npj | digital medicine

www.nature.com/npjdigitalmed

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

## INTRODUCTION: 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 (International Classification of Diseases, Tenth Revision). For example, in Scotland, this means to apply a standard process to classify information about patients into appropriate diagnosis and procedure codes in ICD and OPCS (OPCS Classification of Interventions and Procedures), finally contributing to the Scottish Morbidity Records (SMR01) national data set[1]. The purpose of clinical coding is to provide consistent and comparable clinical information across units of care and over time. The resulting national data are used to support areas, such as health improvement, inform healthcare planning and policy and add to the epidemiological understanding of a wide variety of conditions, so confidence in the data is essential. Also, codes are mainly used for billing purposes in the US[2]. For introductory slides about clinical coding in the UK provided by NHS Digital, see *Clinical coding for non coders*[3].

Clinical coding is a non-trivial task for humans. The process of coding usually includes data abstraction or summarisation[4]. More specifically, an expert clinical coder is expected to decipher a large number of documents about a patient's episode of care, and to select the most accurate codes from a large classification system (or an ontology), according to the contexts in the various documents and the regularly updated coding guidelines. For example, coding in the US adopts the International Classification of Diseases, Tenth Revision, Clinical Modification (ICD-10-CM), which has around 68,000 diagnosis codes[5]; ICD-10 is also the main classification for coding in the UK. There is a standard process for manual coding to ensure data consistency: textual analysis, summarisation, and clearly defined steps to classification into codes (or the four steps of *analyse, locate, assign, and verify* as suggested by the NHS digital in the coding standard of 2021 [6, p.11]). The process minimises the risk of introducing variations caused by artefacts (potentially leading to wrong decision making), thus collecting and analysing data and applying the standard is important. There are regularly updated guidelines and standards for coding (e.g., in Public Health Scotland[6]). Usually, it can take months or longer to train an expert clinical coder in the NHS (National Health Service) in the UK, and worldwide [7].

Automated clinical coding is the idea that clinical coding may be automated by computers using AI techniques, e.g., NLP and machine learning[8]. It is a branch of computer-assisted coding (CAC)[9]. In recent years, AI has been considered a promising approach to transforming healthcare by intelligently processing the increasing amount of data with machine learning and NLP techniques[10]. Automated clinical coding is a potential AI application to facilitate the administration and management of clinical records in the hospital and medical research. There has been a surge of articles for automated clinical coding with deep learning (as the current mainstream approach of AI) in the last few years, as reviewed in recent surveys[11–13].

However, while there is some progress for automated clinical coding, the task is far from solved. For the last two years and more, we have been working on the task and discussing it with practitioners of clinical coding and clinicians from Scotland and the UK. We illustrate the manual and automated clinical coding process, and their potential interactions, in Fig. 1. In this paper, we aim to summarise the technical challenges of clinical coding, mainly related to deep learning, and propose directions for future research in this area.

## WHY DO WE NEED AUTOMATED CLINICAL CODING?

There are some major reasons that automated clinical coding can be helpful. First, manual coding is time-consuming. A clinical

[1]Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK. [2]Department of Computer Science, University of Oxford, Oxford, UK. [3]School of Informatics, University of Edinburgh, Edinburgh, UK. [4]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK. [5]Edinburgh Futures Institute, University of Edinburgh, Edinburgh, UK. [6]Epic Systems Corporation, Verona, WI, USA. [7]University College London Hospitals NHS Foundation Trust, Clinical Research Informatics Unit, London, UK. [8]Department of Computer Science, Aalto University, Espoo, Finland. [9]Institute of Health Informatics, University College London, London, UK. ✉email: hang.dong@cs.ox.ac.uk; honghan.wu@ucl.ac.uk

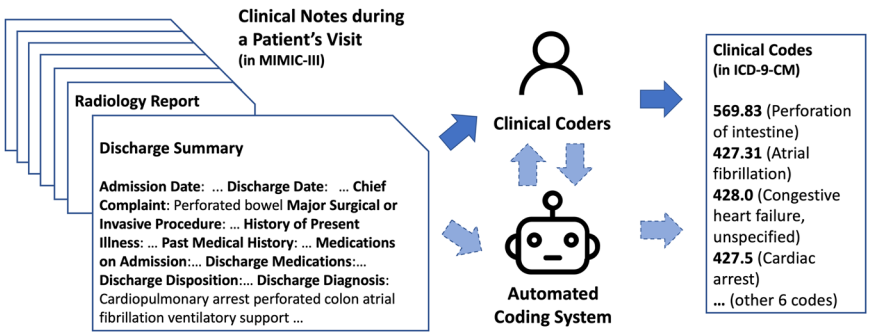


**Fig. 1 An example of clinical coding, manual and automated (linked with solid and dashed arrows, respectively), with ICD-9-CM codes from a clinical note in the MIMIC-III data set[20] of ICU patients in 2001–2012 in a hospital in the US.** Dashed arrows between clinical coders and the automated coding system suggest potential interactions between them, while this is yet to be considered in many clinical coding systems. Note that the format of data and clinical codes does not reflect the situation of other regions in the world—for example, in the UK, where data may be less structured and there is no universal discharge summary format available.

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. Even so, there is usually a backlog of cases to be coded, which can take several months or more (e.g., over a year[14]). Second, manual coding may be prone to errors. This may be due to incompleteness in a patient's data, subjectivity in choosing the diagnosis codes, lack of coding expertise, or data entry errors[4]. The average accuracy of coding in the UK was around 83% with a large variance among studies (50-98%)[15]. In Scotland, the accuracy of coding is very high[16] (e.g., in the assessment during 2019–2020, achieved 92.5% for 3-digit code accuracy and 88.8% for 4-digit code accuracy of main conditions), yet still not perfect and under-coding occurs (for around 20% of the common conditions). On the other hand, computer-assisted coding could improve the accuracy, quality, and efficiency of manual coding, according to a recent, qualitative literature review[9]. We believe that with recent AI technologies (e.g., NLP), automated coding has the potential to better support clinical coders. We mostly focus on the case that AI directly contributes to assigning clinical codes.

## WHY IS AUTOMATED CODING A COMPLEX PROBLEM TO SOLVE?

While humans can achieve high accuracy in clinical coding, the standard procedure, text analysis, text summarisation, and classification into codes, poses immense challenges for computer-based systems. This requires Natural Language Understanding (NLU), one of the classical but largely unsolved areas of AI[17,18], and the linking of natural language to knowledge representations like the ICD-10 classification system. Also, this clinical task poses more specific challenges compared to common NLU tasks. From our experience, these relate mainly to the following difficulties:

(1) 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. They have variable document structures, they can be lengthy (on average around 1500 words[19] in only the discharge summaries in a US intensive care data set, MIMIC-III[20]), and use terse abbreviations and symbols[8,21] (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 in MIMIC-III). Coding also 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.

(2) Classification systems used for coding are complex and dynamic. The ICD-10-CM system (implemented in the US in 2015) has around 68,000 diagnosis codes in a large hierarchy, 5 times more than the previous ICD-9-CM (used in MIMIC-III)[5]. The ICD-11 system[22] (or ICD-11-MMS, ICD-11 for Mortality and Morbidity Statistics, came into effect in early 2022, but is yet to be used in practice in the US or the UK at the time of writing) "contains around 17,000 unique codes for injuries, diseases and causes of death, underpinned by more than 120,000 codable terms" and can code "more than 1.6 million clinical situations" using code combinations[23]. ICD-11 also introduces significant changes in chapter structure, diagnostic categories, diagnostic criteria, etc., for example, in psychiatric classification[24]. ICD-11-MMS has a similar structure as in ICD-10 with more chapters, but distinct from previous versions, ICD-11-MMS has its backbone as a semantic network ("Foundation Component"), a large and deep polyhierarchy (i.e. children can have more than one parents) of medical concepts, where ICD-11-MMS is derived from; coding with ICD-11-MMS also allows "post-coordination" that uses code combinations to express complex phenotypes of a patient[25] and more details and examples are in the ICD-11 reference guide[26]. Besides, to support the localisation of ICD systems, classification standards are updated regularly (e.g., usually every few months in Public Health Scotland[6]). Automated clinical coding needs to work with dynamic and complex classification systems.

(3) The social-technical issues with automated clinical coding systems are still to be explored. 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. How do coders interact with an AI-based CAC system (as modelled in Fig. 1)? How to present the information in an automated coding system so that coders will easily ignore errors and make the most use of the correct automatic codes? Will coders trust such a system? How will the role of coders change (e.g., from coders to coding editors or coding analysts)? What new skills will coders need [9]?

## HOW TO SOLVE AUTOMATED CLINICAL CODING: SYMBOLIC OR NEURAL AI?

The two main schools of thought of AI have been either a *symbolic, knowledge-based* approach or a *neural network* (which

further developed into deep learning) based approach[17]. Putting them into the task of clinical coding, the symbolic AI approach aims at making the use of symbols and rules to represent and model the standard practice that clinical coders apply in their work. The neural network and deep learning approach aims at learning a complex function to match a patient's information to the appropriate set of medical codes. This function is learned from the training data. From the historical perspective, symbolic AI, as the mainstream approach from 1950 to the early 1980s, did not scale up to complex real-world scenarios, for example, to model the natural language that people use in their daily life[17,18]. Neural networks returned in the mid-1980s with *machine learning* in general. *Deep learning* methods became the mainstream of AI after 2011[18], continuing to evolve today [27].

Coming back to automated clinical coding, while the task has been studied for around 50 years (with the earliest studies around 1970[28]), the current deep learning-based methods have a short history. Prior to deep learning, most studies use rules (regular expressions, logic expressions, and keywords) with feature engineering methods for text classification[8,13]. The issue with pure rule-based methods is that it is not straightforward and it can be time-consuming to extend rules to tens of thousands of codes and their varieties, and inter-relations among codes; this thus needs the support of machine learning with textual features for classification, and historically, some of the classifiers were Decision Trees, Support Vector Machine (SVM), etc[8,13,29]. Still, rule-based methods like using regular expressions to match various textual descriptions can result in high precision in coding (yet low recall), and have been used to support human coding to largely improve coding efficiency[30].

Only since around 2017[31,32], deep learning has been applied to automated coding and there are abundant studies in this area (reflected in recent surveys[11–13] and curation of papers in automated medical coding[33]). Unlike rule-based and traditional machine learning methods, pure deep learning methods do not require expert rules and hand-crafted textual features, thus easily applicable, while achieving better overall performance by learning from a sufficient amount of data[32]. Most of the studies formulate the task as a multi-label classification problem[34], while some studies formulate the task as a concept extraction or a Named Entity Recognition and Linking (NER + L) problem[35,36]. Though it seems that deep learning is the main method applied to automated clinical coding, we argue that there is still an important need for knowledge-based approaches in this area, and a better solution is to combine both schools of thought in the design of an automated clinical coding system. A recent trend is *knowledge-augmented deep learning* methods, where several studies used various *embedding*-based approaches to incorporate knowledge graphs into deep learning (to name a few[37–40]) or directly integrated the subsumption relations of codes into the model[41] and the evaluation[42]; however, the knowledge used is usually limited to the definition and hierarchies in the target ontology ICD-9 (except Freebase in Teng et al.[38]), while the other vast number of clinical ontologies (e.g., UMLS, SNOMED-CT, and others) are not leveraged with the multi-label classification approach; also other information in the ontologies like axioms, logical expressions, and class attributes have not been leveraged. Coding standard and guidelines have also not been leveraged to enhance deep learning, where a challenge would be the need to extract and represent the knowledge from them, which varies by locations and requires input from coding experts.

