Automated Clinical Coding

Citation for published version:
Dong, H, Falis, M, Whiteley, W, Alex, B, Ji, S, Chen, J & Wu, H 2022 'Automated Clinical Coding: What, Why, and Where We Are?' ArXiv. https://doi.org/10.48550/arXiv.2203.11092

Digital Object Identifier (DOI):
10.48550/arXiv.2203.11092

Document Version:
Early version, also known as pre-print

Automated Clinical Coding: What, Why, and Where We Are?

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

[1]Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom; [2]School of Informatics, University of Edinburgh, Edinburgh, United Kingdom; [3]Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom; [4]Edinburgh Futures Institute, Edinburgh, United Kingdom; [5]Department of Computer Science, Aalto University, Espoo, Finland; [6]Department of Computer Science, University of Oxford, Oxford, United Kingdom; [7]Institute of Health Informatics, University College London, London, United Kingdom

Abstract

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.

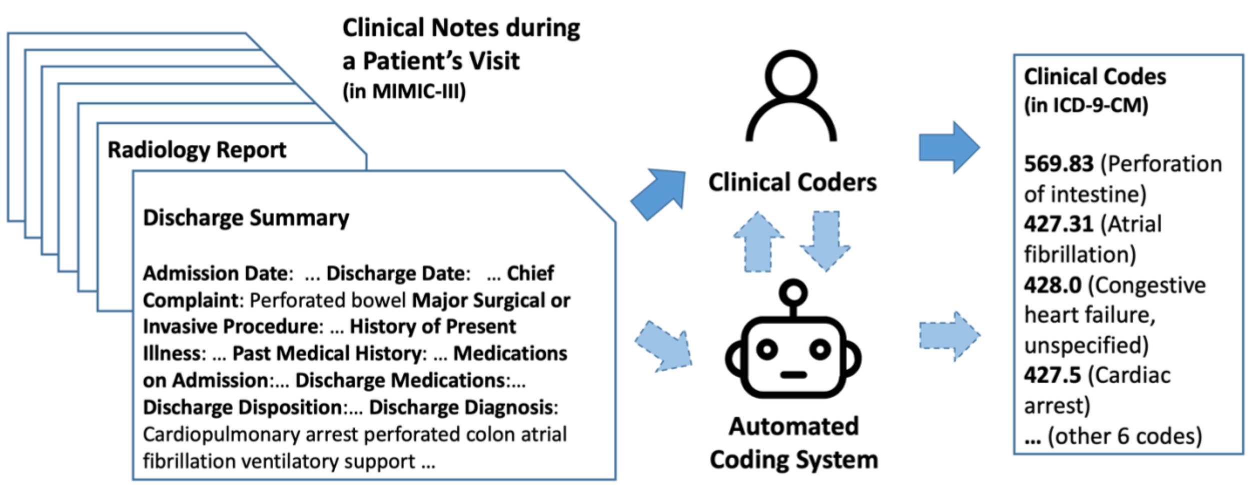

Figure 1[1]. An example of clinical coding, manual and automated (linked with solid and dashed arrows, respectively), with ICD-9-CM codes from a clinical note in the MIMIC-III dataset [9] of ICU patients in 2001-2012 in a hospital in the US. Dashed arrows between clinical coders and the automated coding 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 available.

---

[1] The icon of "Clinical Coders" was from Freepik in Flaticon, https://www.flaticon.com/free-icon/user_747376, under the Flaticon licence (attribution required). The icon of "Automated Coding System" was from https://icon-library.com/png/272370.html, under the Attribution 3.0 Unported (CC BY 3.0) licence.

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). 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, finally contributing to the Scottish Morbidity Records (SMR01) national dataset[2]. The process of coding usually includes data abstraction or summarisation [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 used for billing purposes in the US[3]. For introductory slides about clinical coding in the UK provided by NHS Digital, see *Clinical coding for non coders*[4].

Clinical coding is a non-trivial task for humans. 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 codes. 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 [2, 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[5]). Usually, it can take months and longer to train an expert clinical coder in the NHS (National Health Service) in the UK.

