Leveraging large-language models for medical code list creation: An example using the Charlson Comorbidity Index

RWD118

A Ochs[1], M R Sandu[1], M Duxbury[1], B Behara[2], C Henderson[1], M Danese[3], G Kafatos[1]

(1) Amgen Ltd, Center for Observational Research, Uxbridge, United Kingdom (2) Amgen GmbH, