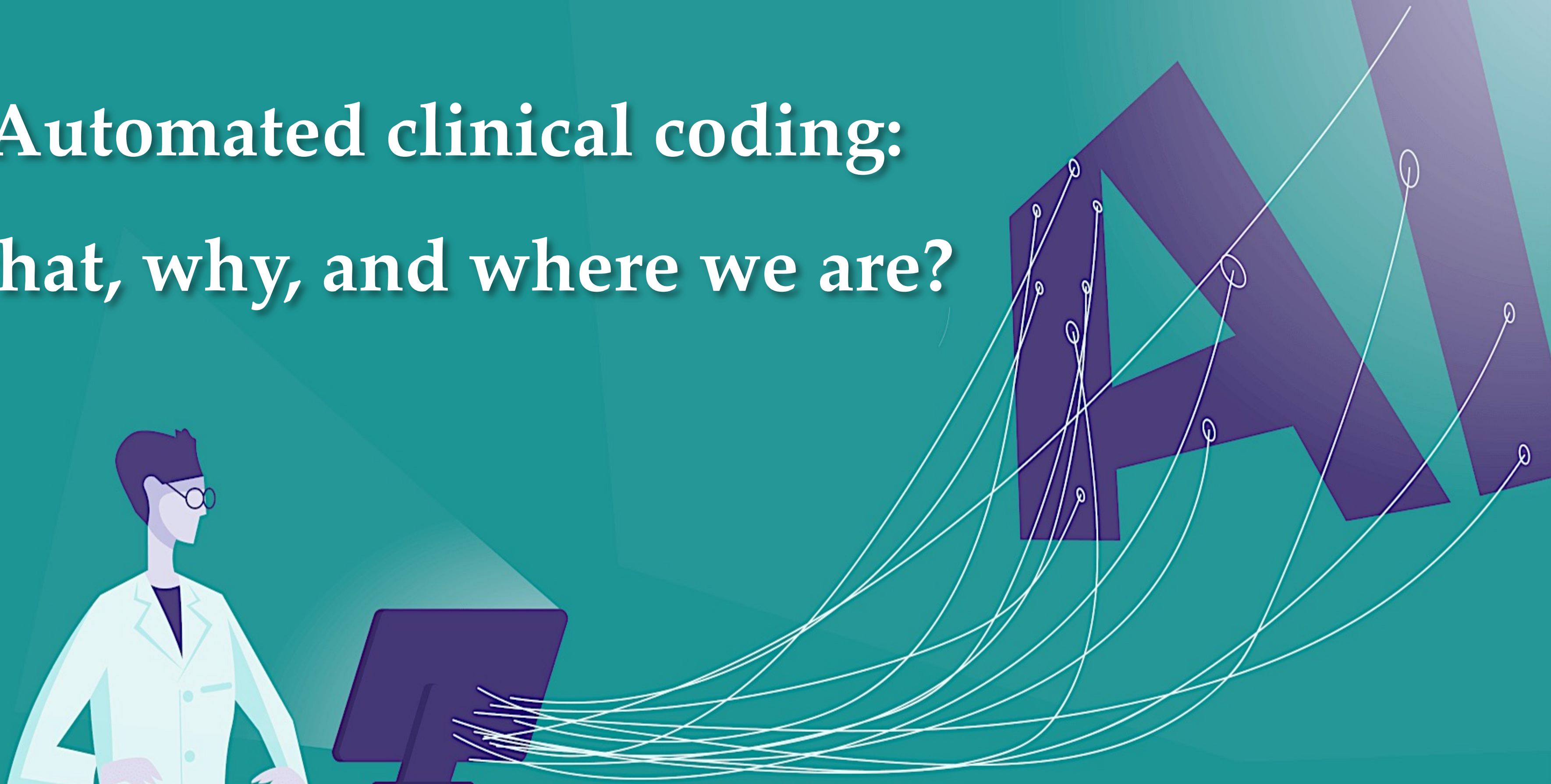
Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

Automated clinical coding: what, why, and where we are?



Teerapong Aramruang

Literature

PERSPECTIVE OPEN

Check for updates

Automated clinical coding: what, why, and where we are?

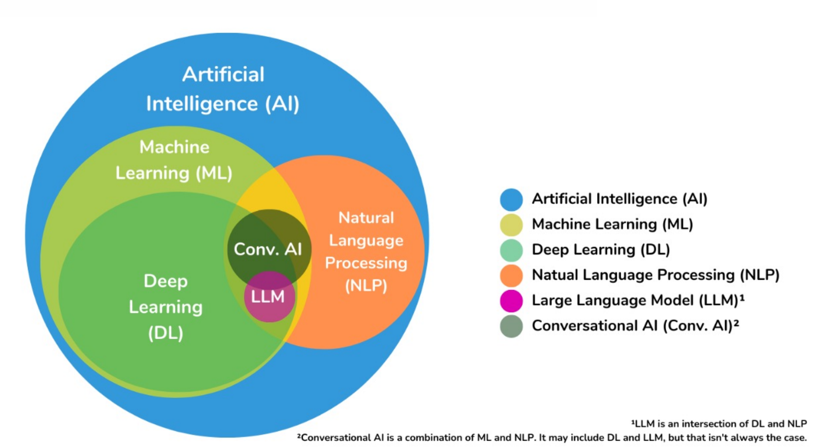
Hang Dong[1,2]✉, Matúš Falis[3], William Whiteley[4], Beatrice Alex[3,5], Joshua Matterson[6,7], Shaoxiong Ji[8], Jiaoyan Chen[2] and Honghan Wu[9]✉

Clinical coding is the task of transforming medical information in a patient's health records into structured codes so that they can be used for statistical analysis. This is a cognitive and time-consuming task that follows a standard process in order to achieve a high level of consistency. Clinical coding could potentially be supported by an automated system to improve the efficiency and accuracy of the process. We introduce the idea of automated clinical coding and summarise its challenges from the perspective of Artificial Intelligence (AI) and Natural Language Processing (NLP), based on the literature, our project experience over the past two and half years (late 2019–early 2022), and discussions with clinical coding experts in Scotland and the UK. Our research reveals the gaps between the current deep learning-based approach applied to clinical coding and the need for explainability and consistency in real-world practice. Knowledge-based methods that represent and reason the standard, explainable process of a task may need to be incorporated into deep learning-based methods for clinical coding. Automated clinical coding is a promising task for AI, despite the technical and organisational challenges. Coders are needed to be involved in the development process. There is much to achieve to develop and deploy an AI-based automated system to support coding in the next five years and beyond.

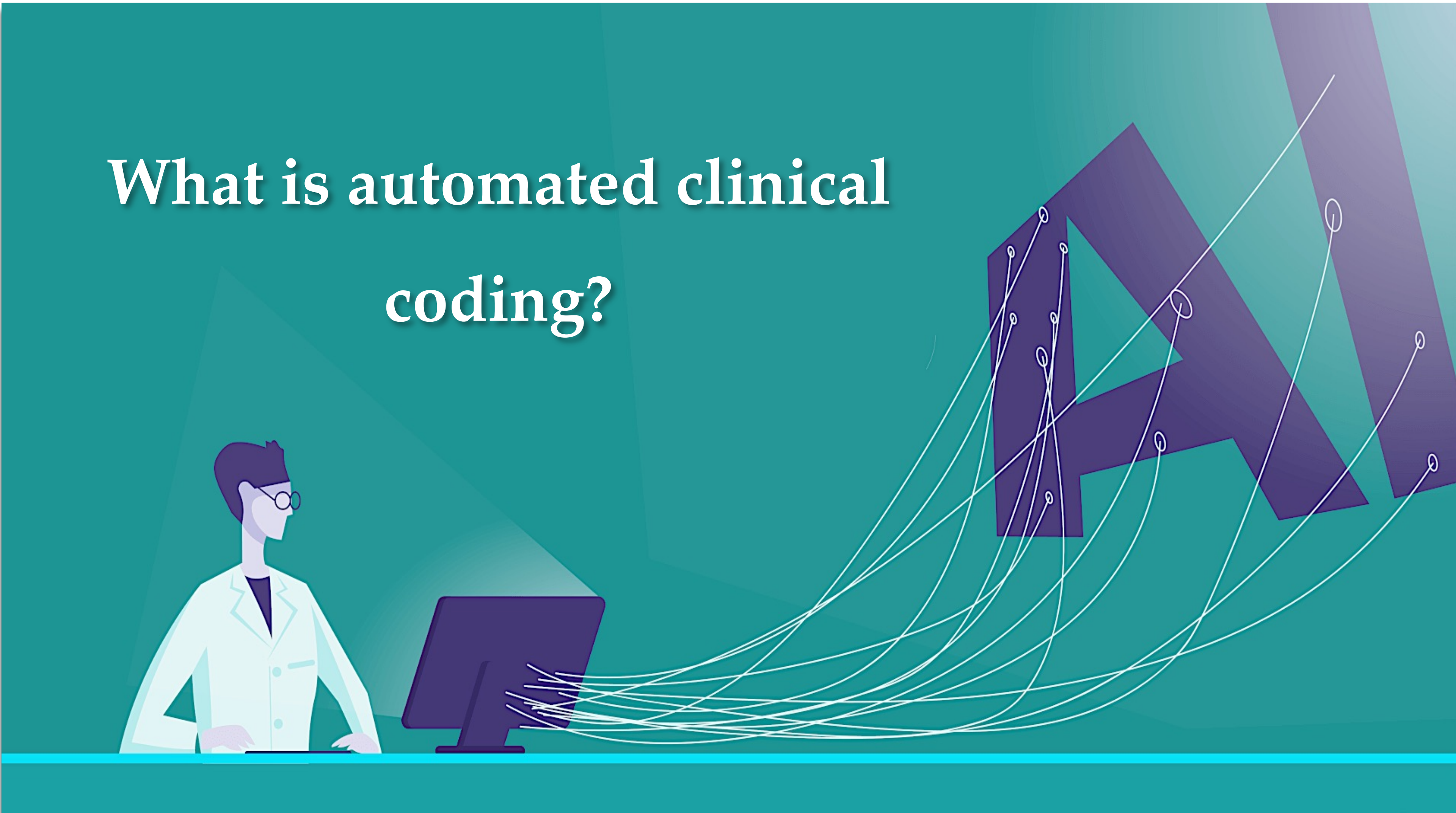
npj Digital Medicine (2022)5:159 ; https://doi.org/10.1038/s41746-022-00705-7

Outline

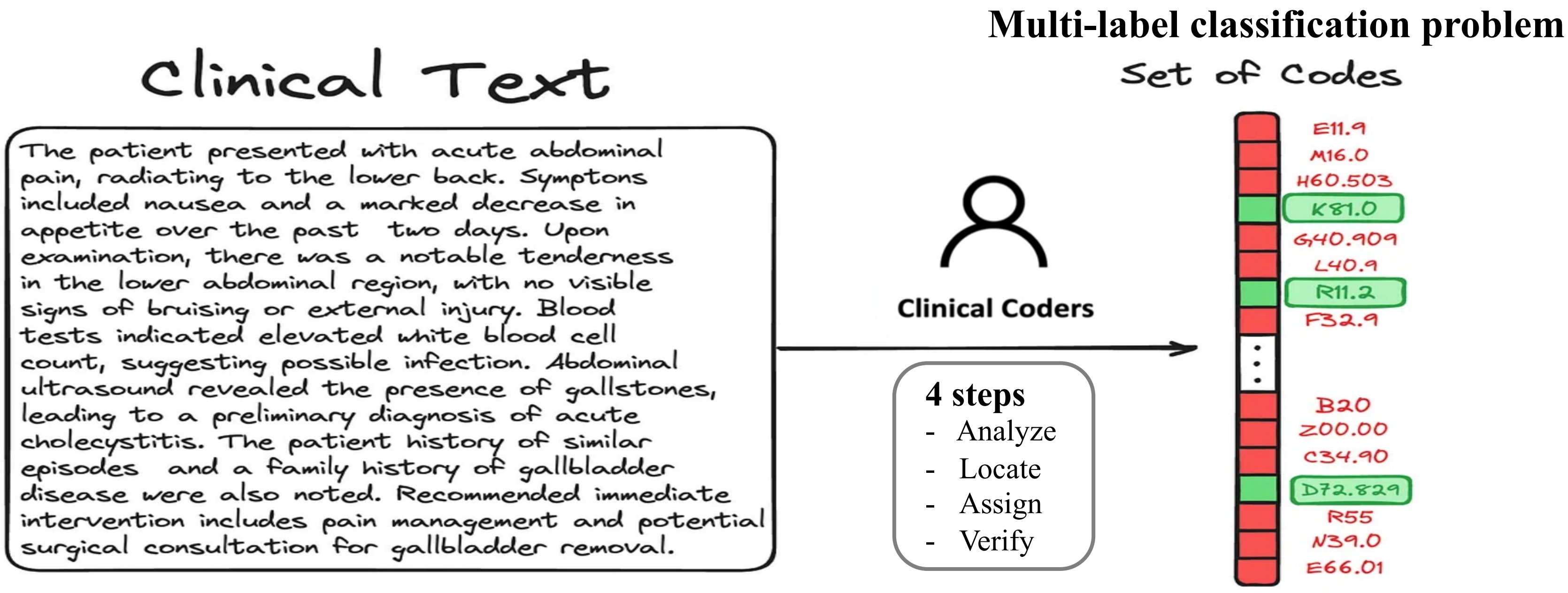
- What is automated clinical coding?
- Why do we need automated clinical coding?
- Why is automated coding a complex problem to solve?
- What are the potential challenges to address for automated clinical coding?
- How do state-of-the-art deep learning models work so far?
- Recently approach: Generative artificial intelligence (AI)



What is automated clinical coding?



What is automated clinical coding?



- **Clinical coding** is the task of transforming medical records, usually presented as free texts written by clinicians, into structured codes in a classification system like ICD-10.
- **The purpose is** to provide consistent and comparable clinical information across units of care: health improvement, healthcare planning and policy, epidemiology, and billing.
- **Automated clinical coding** is the idea that clinical coding may be automated by computers using AI techniques, e.g., natural language processing (NLP) and machine learning.

Why do we need automated clinical coding?

Why do we need automated clinical coding?

I. Manual coding is time-consuming

- A clinical coder in NHS Scotland usually codes about **60 cases a day (equivalent to 7-8 min for each case)** and an NHS coding department of around 25-30 coders usually codes **over 20,000 cases per month**.
- Huge number of cases must be coded, **which can take several months (e.g., over a year).**

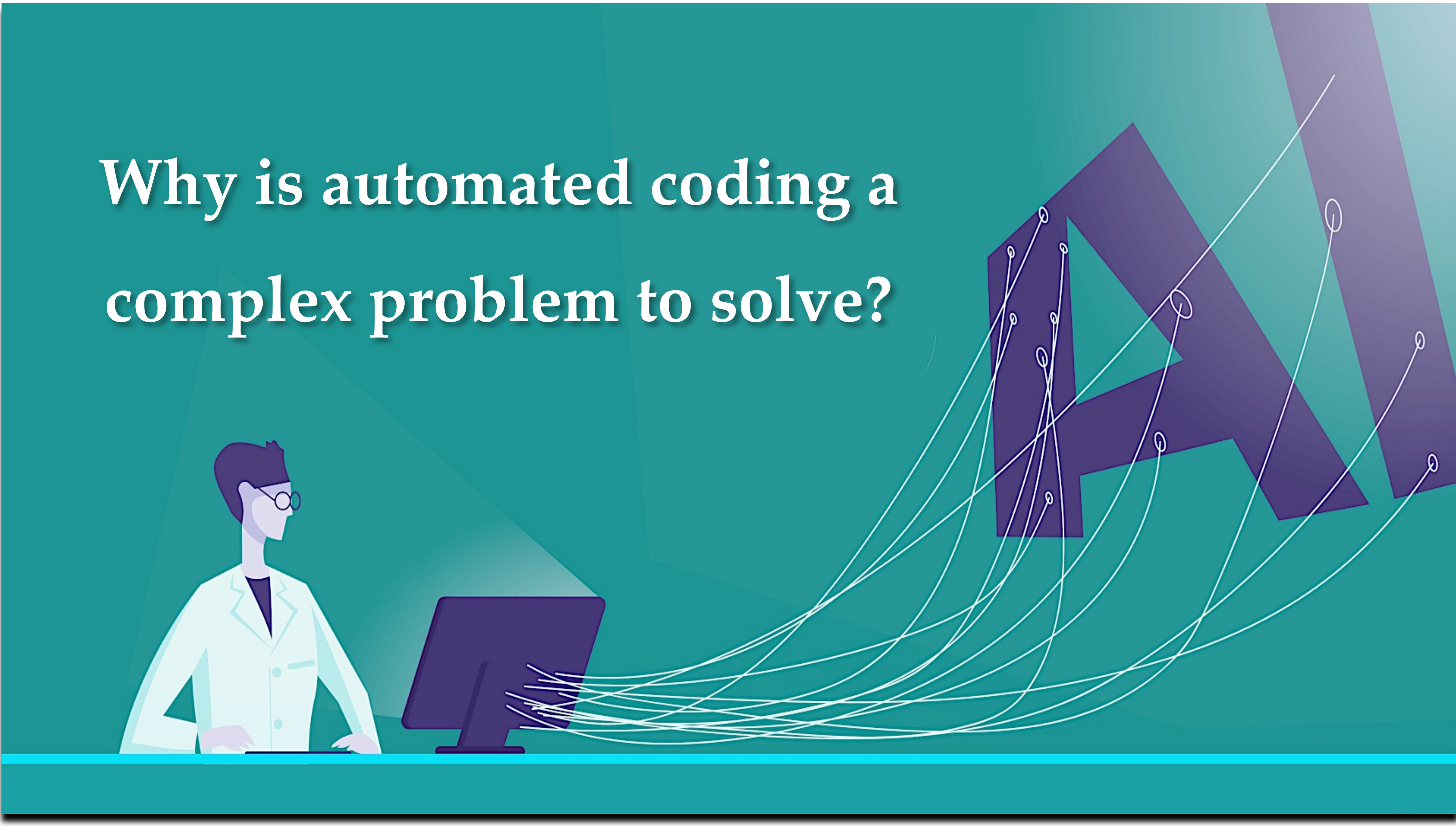
II. Manual coding may be prone to errors

- In UK, **the accuracy is 83%** with a large variance among studies (50-98%).
- In Scotland, **the accuracy is 92.5% for 3-digit code** and **88.8% for 4-digit code.**
- **Not perfect and under-coding issues:** incompleteness in a data of patient, subjectivity in choosing the diagnosis codes, lack of coding expertise, or data entry errors.

AI technologies (e.g., NLP), automated coding has the potential to better support clinical coders to facilitate the administration and management of clinical records in the hospital and medical research.

NHS, National Health Service

npj Digital Medicine. 2022;5(1):159.


Why is automated coding a complex problem to solve?
AI

Why is automated coding a complex problem to solve?

I. Clinical documents are variously structured, notational, lengthy, and incomplete

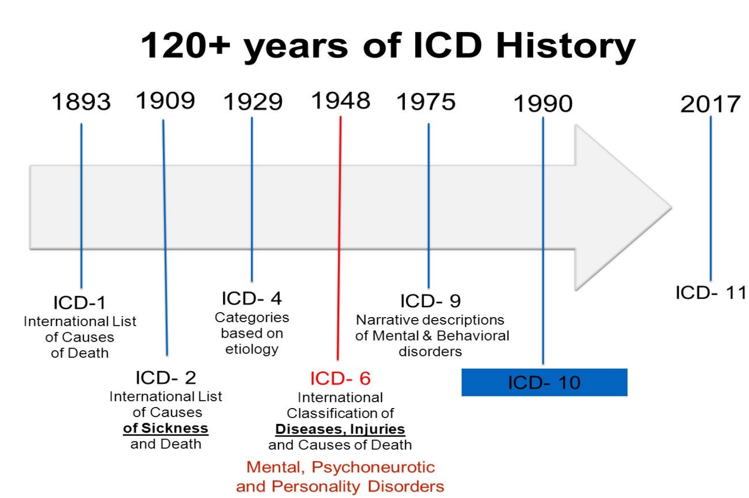
- Clinical coding requires the understanding of texts in clinical documents, which is usually **different from other types of documents** like publications or texts from social media.
- **Various document structures**, they can be lengthy (e.g., on average around 1500 words in only the discharge summaries in a US intensive care data)
- **Various abbreviations and symbols** (e.g., a [xx] y/o M w/ Hep C, HTN, CKD, a/w HTN emergency in a discharge summary and the use of "?" to denote uncertainty and "+" to denote a positive test).
- Coding requires the **understanding of the entirety of a patient's records**, which includes multiple types of documents (e.g., discharge summaries, radiology reports, pathology reports, etc.). These documents are not always in a structured format and are sometimes incomplete or missing.

npj Digital Medicine. 2022;5(1):159.

Why is automated coding a complex problem to solve?
(cont.)

II. Classification systems used for coding are complex and dynamic

- **Various codes involved and each code is used for different purposes** (e.g., DRGs for billing, SNOMED CT, ORDO, ICD-11 for diagnosis)
- **The ICD-10-CM system has around 68,000 diagnosis codes** in a large hierarchy, 5 times more than the previous ICD-9-CM.
- **The ICD-11 system** (or ICD-11-MMS) is launched in early 2022, but is yet to be used in practice, contains around 17,000 unique codes for injuries, diseases and causes of death, under-pinned by more than 120,000 codable terms and can code more than 1.6 million clinical situations using code combinations.
- Besides, the ICD systems are **updated regularly**.
- Thus, automated clinical coding needs to work with dynamic and complex classification systems.

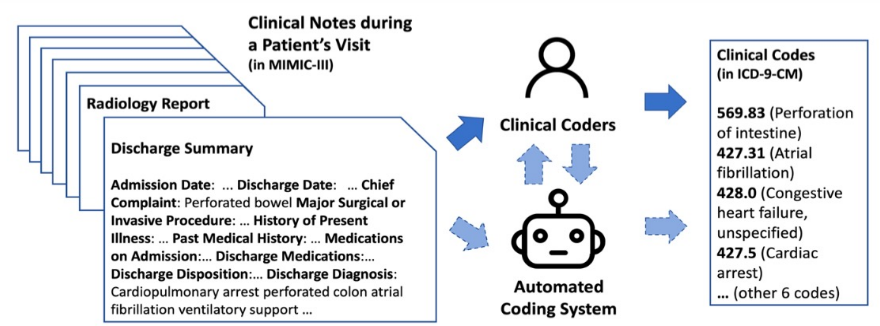


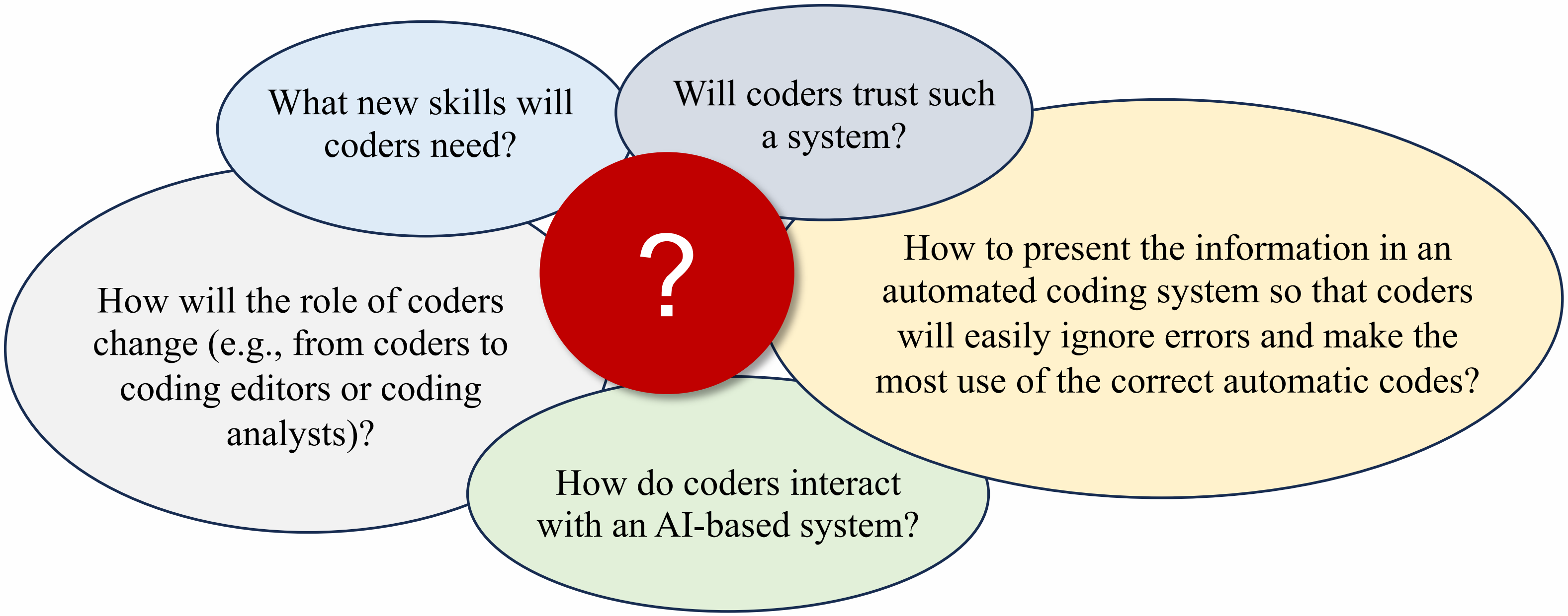
DRGs, Diagnosis-Related Groups; ICD, International Classification of Diseases

Why is automated coding a complex problem to solve? (cont.)

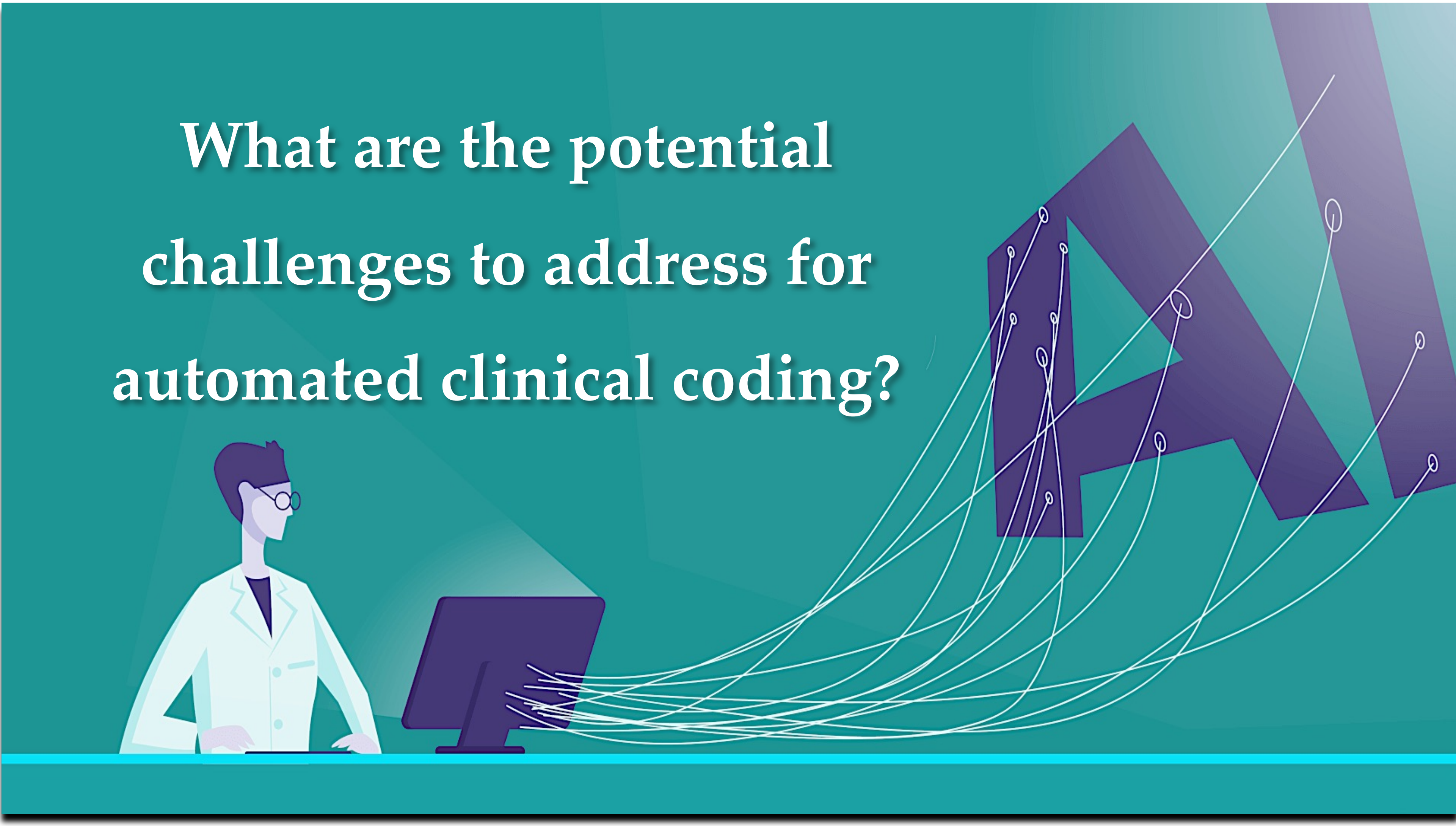
III. The social-technical issues

- From the perspective of information systems, transitioning to a (semi-) automated coding environment in a national healthcare system is more challenging than the technical issues themselves.





npj Digital Medicine. 2022;5(1):159.


What are the potential challenges to address for automated clinical coding?
AI

What are the challenges to address for automated clinical coding?

I. Creating gold standard coding data sets

- The current widely used benchmark: **MIMIC-III may have been significantly under-coded**.
- There is a **lack of large, openly available, and expert-labelled data sets from EHRs**, and models trained on MIMIC-III may not simply generalize to other data sets due to the difference in length, style, and language (e.g., China, Spain, or UK).
- **Various expert-labelled coding data sets are also needed for different purposes** of using clinical codes. For example, for epidemiology studies to identify deep phenotypes (potentially link to nuanced terminologies like SNOMED CT).
- Ensuring accurate and publicly available data sets from more healthcare systems for various purposes will better support the clinical NLP community.

MIMIC-III (Medical Information Mart for Intensive Care)

- is a large, single-center database comprising information relating to **patients admitted to critical care units** at a large tertiary care hospital. Data includes vital signs, medications, laboratory measurements, observations and notes charted by care providers, fluid balance, procedure codes, diagnostic codes, imaging reports, hospital length of stay, survival data, and more.
- contains 52,728 discharge summaries and 8,922 labels (ICD-10) in the dataset.

EHRs, Electronic health records

npj Digital Medicine. 2022;5(1):159.

What are the challenges to address for automated clinical coding?

II. Coding from heterogeneous, incomplete, and noisy sources (cont.)

- Clinical coding should be based on **all the relevant documents of a patient**, rather than just discharge summaries as in the majority of recent studies.
- This brings the challenges of long documents as discussed previously. **Structured data** (e.g., laboratory results) can also be included as a source for coding, also and **unstructured data** (e.g., radiographs) can be useful for coding as well.
- Besides, real-world data for clinical coders are usually **incomplete and noisy**, even for the same type of document (e.g., discharge summary), there is no guarantee that the document is available for all cases and presented in a unified format (i.e., can be hand-written or typed, with various levels of completeness).

What are the challenges to address for automated clinical coding? (cont.)

III. Explainability of clinical coding

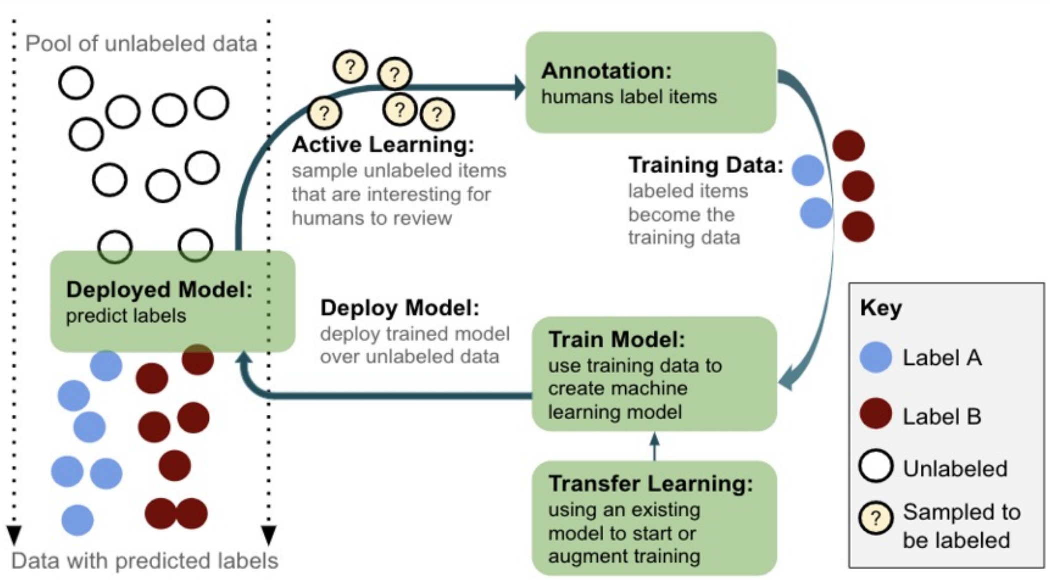
- Coders need to understand how the decisions are made by the system.
- The challenge is more related to the deep learning based multi-label classification approach. Work in this area so far uses **label-wise attention mechanisms to highlight key n-grams, words, and sentences**.
- However, **the highlighted texts mostly indicate associations instead of causality**. Further studies are needed to evaluate the usefulness of highlights for clinical coders and also to integrate more inherently explainable methods, for example, integrating symbolic representations of the coding steps with deep learning.

npj Digital Medicine. 2022;5(1):159.

What are the challenges to address for automated clinical coding? (cont.)

IV. Human-in-the-loop learning with coders' feedback

- To better deploy an automated coding tool into practice, it is essential to involve coders' feedback in the system.
- The feedback may take different forms, for example, manual corrections, highlights, and rules. The feedback may need to be incorporated into a deep learning system for coding.
- A relevant area to **human-in-the-loop learning is active learning**, with is about selecting the minimum set of most important data for humans to provide annotation feedback.



What are the challenges to address for automated clinical coding? (cont.)

V. Few-shot and zero-shot learning

- Many codes have a **low frequency or even no occurrence (or unseen) in the training data**, this is a key problem for multi-label classification with many labels (e.g., 68,000 codes in ICD-10).
- The best systems so far to work with low-frequent (<5 times) codes on the MIMIC-III data set are still below or around 40% recall.
- Better support for few-shot and zero-shot learning will improve the overall coding performance and usage. Knowledge (e.g., descriptions, properties, relations from multiple linked sources, and coding rules) can bridge the gap between the seen and unseen codes, as reviewed in the general domain.

npj Digital Medicine. 2022;5(1):159.

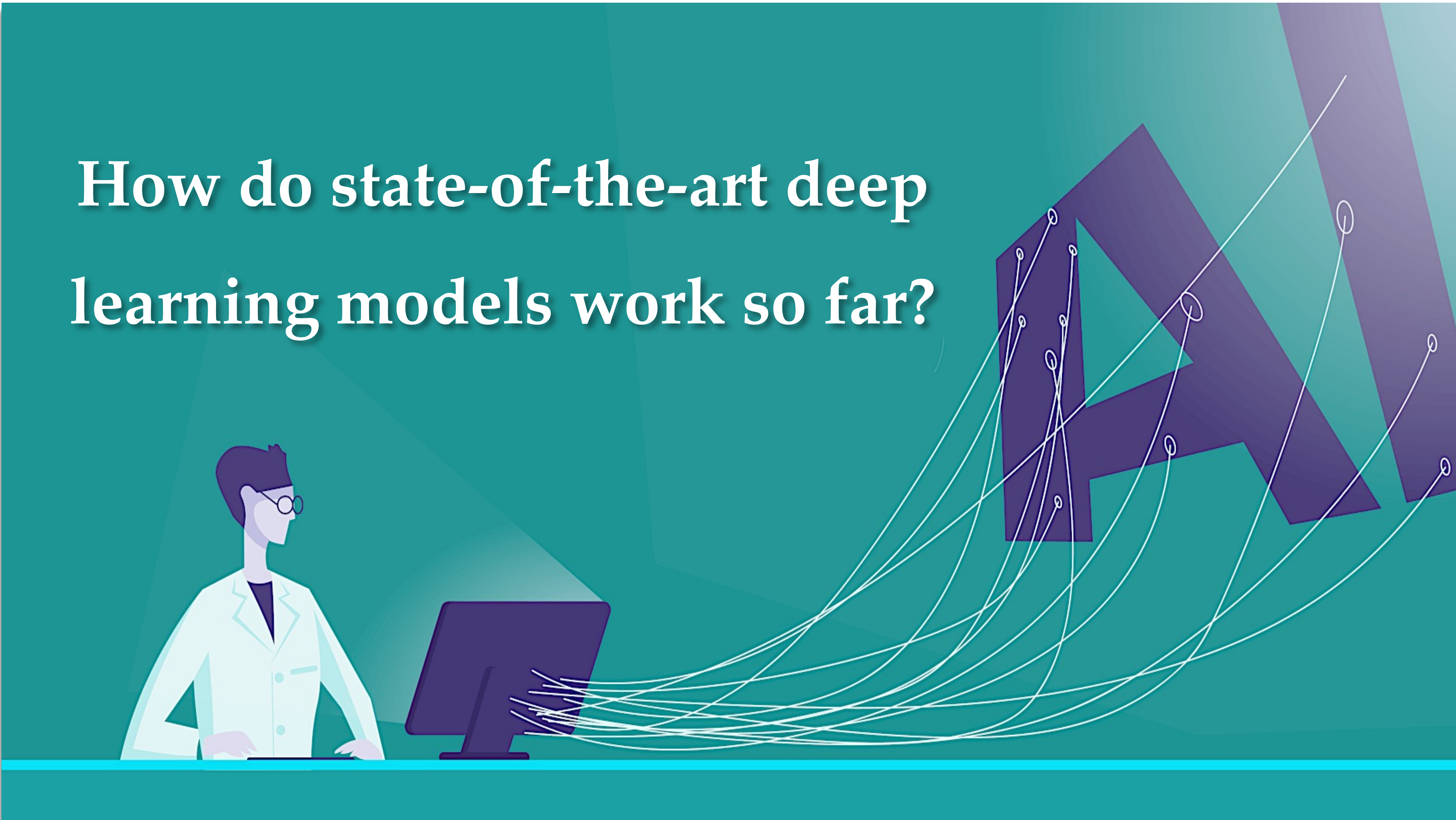
What are the challenges to address for automated clinical coding? (cont.)

VI. Adaptation to terminology changes

- How a trained model can be **adapted to modified standards** for coding or a completely new ontology (e.g., from ICD-10 to ICD- 11)?
- As described earlier, ICD-11 is semantically more complex than ICD-10 with a polyhierarchical backbone structure and the post-coordination of codes.
- The transition of terminologies may require novel paradigms in deep learning (e.g., self-supervised learning, transfer learning, and meta-learning), accurate ontology matching, concept drift handling, and few-shot and zero-shot learning for **new codes with no or few training data.**

npj Digital Medicine. 2022;5(1):159.

How do state-of-the-art deep learning models work so far?