## HOW DO STATE-OF-THE-ART DEEP LEARNING MODELS WORK SO FAR?

*Coding tasks involving complex reasoning, such as those in which disparate pieces of information must be connected, are a difficult challenge for current NLP systems*.—Kukafka et al.[43], and also quoted in Stanfill et al.[8].

Clinical coding is a complex testbed for contemporary AI, especially for machine learning and deep learning applied to NLP. During the last few years, the problem itself elicits applied and theoretical research on text representation learning[19,44], multi-task learning[41,45], zero-shot learning[37,46], meta-learning[47], multi-modal learning[48], etc. The pursuit of a full-fledged deep learning-based clinical coding system, however, is far from being achieved: at the time of writing, the best Micro-F1 score (a harmonic mean of precision and recall evaluated based on pairs of a patient's information and a code) on the full 8932 ICD-9 codes for the MIMIC-III data was under 60% (between 58–60%)[45,49–52]. MIMIC-III discharge summaries[20], although coded with the older and obsolete version of ICD (ICD-9-CM, the ninth version, Clinical Modification), are the main data sets used for benchmarking[19]. This data set is also now older (collected over 10 years ago, from 2001 to 2012), and only represents an intensive care data set in the US, thus not representative of the documents available in the UK or other regions.

The main principle of the current deep learning approach is to find a complex function (non-linear and constructed by multiple layers) to match a clinical note of a patient's visit to a set of codes. As we introduced earlier, this is the multi-label classification setting. This approach, however, has several major limitations when applied to clinical coding:

(1) 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[37]. 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.

(2) 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[8,43]. 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[53]. 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 formalising coding guidelines into logical expressions[29] and creating regular expressions to capture various diagnosis descriptions of a code[30], and leveraging various semantics in knowledge graphs constructed from several linked ontologies including the target ICD hierarchy.

(3) Handling long documents: Looking for the relevant information of a code from a long document poses a "needle-in-the-haystack" problem. The recent Transformer-based pre-trained language models (e.g., BERT, Bi-directional Encoding Representations from Transformers[54]) usually require a limited length of up to 512 sub-word tokens (where a word can be tokenised 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[19] and up to over 10,000 tokens, not counting other types of clinical notes. More recent studies applied Longformer[55], TransformerXL[56], BigBird[57] to clinical coding to process documents of up to 4,096 tokens, but this is still insufficient

for the clinical notes. On the other hand, text redundancy (or "Note Bloat" problem[58]) is prevalent in clinical note creation, as measured in recent studies[58,59]. This may impede the performance of deep learning models for code prediction, which may be alleviated through text de-duplication based on text similarity measures[58].

## WHAT ARE THE POTENTIAL CHALLENGES TO ADDRESS FOR AUTOMATED CLINICAL CODING?

An empirical fact is that the current BERT-based approaches still do not achieve better performance than CNN-based methods for multi-label classification applied to clinical coding[44,60,61], except for the study[52]. The limitation of BERT may be due to its inefficiency in modelling concept-level information (usually represented in a few keywords or phrases instead of complex relations of tokens in the context) and long documents[60].

Besides, as we stated previously, manual coding is largely based on a standard and implied process with rules applied to the healthcare system, e.g., priority of certain codes, hypothetical mentions, code definitions, mutual exclusion, etc. Future deep learning-based systems need to integrate knowledge reasoning with rules and ontologies to achieve improved and more explainable results.

We list the technical challenges from our work in clinical coding and suggest relevant references below. Some of the challenges are also presented in a different way in a recent, concurrent review in Teng et al.[13]. The challenges of explainability and few- and zero-shot learning are more relevant to the multi-label classification approach but may be alleviated by the NER + L approach.

- Creating gold standard coding data sets—the current widely used benchmark data set MIMIC-III may have been significantly under-coded[62]. There is a lack of large, openly available, and expert-labelled data sets from Electronic Health Records in this area, and models trained on MIMIC-III may not simply generalise to other data sets due to the difference in length, style, and language (for example, clinical notes in China, Spain, or even the UK). Various expert-labelled coding data sets are also needed for different purposes of using clinical codes (for decision making, diagnosis, epidemiology, etc.), for example, for epidemiology studies to identify deep phenotypes (potentially link to nuanced terminologies like SNOMED CT) from multimodal and multi-source clinical data. Ensuring accurate and publicly available data sets from more healthcare systems for various purposes will better support the clinical NLP community.
- Coding from heterogeneous, incomplete, and noisy sources—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, as discussed in Alonso et al.[14]. This brings the challenges of long documents as discussed previously. *Structured data*, such as laboratory results, can also be included as a source for coding[48]. *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).
- 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[19], words, and sentences[61,63]. 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.
- 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[9]. 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. There may be many rounds of updating the system based on coders' feedback. There were examples in NER + L tools, which are yet to be deployed for clinical coding: in MedCATTrainer[64], a dedicated interface is deployed for users to add new concepts, new synonyms and abbreviations, corrections of concepts (of samples selected using active learning), and binary annotations of temporality and phenotyping, then the model is re-run with the feedback; an interface is also designed in SemEHR[65] to allows users to add labels for mentions, which is used to either train a confidence model or to form post-processing rules to refine the results; manually added rules may also be integrated with weak supervision to generate coded data for training[66,67]. 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; active learning is deployed in NER + L in MedCATTrainer[64], and evaluated in automated coding to potentially reduce human annotations[68].
- 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)[37]. 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 at $K$ (or the percentage of correct codes in top-$K$ predictions, $K = 10$ or 15)[37,46,47]. 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[69].
- Adaptation to terminology changes—how a trained model can be adapted to modified standards for coding or a completely new ontology (for example from ICD-10 to ICD-11[24])? As we described earlier, ICD-11 is semantically more complex than ICD-10 with a poly-hierarchical 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 the above-mentioned robust few-shot and zero-shot learning for new codes with no or few training data.
- Knowledge representation and reasoning in coding—finally and most fundamentally, many of the above technical directions suggest to integrate knowledge or semantic information in coding classification systems and ontologies. ICD code descriptions[19,55] and hierarchies[41,42] have been considered in recent studies (and see the blog about hierarchical evaluation[70] for ref. [41]). Other ontologies, such as CCS[71] and code synonyms in UMLS, have been adopted recently to achieve state-of-the-art performance[45,51]. Also, manual coding is mainly based on a standard process and coding guidelines, potentially formalised as a set of rules and terminologies deployed in the healthcare system, for example, the priority of certain codes, the number of codes for each case, the mutual exclusion among certain codes, the rules to code hypothetical cases (e.g., possible and probable), the locally defined specific codes, etc. An example of formalising and integrating rules regarding the mutual exclusion of codes

and hypothetical cases with machine learning is presented in the study[29]. These guidelines need to be formally represented in a machine-readable way and to be iteratively integrated into the deep learning-based automated coding system.

While multi-label classification is a straightforward formulation of clinical coding, another approach is through named entity extraction and linking or NER + L (for example in the work of MedCAT[35] and the study of rare disease identification[66,67] with SemEHR[65]), although less adopted in the recent literature. NER + L is based on the general approach of clinical information extraction, which is also more recently enhanced by deep learning[72]. NER + L is explainable and feasible, as it inherently links the code to the piece of text in the document and helps handle the long document problem, but the extracted codes still need to be summarised to the final set of codes, and abide by the standard process and guidelines of coding. NER + L methods may help alleviate the coding of few-shot and zero-shot codes by extracting the concepts in the target ontology from clinical notes. A downside of NER + L-based coding is that it requires contextual understanding, i.e., the negation, temporality, and experiencer of the extracted concept or code[35,65], which are not needed using the multi-label classification approach. These two formulations (multi-label classification and NER + L) may be combined in the design of a clinical coding system. A recent attempt is to use either text enrichment or multi-task learning to integrate NER + L identified concepts[36], which however does not improve over the multi-label classification approach, and warrants future studies for alternative methods. The study[73] uses NER + L and ontologies to help synthesise clinical notes by replacing words with synonyms or with names of sibling codes (thus to predict the sibling codes) to potentially improve few- and zero-shot coding. Also, the study[62] used NER + L to explore the under-coded problem of clinical coding. The study[74] proposes to rank ICD-10 codes extracted from an off-the-shelf NER + L system for billing code prediction, which better addresses the few- and zero- shot problem than multi-label classification. More benchmarking results for NER + L enhanced methods are needed for comparison.

Automated clinical coding systems also need to be tailored for different purposes (e.g., billing vs. health-related research) and contexts (e.g., countries). For billing purposes, automated coding systems aim at predicting Diagnosis-Related Groups (DRGs) in the US (and Healthcare Resource Groups, HRGs in the UK), which have a smaller number of codes, usually grouped from the full set of ICD codes but can potentially be predicted prior to the ICD coding[75]. For health-related research, automated coding task needs a variety of classification systems (usually with high granularity) for use in case detection or phenotyping, thus other terminologies (e.g., SNOMED CT[76], ORDO[67,77], and ICD-11 in the near future[25]) and customised terminologies (e.g., for sub-stroke phenotyping[78]), and also see the surveys[8,79]. NER + L systems with rule-based inference can help improve the phenotyping when data are scarce to be used for supervised learning[67,80]. Automated coding systems can also be jointly designed with clinical outcome predictions (e.g., readmission and mortality) using deep learning in an end-to-end manner[81]. Also, case detection in some health-related research may favour precision (PPV) than recall (sensitivity) for evaluation[82], which needs to be considered in building and tuning the automated coding system. In terms of other country-related factors, a known issue mentioned earlier in the US is "Note Bloat", where content-importing shortcuts like copy-and-paste are used, which may reduce the time of documentation[58]. The "Note Bloat" phenomenon exacerbates the redundant entry of data in notes that is pulled in or copy-and-pasted from discrete places (e.g., various charts) in the Electronic Health Record (EHR). Training a model to fill codes to the charts needs to remove information from the notes (e.g., ICD codes) that is already present in the charts in the EHR. Also, it is shown that de-duplication of clinical notes improves the performance of prediction tasks, including predicting codes in the DRGs for billing[58]. More country related factors, e.g., billing and insurance, may also affect the system design and would warrant future studies.

Besides, industry organisations, beyond healthcare institutions and academia, play a key role related to automated clinical coding. There are also increasing collaborations between industry and academia. The Epic EHR system is deployed in the University College London Hospital (UCLH) for the management of EHRs. Recently, the CogStack team (including King's College Hospital (KCH), NIHR Mausley Biomedical Research Centre, and UCLH) is collaborating with the UCLH Epic team to integrate an NLP component into the NoteReader interface in the Epic system. The NER + L tool MedCAT is planned to be deployed to populate structured information (by extracting concepts including diagnosis, symptoms, medications, etc.) from newly-created clinical notes to reduce documentation time and verify and complement structured information[83]. Working with five NHS Trusts in England, the CogStack team has also received an AI award from the National Institute of Health Research for developing AI-based clinical coding of medical records (see news from KCH[84]). The project aims to enable more efficient and accurate analysis, free up staff time, and improve research. Industry NER + L APIs (e.g., Amazon Comprehend Medical InferICD10CM[85], Microsoft Text Analytics for health[86] and Google Healthcare Natural Language API[87]) have been released during the last two to three years[88,89] to support clinical concept extraction from texts with price charges. Many technology companies in the industry also provide proprietary solutions and paid services for (semi-)automated clinical coding including Deloitte[90], Optum[91], Capita[92] and CHKS[93]. However, the research access and inner working of the systems are usually not available, leaving it hard to contrast and compare technically. Due to its promising potentials both clinically and financially, the automated coding also attracts great attentions from start-up companies. For example, AKASA in the US is developing a deep-learning based solution, aiming to tackle automated clinical coding adapting a multi-label classification approach. They reported performances with the state-of-the-art results on MIMIC-III full codes, better than human coding in the experiments[50] (also see news[94]). These contribute to the overall picture of the promising potential of automated clinical coding.

## CONCLUSION

In this paper, we reviewed the task of automated clinical coding from the perspectives of AI researchers and clinical coding professionals, what it is and why it is an important task, and summarised the challenges of the recent deep learning methods for the task. We then position several key directions for future studies.

While we summarised the *technical* challenges, there are many *organisational* challenges to be addressed to deploy an AI-based coding tool into the clinical coding environment, as reviewed in Campbell and Giadresco[9], where an essential idea is that coders need to be involved in the model development and deployment stage. Coders are usually occupied with their coding work and it may not be easy to engage them for system testing. Further research support on projects in medical informatics and computer science is needed to address these challenges.

How far are we from automated clinical coding that is human-centred, explainable, intelligent, and robust to complex real-world scenarios? We cannot give a concrete estimation, but it seems we now have a clearer path and a list of challenges to address. With the growing number of studies and projects in academia and the industry, we look forward to seeing more advances in AI-assisted clinical coding in the next five years and beyond and its application into practice in the near future.

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## DATA AVAILABILITY

The authors declare that all data supporting the findings of this perspective article are available within the paper.

Received: 30 March 2022; Accepted: 7 October 2022;
Published online: 22 October 2022