Automated clinical coding is the idea that clinical coding may be automated by computers using AI technologies [3]. It is a branch of computer-assisted coding (CAC) [4]. In recent years, AI has been considered as a promising approach to transforming healthcare by intelligently processing the increasing amount of data with machine learning and NLP techniques [5]. 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) for the last few years, as reviewed in [6-8].

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 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 Figure 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 coder in NHS Scotland usually codes about 60 cases a day (equivalent to 7-8min for each case) and an NHS coding department of around 25 to 30 coders usually codes over 20,000 cases per month. Even so, there is a backlog of cases to be coded, which can take several months or more (e.g., over a year [10]). 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 [1]. The average accuracy of coding in the UK was around 83% with a large variance among studies (50-98%) [11]. In Scotland, the accuracy of coding is very high[6] (e.g. in 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

[2] https://www.ndc.scot.nhs.uk/National-Datasets/data.asp?SubID=5
[3] https://www.aapc.com/medical-coding/medical-coding.aspx
[4] https://hscic.kahootz.com/gf2.ti/f/762498/30719205.1/PPSX/-/Coding_for_non_coders_automaticnew.ppsx
[5] https://www.isdscotland.org/Products-and-services/Terminology-services/Clinical-coding-guidelines/
[6] https://beta.isdscotland.org/media/7465/assessment-of-smr01-data-scotland-report-2019-v1.pdf

review [4]. We believe that with recent AI technologies (for example, NLP), automated coding has the potential to better support clinical coders. We mostly focus on the case that AI directly contributes to 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 [12-13], 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 [14] in only the discharge summaries in a US intensive care dataset, MIMIC-III [9]), and use variable abbreviations (e.g., "a [xx] y/o M w/ Hep C, HTN, CKD, a/w HTN emergency" in a discharge summary in MIMIC-III) and terse symbols (e.g., the use of "?" to denote uncertainty and "+" to denote a positive test). 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 has around 68,000 codes in a large hierarchy. The ICD-11 system (started in Jan 2022, but yet to be used in practice at the time of writing) introduces significant changes in chapter structure, diagnostic categories, diagnostic criteria, etc. [15] Besides, classification standards are updated regularly (e.g. usually every few months in Public Health Scotland[7]). 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 Figure 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? [4]

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 [12]. 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 [12-13]. Neural networks returned in the mid-1980s with *machine learning* in general. *Deep learning* methods became the mainstream of AI after 2011 [13], continuing to evolve today [16].

Coming back to automated clinical coding, while the task has been studied for around 50 years (with the earliest studies around 1970 [17]), the current deep learning-based methods have a short history. Only since around 2017 [18-19], deep learning has been applied to automated coding and there are abundant studies in this area (reflected in recent surveys in [6-8] and curation of papers in automated medical coding[8]). Most of the studies formulate the task as a multi-label classification problem [20], while some studies formulate the task as a concept extraction or a Named Entity Recognition and Linking (NER+L) problem [21-22]. 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.

---

[7] https://www.isdscotland.org/Products-and-services/Terminology-services/Clinical-coding-guidelines/

[8] https://github.com/acadTags/Awesome-medical-coding-NLP

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., 2006 [23], and also quoted in Stanfill *et al.*, 2010 [1]

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 [14,24], multi-task learning [25-26], zero-shot learning [27-28], meta-learning [29], multi-modal learning [30], 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 8,932 ICD-9 codes for the MIMIC-III data was under 60% (between 58%-59%)** [25,31-33]. MIMIC-III discharge summaries [9], although coded with ICD-9-CM (the ninth version of ICD, Clinical Modification), are the main dataset used for benchmarking [14]. This dataset is also now older (collected over ten years ago, from 2001-2012), and only represents an intensive care dataset 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. **Handing unseen, infrequent, and imbalanced labels**: In the MIMIC-III dataset, around 5,000 codes appear fewer than 10 times in the training data and over 50% of codes never appear [27]. 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 [3,23]. 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 [2]. 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.

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 [34]) usually require a limited length of up to 512 tokens as input due to the memory-demanding self-attention mechanism, while discharge summaries *alone* in MIMIC-III have on average around 1,500 tokens [14] (and up to over 10,000 tokens). More recent studies applied Longformer [35] and TransformerXL [36] to clinical coding to process documents of over 3,000 tokens, but this is still insufficient for clinical notes.

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 clinical coding [24,37-38]. 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 [37].

Besides, as we stated previously, manual coding is largely based on a standard and implied process with rules applied to the healthcare system. 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 [8].

- **Creating gold standard coding datasets** – the current widely used benchmark dataset MIMIC-III may have been significantly under-coded [39]. There is a lack of large, openly available, and expert-labelled datasets from Electronic Health Records in this area, and models trained on MIMIC-III may not simply generalise to other datasets due to the difference of length, style, and language (for example, clinical notes in China, Spain, or even the UK). Various expert-labelled coding datasets 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 datasets 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 [10]. 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 [30]. *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. Work in this area so far uses label-wise attention mechanisms to highlight key *n*-grams [14], words, and sentences [38,40]. 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 into the system [4]. 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.
- **Few-shot and zero-shot learning** - many codes have a low frequency or even no occurrence in the training data, this is a key problem for multi-label classification with many labels (e.g. 68,000 codes in ICD-10) [27]. The best systems so far to work with low-frequent (<5 times) codes on the MIMIC-III dataset are still below or around 40% recall at K (or the percentage of correct codes in top-K predictions, K=10 or 15) [27-29]. Better support for few-shot and zero-shot learning will improve the overall coding performance and usage.
- **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 [15])? This may require novel paradigms in deep learning (e.g. self-supervised and transfer learning), accurate ontology matching, and robust zero-shot learning.
- **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 [14,35] and hierarchies [26,41] have been considered in recent studies (and see the blog about hierarchical evaluation[9] for [41]). Other ontologies, such as CCS[10] used in [25] and code synonyms in UMLS used in [33], have been adopted recently to achieve state-of-the-art performance. 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 locally defined specific codes, etc. These guidelines need to be formally represented in a machine-readable way and to be iteratively integrated into the 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 in [21] and the study of rare disease coding [42] with SemEHR [43]), although less adopted in the recent literature. NER+L is a more explainable and feasible approach, 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, abide by the standard

[9] https://www.ltg.ed.ac.uk/blood-is-thicker-than-water/
[10] https://www.hcup-us.ahrq.gov/toolssoftware/ccs/ccs.jsp

process and guidelines of coding. These two formulations (classification and NER+L) may be combined in the design of a clinical coding system. A recent attempt at this is presented in [22].

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 [4], 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. A recent, large project in the UK to consider these issues in coding is the Artificial Intelligence in Health and Care Award to CogStack by King's College Hospital (KCH), NIHR Mausley Biomedical Research Centre, and University College London Hospital (see news from KCH[11]).

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, 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.

Ethical declaration: The opinions expressed in this article are the authors' own and do not reflect the view of Public Health Scotland or the National Health Service (NHS) in the UK. This work is not to be considered as research that directly involves human subjects (e.g., clinical coders or relevant practitioners), as we do not explicitly collect data from them, instead, we received informal comments and high-level information that does not disclose any personal information. We acknowledge all the people who provided support for this study below.

Acknowledgement: The work is supported by WellCome Trust iTPA Awards (PIII009, PIII032), Health Data Research UK National Phenomics and Text Analytics Implementation Projects. We thank the 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 and Paul Mitchell, Nicola Symmers, and Barry Hewit in Edinburgh Cancer Informatics. Thanks to the discussions with Dr. Kristiina Rannikmae regarding the research on clinical coding and with Ruohua Han regarding the social and qualitative aspects of this research.

References

1. Enrico C. Chapter 24 Natural language and formal Terminology. In: Guide to Health Informatics, *Taylor & Francis Group*, 2015: 401-414.
2. Terminology and Classifications Delivery Service, National Health Service Digital. *National Clinical Coding Standards ICD-10 5th Edition*. NHS Digital. 2021:11, https://classbrowser.nhs.uk/ref_books/ICD-10_2021_5th_Ed_NCCS.pdf
3. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association*. 2010 Nov 1;17(6):646-51.
4. 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. *Health Information Management Journal*. 2020 Jan;49(1):5-18.
5. Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, Wang Y, Dong Q, Shen H, Wang Y. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology*. 2017 Dec 1;2(4).
6. Kaur R, Ginige JA, Obst O. A Systematic Literature Review of Automated ICD Coding and Classification Systems using Discharge Summaries. *arXiv preprint* arXiv:2107.10652. 2021 Jul 12.
7. Ji S, Sun W, Dong H, Wu H, Marttinen P. A Unified Review of Deep Learning for Automated Medical Coding. *arXiv preprint* arXiv:2201.02797. 2022 Jan 8.
8. Teng F, Liu Y, Li T, Zhang Y, Li S, Zhao Y. A review on deep neural networks for ICD coding. *IEEE Transactions on Knowledge and Data Engineering*. 2022 Feb 4.

[11] https://www.kch.nhs.uk/news/public/news/view/34965

9. Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. *Scientific Data*. 2016 May 24;3(1):1-9.
10. Alonso V, Santos JV, Pinto M, Ferreira J, Lema I, Lopes F, Freitas A. Problems and barriers during the process of clinical coding: a focus group study of coders' perceptions. Journal of *Medical Systems*. 2020 Mar;44(3):1-8.
11. Burns EM, Rigby E, Mamidanna R, Bottle A, Aylin P, Ziprin P, Faiz OD. Systematic review of discharge coding accuracy. *Journal of Public Health*. 2012 Mar 1;34(1):138-48.
12. Wooldridge M. The road to conscious machines: The story of AI. *Penguin UK*; 2020 Mar 5.
13. Russell SJ, Norvig P. Artificial Intelligence: A Modern Approach. United Kingdom: *Pearson*, 2021.
14. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long Papers) 2018 Jun (pp. 1101-1111).
15. Gaebel W, Stricker J, Kerst A. Changes from ICD-10 to ICD-11 and future directions in psychiatric classification. *Dialogues in Clinical Neuroscience*. 2020 Mar;22(1):7.
16. Bengio Y, Lecun Y, Hinton G. Deep learning for AI. Communications of the ACM. 2021 Jun 21;64(7):58-65.
17. Dinwoodie HP, Howell RW. Automatic disease coding: the 'fruit-machine' method in general practice. *British Journal of Preventive & Social Medicine*. 1973 Feb;27(1):59.
18. Shi H, Xie P, Hu Z, Zhang M, Xing EP. Towards automated ICD coding using deep learning. *arXiv preprint* arXiv:1711.04075. 2017 Nov 11.
19. 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* 2017 Aug (pp. 328-332).
20. Nam J, Kim J, Loza Mencía E, Gurevych I, Fürnkranz J. Large-scale multi-label text classification—revisiting neural networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 2014 Sep 14 (pp. 437-452). Springer, Berlin, Heidelberg.
21. Kraljevic Z, Searle T, Shek A, Roguski L, Noor K, Bean D, Mascio A, Zhu L, Folarin AA, Roberts A, Bendayan R. Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit. *Artificial Intelligence in Medicine*. 2021 Jul 1;117:102083.
22. Wiegreffe S, Choi E, Yan S, Sun J, Eisenstein J. Clinical Concept Extraction for Document-Level Coding. In *Proceedings of the 18th BioNLP Workshop and Shared Task* 2019 Aug (pp. 261-272).
23. Kukafka R, Bales ME, Burkhardt A, Friedman C. Human and automated coding of rehabilitation discharge summaries according to the International Classification of Functioning, Disability, and Health. *Journal of the American Medical Informatics Association*. 2006 Sep 1;13(5):508-15.
24. Ji S, Hölttä M, Marttinen P. Does the magic of BERT apply to medical code assignment? A quantitative study. *Computers in Biology and Medicine*. 2021 Dec 1;139:104998.
25. [preprint]Sun W, Ji S, Cambria E, Marttinen P. Multitask Balanced and Recalibrated Network for Medical Code Prediction. *arXiv preprint* arXiv:2109.02418. 2021 Sep 6.
26. Falis M, Pajak M, Lisowska A, Schrempf P, Deckers L, Mikhael S, Tsaftaris S, O'Neil A. Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)* 2019 Nov (pp. 168-177).
27. Rios A, Kavuluru R. Few-Shot and Zero-Shot Multi-Label Learning for Structured Label Spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing 2018* (pp. 3132-3142).
28. Chalkidis I, Fergadiotis M, Kotitsas S, Malakasiotis P, Aletras N, Androutsopoulos I. An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* 2020 Nov (pp. 7503-7515).
29. Wang R, Long S, Dai X, Huang S, Chen J. Meta-LMTC: Meta-Learning for Large-Scale Multi-Label Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* 2021 Nov (pp. 8633-8646).
30. Xu K, Lam M, Pang J, Gao X, Band C, Mathur P, Papay F, Khanna AK, Cywinski JB, Maheshwari K, Xie P. Multimodal machine learning for automated ICD coding. In *Machine Learning for Healthcare Conference* 2019 Oct 28 (pp. 197-215). PMLR.
31. Liu Y, Cheng H, Klopfer R, Gormley MR, Schaaf T. Effective Convolutional Attention Network for Multi-label Clinical Document Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* 2021 Nov (pp. 5941-5953).
32. Kim BH, Ganapathi V. Read, Attend, and Code: Pushing the Limits of Medical Codes Prediction from Clinical Notes by Machines. In *Machine Learning for Healthcare Conference* 2021 Oct 21 (pp. 196-208). PMLR.