How do state-of-the-art DL models work so far?

Researches related to automated clinical coding

Year	Study	ML classifiers & DL models
2013	Perotte et al. [65]	Flat SVM, Hierarchy-based SVM
2014	Marafino et al. [55]	SVM
2014	Subotin and Davis [86]	Two-level hierarchical classification
2015	Kavuluru et al. [43]	SVM, LR, MNB
2016	Ayyar and Oliver [4]	LSTM
2017	Prakash et al. [67]	C-MemNN and A-MemNN
2017	Lin et al. [52]	CNN
2017	Berndorfer and Henriksson [8]	Flat SVM and Hierarchical SVM
2018	Amoia et al. [3]	LR and CNN
2018	Catling et al. [13]	RNN-GRU
2018	Baumel et al. [5]	SVM, CBOW, CNN, HA-GRU
2018	Mullenbach et al. [61]	CAML, DR-CAML
2018	Samonte et al. [74]	EnHAN
2018	Xie and Xing [93]	Tree-of-sequences LSTM
2018	Rios and Kavuluru [71]	ZACNN, ZAGCNN
2018	Kaur and Ginige [39]	SVM, NB, Decision Tree, kNN, RF, AdaBoost, and MLP
2019	Zeng et al. [97]	Deep transfer learning using multi-scale CNN and Batch normalisation
2019	Kaur and Ginige [40]	Binary relevance, Label Power set, ML-kNN
2019	Xie et al. [94]	MSATT-KG
2019	Falis et al. [19]	Multi-view CNN (Ontological attention ensemble mechanism)
2019	Huang et al. [36]	CNN, LSTM, GRU
2019	Li et al. [51]	DeepLabeler (CNN and D2V)
2019	Rios and Kavuluru [72]	CNNs
2019	Schäfer and Friedrich [78]	FastText
2019	Xu et al. [95]	Text-CNN
2019	Du et al. [18]	ML-Net, ML-CNN, ML-HAN
2020	Cao et al. [12]	HyperCore

Year	Study	ML classifiers & DL models
2020	Guo et al. [28]	BiLSTMs
2020	Vu et al. [90]	LAAT and JointLAAT
2020	Song et al. [84]	ZAGRNN, ZAGRNN with LDAM loss
2020	Sonabend W et al. [83]	UNITE
2020	Mascio et al. [56]	ANN, CNN, Bi-LSTM
2020	Li and Yu [50]	MultiResCNN
2020	Teng et al. [88]	G_Coder (Multi-CNN, Graph Presentation, Attention Matching, Adversarial Learning)
2020	Moons et al. [58]	CNN, Bi-GRU, DR-CAML, MVC-LDA, MVC-RLDA
2020	Ji et al. [37]	DCAN
2020	Chang and Chang [14]	CNN, LSTM, GRU, HAN
2020	Zhang et al. [99]	AttentionXML (BERT-XML)

Kaur R, et al. 2021

How do state-of-the-art DL models work so far? (cont.)

Several major limitations when DL applied to clinical coding

I. Handling unseen, infrequent, and imbalanced labels

- In the MIMIC-III data set, **around 5000 codes appear fewer than 10 times in the training data and over 50% of codes never appear**.
- Vanilla deep learning models rely on large amounts of data for training and fail completely for new or unseen labels.
- Multi-label classification is also very challenging, especially when there are many labels or when the labels are imbalanced.

II. Handling long documents

- The recent Transformer-based pre-trained language models (e.g., BERT) usually require a **limited length of up to 512 sub-word tokens** (where a word can be tokenized into several sub- words) as input due to the memory-demanding self-attention mechanism, while discharge summaries alone in MIMIC-III have on average around 1500 tokens or words and up to over 10,000 tokens, not counting other types of clinical notes.
- More recent studies applied Longformer, TransformerXL, BigBird to clinical coding to **process documents of up to 4,096 tokens**, but this is still insufficient for the clinical notes.

BERT, Bidirectional Encoding Representations from Transformers

npj Digital Medicine. 2022;5(1):159.

How do state-of-the-art DL models work so far? (cont.)

Several major limitations when DL applied to clinical coding

III. Domain mismatch

- Large Language Models (LLM) are typically pre-trained with billions of tokens from a wide variety of general-domain texts, such as Wikipedia, novels, web pages, and forums.
- This broad spectrum often leads to a domain mismatch when these models are applied to specialized fields, which has been shown to undermine performance on downstream tasks.
- To tackle domain-specific problems, prior work adapted such **models to scientific and biomedical domains**, including BioBERT, ClinicalBERT, PubMed-BERT, and RoBERTa-PM.

BERT, Bidirectional Encoding Representations from Transformers

npj Digital Medicine. 2022;5(1):159.

How do state-of-the-art DL models work so far?
(cont.)

Several major limitations when DL applied to clinical coding

IV. Lack of symbolic reasoning capabilities

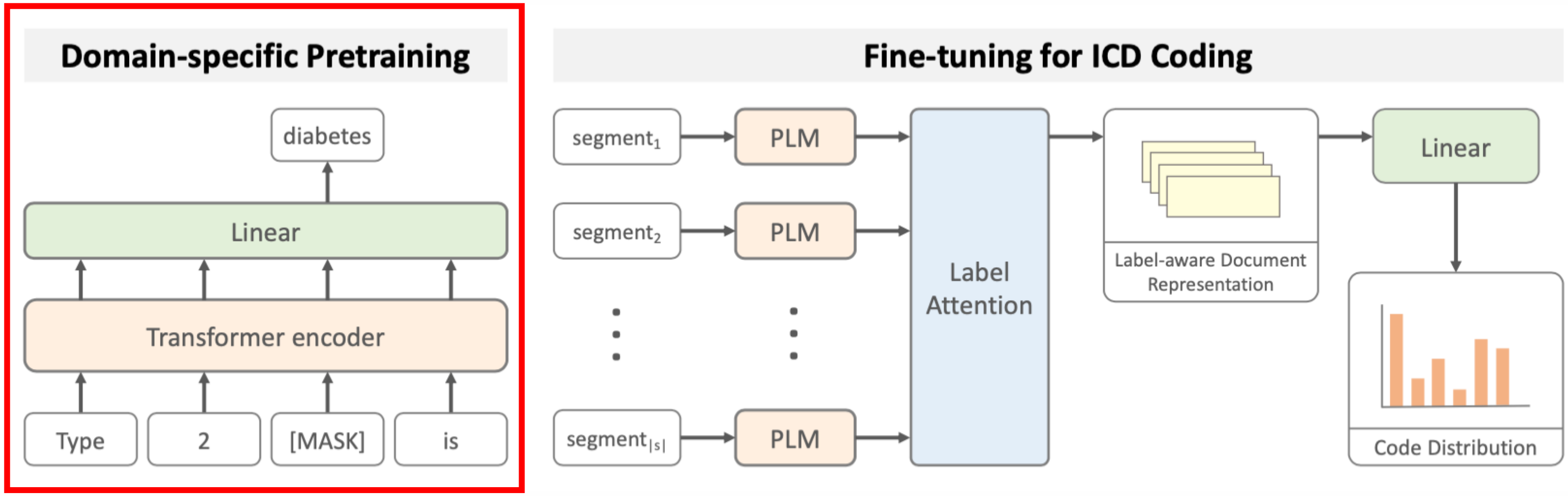
- ❖ Manual coding involves reasoning beyond just locating concepts in the notes. The coders sometimes need to connect different pieces of information together.
- ❖ The information from different sources may even be contradictory to each other for the same patient. Their decisions are based on a standard coding process, aided by coding guidelines.
- ❖ Deep learning, on the other hand, tries to simply **learn from the labelled data the association between texts and codes in different (pre-trained) embedding spaces**, without explicitly modelling the reasoning process.
- ❖ Human-like reasoning may be supported by **knowledge-based techniques**, which can potentially boost the performance and explainability of coding of deep learning methods.
- ❖ The reasoning may include formalizing coding guidelines into logical expressions and creating regular expressions to capture various diagnosis descriptions of a code and leveraging various semantics in knowledge graphs constructed from several linked ontologies including the target ICD hierarchy.

npj Digital Medicine. 2022;5(1):159.

How do state-of-the-art DL models work so far?
(cont.)

The state-of-the-art model for the task of ICD coding

The Pretrained-Language-Model framework (PLM-ICD)



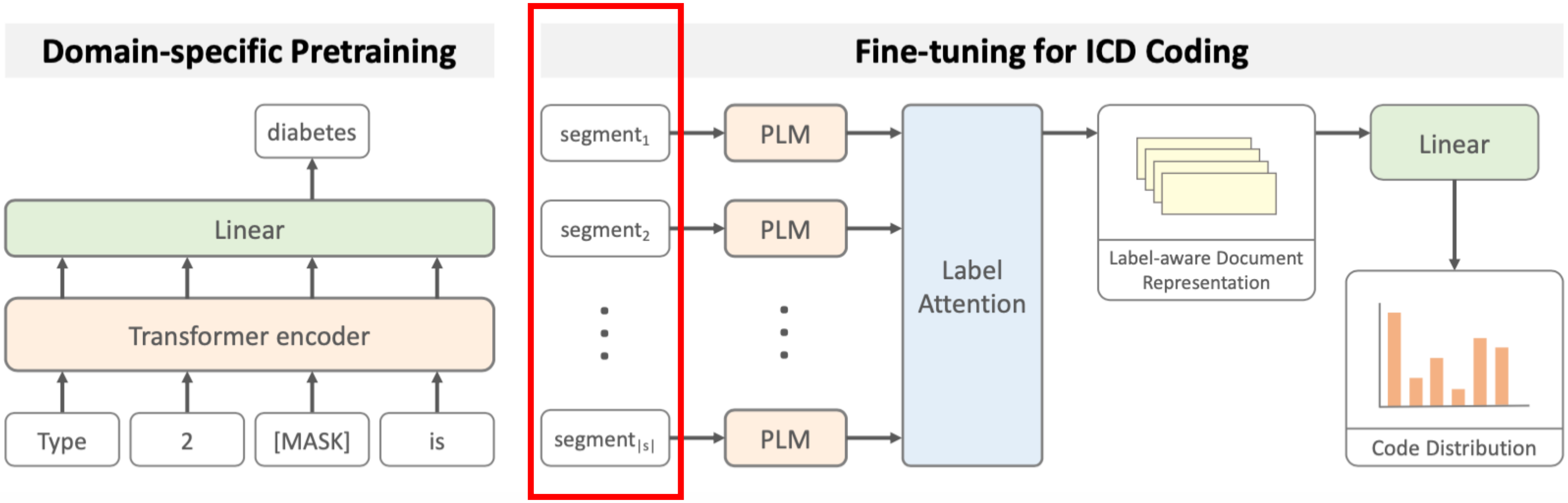
❖ **Domain-Specific Pre-training**: employing pre-trained language models on biomedical and clinical texts, like BioBERT, PubMedBERT, and RoBERTa-PM. These specialized models have a better grasp of the nuanced biomedical terminology prevalent in clinical notes, crucial for accurate ICD code assignment. By fine-tuning these domain-specific PLMs on ICD coding tasks, we leverage even more their inherent design and objectives, similar to general-domain models.

Huang C-W, et al. 2022

How do state-of-the-art DL models work so far? (cont.)

The state-of-the-art model for the task of ICD coding

The Pretrained-Language-Model framework (PLM-ICD)



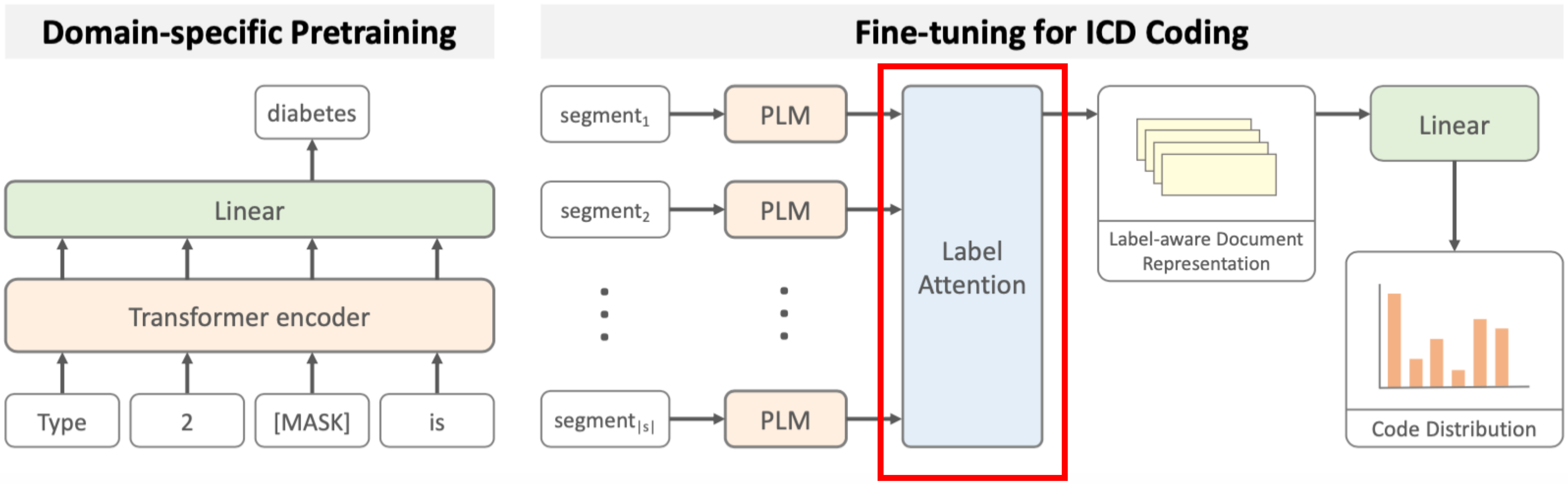
- **Segment Pooling**: dividing the lengthy clinical documents into smaller segments, each fitting within the maximum token limit of the PLMs. Each segment is independently encoded to generate representations, which are then concatenated to form a comprehensive representation of the entire document (H). By breaking down and then aggregating these segment representations, the model can effectively process and predict based on the full extent of the clinical notes, bypassing the token length constraints of standard PLMs.

Huang C-W, et al. 2022

How do state-of-the-art DL models work so far? (cont.)

The state-of-the-art model for the task of ICD coding

The Pretrained-Language-Model framework (PLM-ICD)



❖ **Label-Aware Attention**: designing a mechanism to enhance the translation task by focusing on parts of the text that are particularly relevant to specific labels. After processing the text to capture the context within hidden representations (H), label-aware attention selectively emphasizes the information from these representations that is most pertinent to each label. This way, even when faced with a large number of possible labels, the model can efficiently identify and focus on the text segments that contribute most meaningfully to the coding task, enabling more precise predictions.

How do state-of-the-art DL models work so far?
(cont.)

The state-of-the-art model for the task of ICD coding

The Pretrained-Language-Model framework (PLM-ICD)

❖ Results on the MIMIC-3 full test set (%).

Model	AUC		F1		P@k		
	Macro	Micro	Macro	Micro	P@5	P@8	P@15
CAML (2018)	89.5	98.6	8.8	53.9	-	70.9	56.1
DR-CAML (2018)	89.7	98.5	8.6	52.9	-	69.0	54.8
MultiResCNN (2020)	91.0	98.6	8.5	55.2	-	73.4	58.4
LAAT (2020)	91.9	98.8	9.9	57.5	81.3	73.8	59.1
JointLAAT (2020)	92.1	98.8	**10.7**	57.5	80.6	73.5	59.0
EffectiveCAN (2021)	91.5	98.8	10.6	58.9	-	75.8	60.6
PLM-ICD (Ours)	**92.6** (.2)	**98.9** (.1)	10.4 (.1)	***59.8***† (.3)	***84.4***† (.2)	***77.1***† (.2)	***61.3***† (.1)
Models with Special Code Description Modeling							
HyperCore (2020)	93.0	98.9	9.0	55.1	-	72.2	57.9
ISD (2021)	93.8	99.0	11.9	55.9	-	74.5	-
RAC (2021)	*94.8*	*99.2*	*12.7*	58.6	82.9	75.4	60.1

Huang C-W, et al. 2022

How do state-of-the-art DL models work so far? (cont.)

The state-of-the-art model for the task of ICD coding

The Pretrained-Language-Model framework (PLM-ICD)

- Ablation results on the MIMIC-3 full test set (%).

	Model	Macro-F	Micro-F
	PLM-ICD	10.4	59.8
(a)	- domain pretraining	8.9	54.2
(b)	- segment pooling	7.2	54.6
(c)	- label attention	4.6	48.0

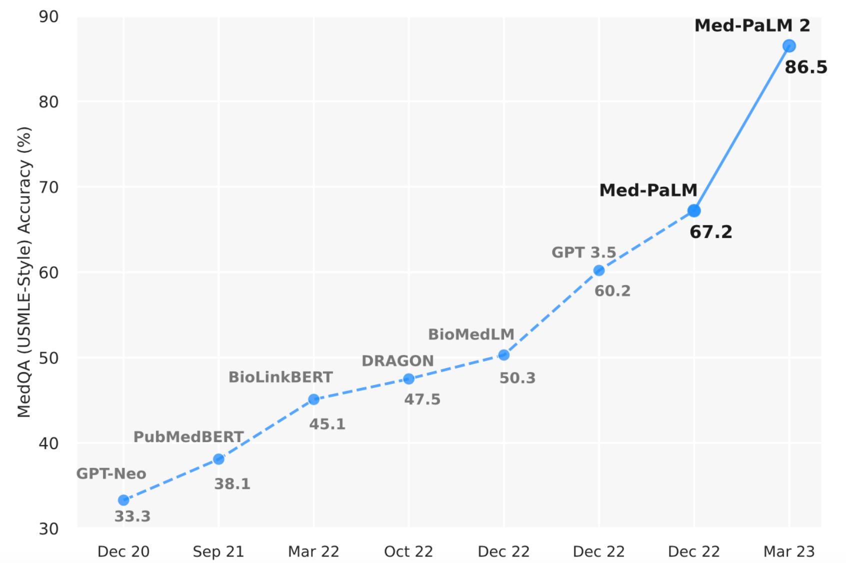
- **The label attention mechanism** is indeed the most pivotal component in bridging clinical text and medical codes, as it is the core translator in the PLM-ICD framework.
- This model marks a significant advancement in the field of automated medical coding, providing a sophisticated solution by leveraging pre-trained language models for this task. Combined with the ingenious translation component, the label-aware attention mechanism, it sets a new benchmark for precision and efficiency, as **demonstrated by its state-of-the-art performance.**

Huang C-W, et al. 2022

How do state-of-the-art DL models work so far? (cont.)

Google's New Chatbot Passed The US Medical Exam (But Only Just)

TECH 13 July 2023 By AFP



Google's Med-PALM 2 achieved 86.5% score on the MedQA dataset[1]

A 90% score achieved using GPT-4 combined with a set of **innovative prompt engineering techniques**, called **Medprompt**[2].