## REFERENCES

1. Public Health Scotland. National Data Catalogue. General acute inpatient and day case - Scottish Morbidity Record (SMR01). https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=5 (2020).
2. American Academy of Professional Coders (AAPC). What is medical coding? https://www.aapc.com/medical-coding/medical-coding.aspx (2022).
3. NHS Digital. Clinical coding for non coders. https://hscic.kahootz.com/gf2.ti/f/762498/30719205.1/PPSX/-/Coding_for_non_coders_automaticnew.ppsx (2017).
4. Enrico, C. In *Guide to Health Informatics* Ch. 24 (Taylor & Francis Group, 2015).
5. National Center for Health Statistics. International Classification of Diseases, (ICD-10-CM/PCS) transition – background. https://www.cdc.gov/nchs/icd/icd10cm_pcs_background.htm (2015).
6. Public Health Scotland. Terminology Services. Scottish Clinical Coding Standards. https://www.isdscotland.org/Products-and-services/Terminology-services/Clinical-coding-guidelines/ (2022).
7. Otero Varela, L. et al. International Classification of Diseases clinical coding training: an international survey. *Health Inf. Manag.* https://doi.org/10.1177/18333583221106509 (2022)
8. Stanfill, M. H., Williams, M., Fenton, S. H., Jenders, R. A. & Hersh, W. R. A systematic literature review of automated clinical coding and classification systems. *J. Am. Med Inf. Assoc.* **17**, 646–651 (2010).
9. Campbell, S. & Giadresco, K. Computer-assisted clinical coding: a narrative review of the literature on its benefits, limitations, implementation and impact on clinical coding professionals. *HIM J.* **49**, 5–18 (2020).
10. Jiang, F. et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc. Neurol*. https://doi.org/10.1136/svn-2017-000101 (2017)
11. Kaur, R., Ginige, J. A. & Obst, O. AI-based ICD coding and classification approaches using discharge summaries: A systematic literature review. *Expert Syst. Appl*. 118997 https://doi.org/10.1016/j.eswa.2022.118997 (2022).
12. Ji, S., Sun, W., Dong, H., Wu, H. & Marttinen, P. A unified review of deep learning for automated medical coding. Preprint at *arXiv* http://arxiv.org/abs/2201.02797 (2022).
13. Teng, F. et al. A review on deep neural networks for ICD coding. In *IEEE Transactions on Knowledge and Data Engineering* 1–19 (IEEE, 2022)
14. Alonso, V. et al. Problems and barriers during the process of clinical coding: a Focus Group Study of coders' perceptions. *J. Med. Syst.* **44**, 62 (2020).
15. Burns, E. M. et al. Systematic review of discharge coding accuracy. *J. Public Health* **34**, 138–148 (2012).
16. Public Health Scotland. Data quality assurance. Assessment of SMR01 Data Scotland Report 2019 V1. https://beta.isdscotland.org/media/7465/assessment-of-smr01-data-scotland-report-2019-v1.pdf (2019).
17. Wooldridge, M. *The Road to Conscious Machines: The Story of AI* (Penguin UK, 2020).
18. Russell, S. J. & Norvig, P. *Artificial Intelligence: A Modern Approach, Global Edition* (Pearson, 2021).
19. Mullenbach, J., Wiegreffe, S., Duke, J., Sun, J. & Eisenstein, J. Explainable prediction of medical codes from clinical text. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* 1101–1111 (Association for Computational Linguistics, 2018).
20. Johnson, A. E. W. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 1–9 (2016).
21. Barrows Jr, R. C., Busuioc, M. & Friedman, C. Limited parsing of notational text visit notes: ad-hoc vs. NLP approaches. In *Proc. AMIA Symposium* 51 (American Medical Informatics Association, 2000).
22. World Health Organization. *ICD-11 for Mortality and Morbidity Statistics* (WHO, 2022).
23. World Health Organization. WHO's new International Classification of Diseases (ICD-11) comes into effect. https://www.who.int/news/item/11-02-2022-who-s-new-international-classification-of-diseases-(icd-11)-comes-into-effect (2022).
24. Gaebel, W., Stricker, J. & Kerst, A. Changes from ICD-10 to ICD-11 and future directions in psychiatric classification. *Dialogues Clin. Neurosci.* **22**, 7–15 (2020).
25. Chute, C. G. The rendering of human phenotype and rare diseases in ICD-11. *J. Inherit. Metab. Dis.* **41**, 563–569 (2018).
26. World Health Organization. ICD-11 Reference Guide. 2.10 Precoordination and postcoordination. https://icdcdn.who.int/icd11referenceguide/en/html/index.html#precoordination-and-postcoordination (2022).
27. Bengio, Y., Lecun, Y. & Hinton, G. Deep learning for AI. *Commun. ACM* **64**, 58–65 (2021).
28. Dinwoodie, H. P. & Howell, R. W. Automatic disease coding: the 'fruit-machine' method in general practice. *Br. J. Prev. Soc. Med.* **27**, 59–62 (1973).
29. Farkas, R., & Szarvas, G. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics* **9**, 1–9 (2008).
30. Zhou, L., Cheng, C., Ou, D. & Huang, H. Construction of a semi-automatic ICD-10 coding system. *BMC Med. Inform. Decis. Mak.* **20**, 1–12 (2020).
31. Shi, H., Xie, P., Hu, Z., Zhang, M. & Xing, E. P. Towards automated ICD coding using deep learning. Preprint at *arXiv* https://arxiv.org/abs/1711.04075 (2017).
32. Karimi, S., Dai, X., Hassanzadeh, H. & Nguyen, A. Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. in *BioNLP 2017* 328–332 (Association for Computational Linguistics, 2017).
33. acadTags. Awesome-medical-coding-NLP. https://github.com/acadTags/Awesome-medical-coding-NLP (2022).
34. Nam, J., Kim, J., Loza Mencía, E., Gurevych, I. & Fürnkranz, J. In *Machine Learning and Knowledge Discovery in Databases* (eds. Calders, T., Esposito, F., Hüllermeier, E. & Meo, R.) 437–452 (Springer, 2014).
35. Kraljevic, Z. et al. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artif. Intelligence Med.* **117**, 102083 (2021).
36. Wiegreffe, S., Choi, E., Yan, S., Sun, J. & Eisenstein, J. Clinical concept extraction for document-level coding. In *Proc. 18th BioNLP Workshop and Shared Task* 261–272 (Association for Computational Linguistics, 2019)
37. Rios, A. & Kavuluru, R. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing* 3132–3142 (Association for Computational Linguistics, 2018).
38. Teng, F., Yang, W., Chen, L., Huang, L. & Xu, Q. Explainable prediction of medical codes with knowledge graphs. *Front. Bioeng. Biotechnol.* **8**, 867 (2020).
39. Xie, X., Xiong, Y., Yu, P. S. & Zhu, Y. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proc. 28th ACM International Conference on Information and Knowledge Management* 649–658 (ACM, 2019).
40. Cao, P. et al. Hypercore: hyperbolic and co-graph representation for automatic ICD coding. In *Proc. 58th Annual Meeting of the Association for Computational Linguistics* 3105–3114 (Association for Computational Linguistics, 2020).
41. Falis, M. et al. Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text. In *Proc. Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)* 168–177 (Association for Computational Linguistics, 2019).
42. Falis, M., Dong, H., Birch, A. & Alex, B. CoPHE: a count-preserving hierarchical evaluation metric in large-scale multi-label text classification. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 907–912 (Association for Computational Linguistics, 2021).
43. Kukafka, R., Bales, M. E., Burkhardt, A. & Friedman, C. Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health. *J. Am. Med. Inform. Assoc.* **13**, 508–515 (2006).
44. Ji, S., Hölttä, M. & Marttinen, P. Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers Biol. Med.* **139**, 104998 (2021).
45. Sun, W., Ji, S., Cambria, E. & Marttinen, P. Multitask balanced and recalibrated network for medical code prediction. *ACM Trans. Intelligent Syst. Technol*. https://doi.org/10.1145/3563041 (2022)
46. Chalkidis, I. et al. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proc. 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 7503–7515 (Association for Computational Linguistics, 2020).
47. Wang, R. et al. Meta-LMTC: meta-learning for large-scale multi-label text classification. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 8633–8646 (Association for Computational Linguistics, 2021).
48. Xu, K. et al. Multimodal machine learning for automated ICD coding. In *Machine Learning for Healthcare Conference* 197–215 (PMLR, 2019).
49. Liu, Y., Cheng, H., Klopfer, R., Gormley, M. R. & Schaaf, T. Effective convolutional attention network for multi-label clinical document classification. In *Proc. 2021 Conference on Empirical Methods in Natural Language Processing* 5941–5953 (Association for Computational Linguistics, 2021).
50. Kim, B. H., & Ganapathi, V. Read, attend, and code: pushing the limits of medical codes prediction from clinical notes by machines. In *Machine Learning for Healthcare Conference* 196–208 (PMLR, 2021).

51. Yuan, Z., Tan, C., & Huang, S. Code synonyms do matter: multiple synonyms matching network for automatic ICD coding. In *Proc. 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* 808–814 (Association for Computational Linguistics, 2022).
52. Huang, C. W., Tsai, S. C., & Chen, Y. N. PLM-ICD: automatic ICD coding with pretrained language models. In *Proc. 4th Clinical Natural Language Processing Workshop* 10–20 (Association for Computational Linguistics, 2022).
53. Terminology and Classifications Delivery Service, National Health Service Digital. National Clinical Coding Standards ICD-10 5th Edition. https://classbrowser.nhs.uk/ref_books/ICD-10_2021_5th_Ed_NCCS.pdf (2021).
54. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 4171–4186 (Association for Computational Linguistics, 2019).
55. Feucht, M., Wu, Z., Althammer, S. & Tresp, V. Description-based label attention classifier for explainable ICD-9 classification. In *Proc. Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* 62–66 (Association for Computational Linguistics, 2021).
56. Yogarajan, V., Pfahringer, B., Smith, T., & Montiel, J. In *Artificial Neural Networks and Machine Learning – ICANN 2022* (eds. Pimenidis, E., Angelov, P., Jayne, C., Papaleonidas, A. & Aydin, M.) 209–221 (Springer Nature Switzerland, 2022).
57. Michalopoulos, G., Malyska, M., Sahar, N., Wong, A. & Chen, H. ICDBigBird: a contextual embedding model for ICD code classification. In *Proc. 21st Workshop on Biomedical Language Processing* 330–336 (Association for Computational Linguistics, 2022).
58. Liu, J., Capurro, D., Nguyen, A. & Verspoor, K. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *J. Biomed. Inform.* **133**, 104149 (2022)
59. Searle, T., Ibrahim, Z., Teo, J. & Dobson, R. Estimating redundancy in clinical text. *J. Biomed. Inform.* **124**, 103938 (2021).
60. Gao, S. et al. Limitations of transformers on clinical text classification. *IEEE J. Biomed. Health Inform.* **25**, 3596–3607 (2021).
61. Dong, H., Suárez-Paniagua, V., Whiteley, W. & Wu, H. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *J. Biomed. Inform.* **116**, 103728 (2021).
62. Searle, T., Ibrahim, Z. & Dobson, R. Experimental evaluation and development of a silver-standard for the MIMIC-III clinical coding dataset. In *Proc. 19th SIGBioMed Workshop on Biomedical Language Processing* 76–85 (Association for Computational Linguistics, 2020).
63. Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M. & Elhadad, N. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*. 409-416 (2018).
64. Searle, T., Kraljevic, Z., Bendayan, R., Bean, D. & Dobson, R. MedCATTrainer: a biomedical free text annotation interface with active learning and research use case specific customisation. In *Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations* 139–144 (Association for Computational Linguistics, 2019).
65. Wu, H. et al. SemEHR: a general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *J. Am. Med. Inform. Assoc.* **25**, 530–537 (2018).
66. Dong, H. et al. Rare disease identification from clinical notes with ontologies and weak supervision. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)* 2294–2298 (EMBC, 2021).
67. Dong, H. et al. Ontology-based and weakly supervised rare disease phenotyping from clinical notes. Preprint at http://arxiv.org/abs/2205.05656 (2022).
68. Ferreira, M. D. et al. Active learning for medical code assignment. In *Workshops from ACM Conference on Health, Inference, and Learning (CHIL) 2021*. Preprint at *arXiv* http://arxiv.org/abs/2104.05741 (2021).
69. Chen, J. et al. Knowledge-aware zero-shot learning: survey and perspective. In *Proc. Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021* 4366–4373 (IJCAI, 2021).
70. Falis, M. Blood is thicker than water, a hierarchical evaluation metric for document classification. https://www.ltg.ed.ac.uk/blood-is-thicker-than-water/ (2021).
71. Healthcare Cost and Utilization Project (HCUP). Clinical classifications software (CCS) for ICD-9-CM. https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp (2017).
72. Hahn, U. & Oleynik, M. Medical information extraction in the age of deep learning. *Yearb. Med. Inform.* **29**, 208–220 (2020).
73. Falis, M., Dong, H., Birch, A. & Alex, B. Horses to zebras: ontology-guided data augmentation and synthesis for ICD-9 coding. In *Proc. 21st Workshop on Biomedical Language Processing* 389–401 (Association for Computational Linguistics, 2022).
74. DeYoung, J., Shing, H.-C., Kong, L., Winestock, C. & Shivade, C. Entity anchored ICD coding. Accepted to American Medical Informatics Association (AMIA) 2022 Annual Symposium. Preprint at *arXiv* http://arxiv.org/abs/2208.07444 (2022).
75. Liu, J., Capurro, D., Nguyen, A. & Verspoor, K. Early prediction of diagnostic-related groups and estimation of hospital cost by processing clinical notes. *NPJ Digital Med.* **4**, 1–8 (2021).
76. Donnelly, K. SNOMED-CT: The advanced terminology and coding system for eHealth. *Stud. Health Technol. Inform.* **121**, 279 (2006).
77. Vasant, D. et al. ORDO: an ontology connecting rare disease, epidemiology and genetic data. In *Bio-Ontology @ ISMB 2014*. 1-4. https://www.researchgate.net/publication/281824026_ORDO_An_Ontology_Connecting_Rare_Disease_Epidemiology_and_Genetic_Data (2014).
78. Alex, B. et al. Text mining brain imaging reports. *J. Biomed. Semant.* **10**, 1–11 (2019).
79. Ford, E., Carroll, J. A., Smith, H. E., Scott, D. & Cassell, J. A. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J. Am. Med. Inform. Assoc.* **23**, 1007–1015 (2016).
80. Rannikmäe, K. et al. Developing automated methods for disease subtyping in UK Biobank: an exemplar study on stroke. *BMC Med. Inform. Decis. Mak.* **21**, 1–9 (2021).
81. Lovelace, J., Hurley, N. C., Haimovich, A. D. & Mortazavi, B. J. Dynamically extracting outcome-specific problem lists from clinical notes with guided multi-headed attention. In *Machine Learning for Healthcare Conference* 245–270 (PMLR, 2020).
82. Rannikmäe, K. et al. Accuracy of identifying incident stroke cases from linked health care data in UK Biobank. *Neurology* **95**, e697–e707 (2020).
83. Noor, K. et al. Deployment of a free-text analytics platform at a UK National Health Service Research Hospital: CogStack at University College London Hospitals. *JMIR Med. Inform.* **10**, e38122 (2022).
84. King's College Hospital NHS Foundation Trust. CogStack wins an artificial intelligence in health and care. https://www.kch.nhs.uk/news/public/news/view/34965 (2021).
85. Amazon Web Services. ICD-10-CM linking. https://docs.aws.amazon.com/comprehend-medical/latest/dev/ontology-icd10.html (2022).
86. Azure. What is text analytics for health in Azure Cognitive Service for Language? https://docs.microsoft.com/en-us/azure/cognitive-services/language-service/text-analytics-for-health/overview?tabs=ner (2022).
87. Google Cloud. Healthcare natural language API. https://cloud.google.com/healthcare-api/docs/concepts/nlp (2022).
88. Bodnari, A. Healthcare gets more productive with new industry-specific AI tools. https://cloud.google.com/blog/topics/healthcare-life-sciences/now-in-preview-healthcare-natural-language-api-and-automl-entity-extraction-for-healthcare (2020).
89. Amazon Web Services. Announcing ICD-10-CM and RxNorm ontology linking for Amazon Comprehend Medical. https://aws.amazon.com/about-aws/whats-new/2019/12/announcing-icd-10-cm-rxnorm-ontology-linking-amazon-comprehend-medical/ (2019).
90. Miranda, M. Automated clinical coding. The AI-based solution to address the critical shortage of clinical coders. https://www2.deloitte.com/au/en/blog/consulting-blog/2020/automated-clinical-coding.html (2020).
91. Optum Inc. Enterprise computer-assisted coding (CAC). https://www.optum360.com/solutions/coding-and-documentation/coding-and-cdi-technology/enterprise-cac.html (2022).
92. Capita plc. Creating better health outcomes with automated clinical coding. https://www.capita.com/expertise/industry-specific-services/health-services/healthcare-business-operations/clinical-coding/automated-clinical-coding (2022).
93. CHKS. Automated clinical coding. https://www.chks.co.uk/Clinical-coding (2022).
94. Mace, S. Making medical coding better and faster with artificial intelligence. Medical Technology Schools. https://www.medicaltechnologyschools.com/health-information-technology/medical-coding-and-artificial-intelligence (2021).

## ACKNOWLEDGEMENTS

The work is supported by WellCome Trust iTPA Awards (PIII009, PIII032), Health Data Research UK National Phenomics and Text Analytics Implementation Projects, and the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. H.D. and J.C. are supported by the Engineering and Physical Sciences Research Council (EP/V050869/1) on "ConCur: Knowledge Base Construction and Curation". HW was supported by Medical Research Council and Health Data Research UK (MR/S004149/1, MR/S004149/2); British Council (UCL-NMU-SEU international collaboration on Artificial Intelligence in Medicine: tackling challenges of low generalisability and health inequality); National Institute for Health Research (NIHR202639); Advanced Care Research Centre at the University of Edinburgh. We thank constructive comments from Murray Bell and Janice Watson in Terminology Service in Public Health Scotland, and information provided by Allison Reid in the coding department in NHS Lothian, Paul Mitchell, Nicola Symmers, and Barry Hewit in Edinburgh Cancer Informatics, and staff in Epic Systems Corporation. Thanks for the suggestions from Dr. Emma Davidson regarding clinical research. Thanks to the discussions with Dr. Kristiina Rannikmäe regarding the research on clinical coding

and with Ruohua Han regarding the social and qualitative aspects of this research. In Fig. 1, the icon of "Clinical Coders" was from Freepik in Flaticon, https://www.flaticon.com/free-icon/user_747376; the icon of "Automated Coding System" was from Free Icon Library, https://icon-library.com/png/272370.html.

## AUTHOR CONTRIBUTIONS

H.D. and H.W. conceived the research on automated clinical coding. H.D., H.W., W.W., B.A., and F.M. discussed the research regularly. H.D., H.W., B.A., and W.W. initiated the writing of this paper. H.D. and F.M. discussed the research with practitioners in clinical coding and received their feedback, with support from H.W. and W.W. S.J. and J.C. provided comments on AI for clinical coding as researchers in the field of AI. J.M. provided comments on the contexts and the design of clinical coding systems from the perspective of the industry. H.D. drafted the manuscript. All authors edited and revised the manuscript and approved the final version of the manuscript.

## COMPETING INTERESTS

The authors declare no competing interests.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-022-00705-7.

**Correspondence** and requests for materials should be addressed to Hang Dong or Honghan Wu.