33. Yuan Z, Tan C, Huang S. Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding. *arXiv preprint* arXiv:2203.01515. 2022 Mar 3.
34. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* 2019 Jun (pp. 4171-4186).
35. Feucht M, Wu Z, Althammer S, Tresp V. Description-based Label Attention Classifier for Explainable ICD-9 Classification. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)* 2021 Nov (pp. 62-66).
36. Yogarajan V, Pfahringer B, Smith T, Montiel J. Improving Predictions of Tail-end Labels using Concatenated BioMed-Transformers for Long Medical Documents. *arXiv preprint* arXiv:2112.01718. 2021 Dec 3.
37. Gao S, Alawad M, Young MT, Gounley J, Schaefferkoetter N, Yoon HJ, Wu XC, Durbin EB, Doherty J, Stroup A, Coyle L. Limitations of transformers on clinical text classification. *IEEE Journal of Biomedical and Health Informatics*. 2021 Feb 26;25(9):3596-607.
38. 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. *Journal of Biomedical Informatics*. 2021 Apr 1;116:103728.
39. Searle T, Ibrahim Z, Dobson R. Experimental Evaluation and Development of a Silver-Standard for the MIMIC-III Clinical Coding Dataset. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing* 2020 Jul (pp. 76-85).
40. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. Multi-label classification of patient notes: case study on ICD code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence* 2018 Jun 20.
41. Falis M, Dong H, Birch A, Alex B. CoPHE: A Count-Preserving Hierarchical Evaluation Metric in Large-Scale Multi-Label Text Classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* 2021 Nov (pp. 907-912).
42. Dong H, Suárez-Paniagua V, Zhang H, Wang M, Whitfield E, Wu H. 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)* 2021 Nov 1 (pp. 2294-2298). IEEE.
43. Wu H, Toti G, Morley KI, Ibrahim ZM, Folarin A, Jackson R, Kartoglu I, Agrawal A, Stringer C, Gale D, Gorrell G. *SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research*. Journal of the American Medical Informatics Association. 2018 May;25(5):530-7.