AI research Dec 4, 2023

Microsoft's Medprompt demonstrates the power of prompting

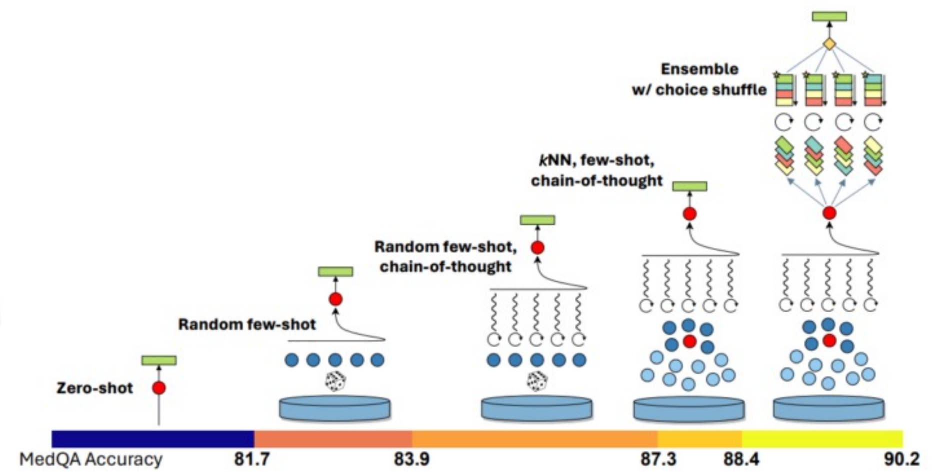
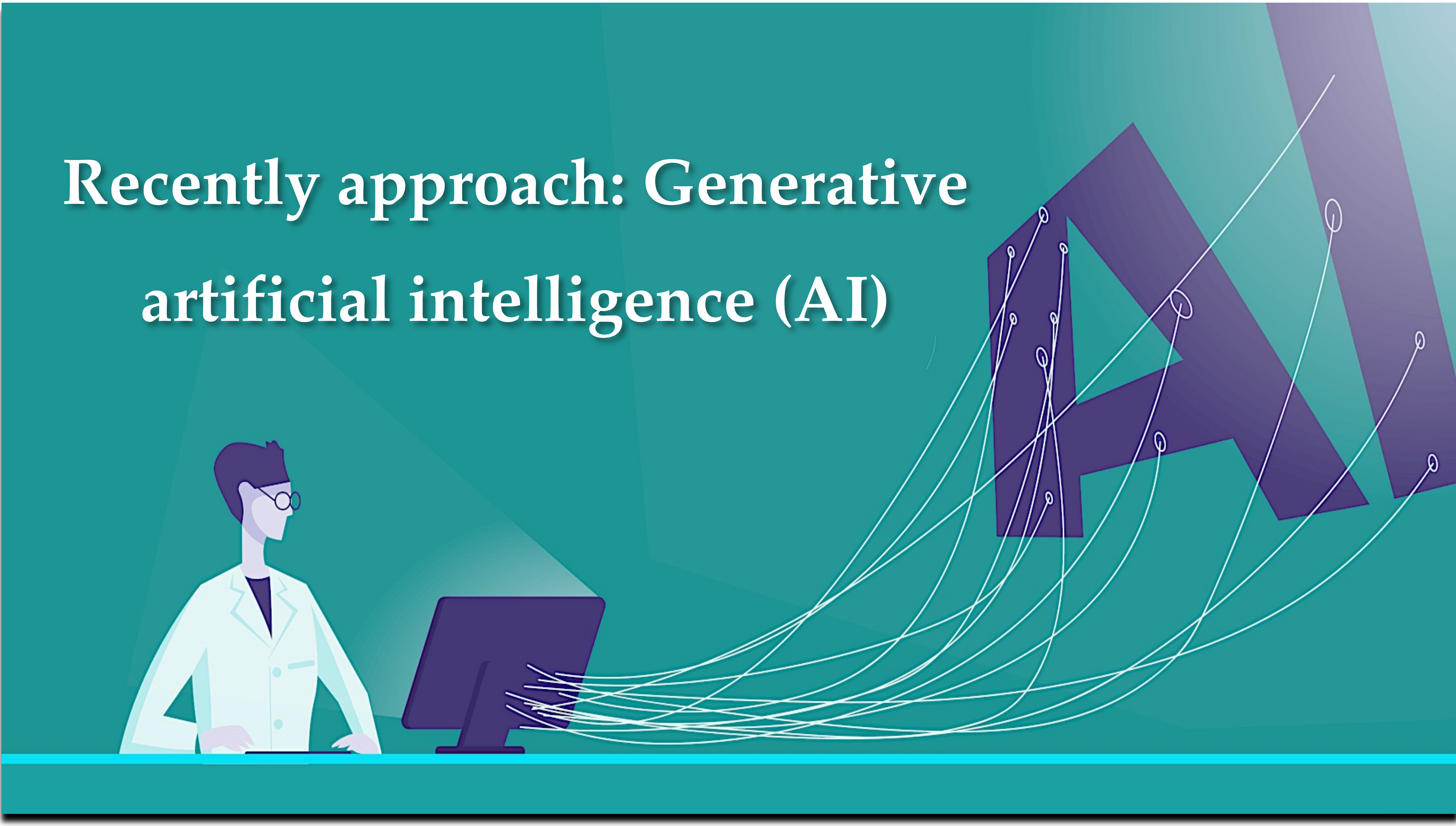


Figure 4: Visual illustration of Medprompt components and additive contributions to performance on the MedQA benchmark. The prompting strategy combines kNN-based few-shot example selection, GPT-4–generated chain-of-thought prompting, and answer-choice shuffled ensembling (see details in Section 4). Relative contributions of each component are shown at the bottom (details in Section 5.2).

1. Singhal K, et al. 2023; 2. Nori H, et al. 2023


Recently approach: Generative
artificial intelligence (AI)
AI

Recently approach: Generative AI

Automated clinical coding using off-the-shelf large language models

Joseph S. Boyle[1,2], Antanas Kascenas[1,3], Pat Lok[1,4], Maria Liakata[2,5,6], Alison Q. O'Neil[1,3]

[1]Canon Medical Research Europe, [2]Queen Mary University of London, [3]University of Edinburgh, [4]Anglia Ruskin University, [5]The Alan Turing Institute, [6]University of Warwick

OpenAI ChatGPT gpt-3.5-turbo

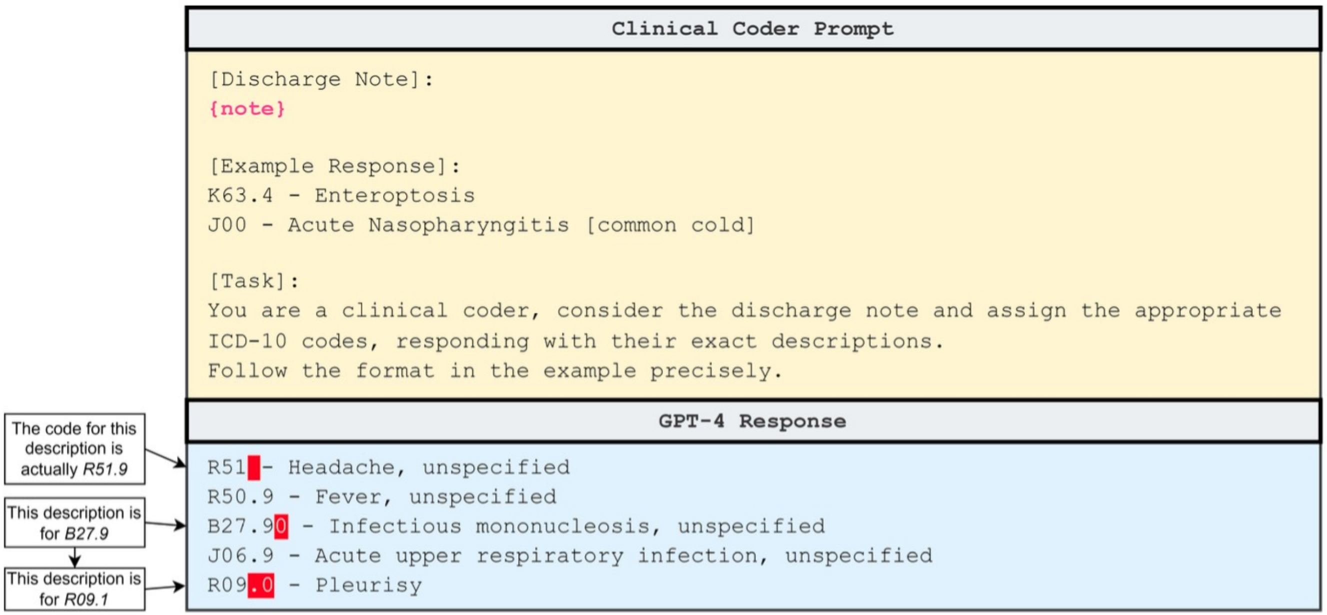
GPT-4

- ❖ Inspired by the **Power of Prompt Engineering**
- ❖ **First method for ICD coding requiring no task-specific training or fine-tuning**
- ❖ **Ability to overcome out-of-the-box ICD coding issue or rare codes**
- ❖ Propose a method that avoids reliance on model knowledge of the target coding ontology via the injection of information into the LLM prompt and the application of a novel search strategy conceptually similar to a multi-label decision tree.

Recently approach: Generative AI

Clinical coder prompt

❖ The prompt used for the out-of-the-box 'You are a clinical coder...' method.



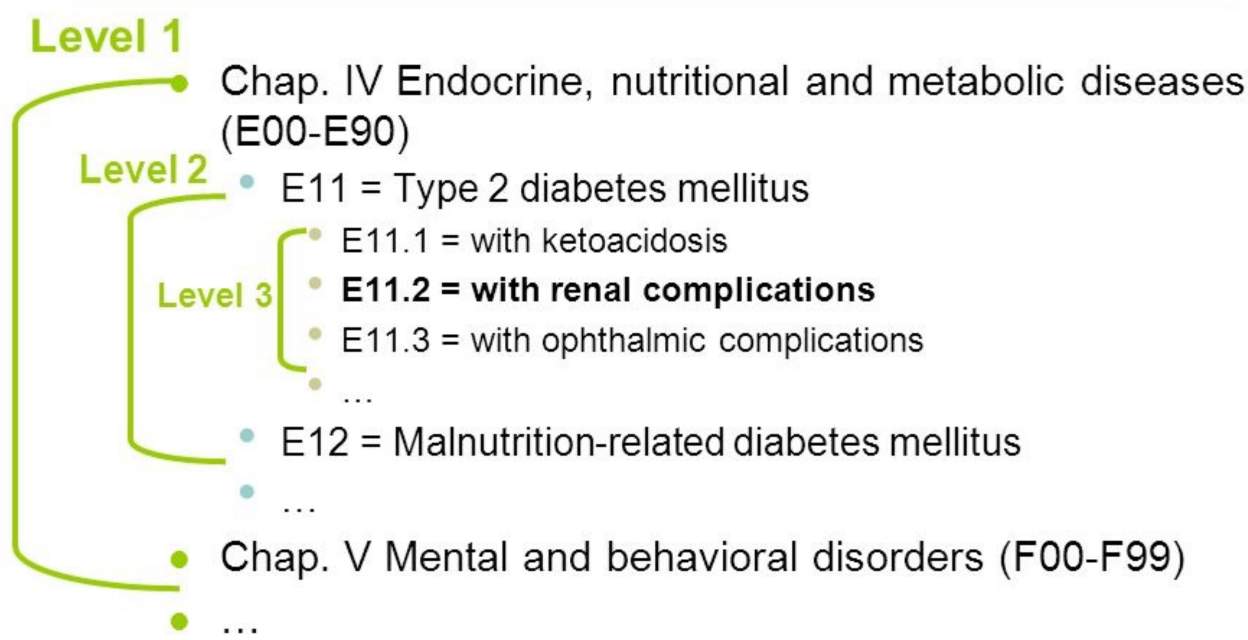
❖ A new approach: we do not assume that the LLM is an expert in medical codes. Instead, we consider that the model is very efficient in abstracting information and cross-referencing it with data it has access to. **Therefore, the problem is reframed as information retrieval.**

Recently approach: Generative AI
(cont.)

Recently approach: Generative AI

LMM Guided Tree-Search of Medical Codes Strategy

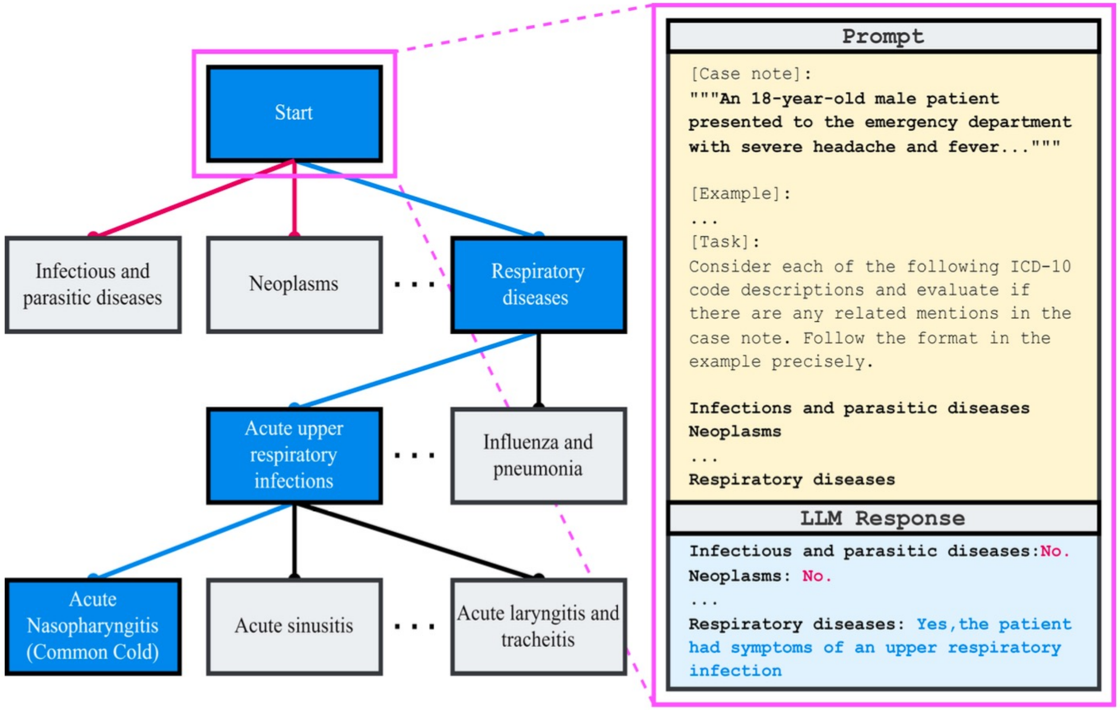
❖ The ICD ontology is structured as a hierarchical tree, where relationships between parent and child codes are defined by ‘is a’ semantics, indicating that one condition is a subtype or specific instance of another.



Joseph S, et al. 2023

Recently approach: Generative AI

LMM Guided Tree-Search of Medical Codes Strategy



- The process starts at the ontology's root, with the LLM guiding the exploration by selecting pertinent chapters or branches based on the input query.
- This selection is similar to navigating decision points, with the model evaluating which branches are relevant and should be pursued further.
- The procedure iterates recursively, traversing down the tree and refining the search scope at each level, based on the model's recommendations.

Recently approach: Generative AI
(cont.)

Recently approach: Generative AI

LMM Guided Tree-Search of Medical Codes Strategy

	Micro			Macro		
Model	**Rec.**	**Prec.**	**F1**	**Rec.**	**Prec.**	**F1**
PLM-ICD	0.213	0.225	**0.219**	0.244	0.237	0.216
Clinical coder (match codes):						
Llama-2	0.011	0.033	0.016	0.007	0.011	0.006
GPT-3.5	0.163	0.155	0.159	0.149	0.161	0.136
GPT-4	0.242	0.161	0.193	0.219	0.214	0.195
Clinical coder (match descriptions):						
Llama-2	0.037	**0.282**	0.065	0.034	0.061	0.040
GPT-3.5	0.147	0.168	0.157	0.135	0.155	0.128
GPT-4	0.217	0.166	0.188	0.187	0.190	0.169
Tree-search:						
Llama-2	0.173	0.039	0.064	0.197	0.152	0.144
GPT-3.5	0.206	0.159	0.179	0.220	**0.241**	0.208
GPT-4	**0.331**	0.087	0.138	**0.381**	0.190	**0.225**

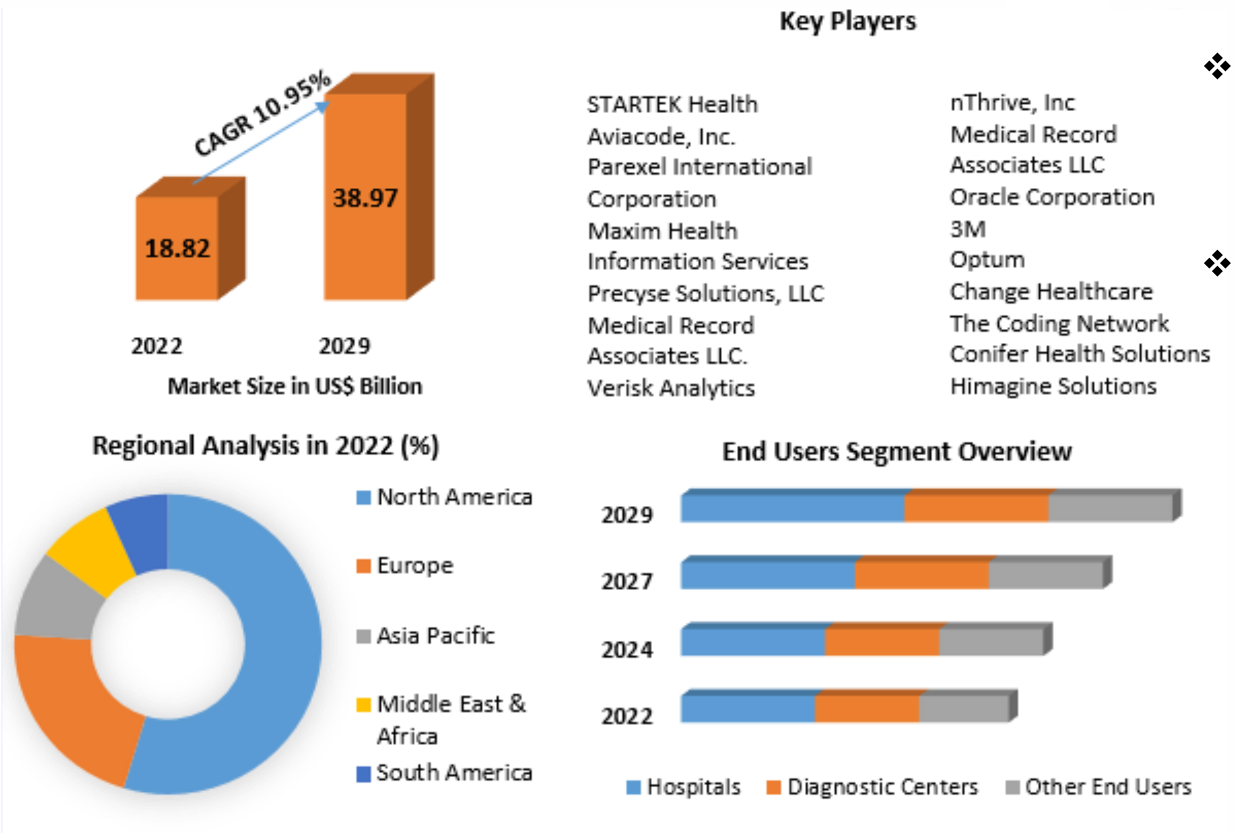
*Performance on the CodiEsp English dataset

Recently approach: Generative AI

LMM Guided Tree-Search of Medical Codes Strategy

- The major drawbacks of this strategy is the high number of queries needed to explore the ICD ontology. This translates into a **high computational cost** for inferring medical codes for each clinical note.
- **No consensus regarding best practices for prompting** in different foundational models.
- **The correct code prediction depends on the path taken in the ICD tree**: if the model fails to identify a high-hierarchy category, it will never have the opportunity to classify the correct code.

Medical Coding Market



- The Epic EHR system is deployed in the University College London Hospital (UCLH) for the management of EHRs.
- Industry NER + L APIs (e.g., **Amazon Comprehend Medical InferICD10CM**, **Microsoft Text Analytics for health and Google Healthcare Natural Language API**) have been released during the last two to three years to support clinical concept extraction from texts with price charges.

1. npj Digital Medicine. 2022;5(1):159; https://www.maximizemarketresearch.com/market-report/global-medical-coding-market/91813/

Thank You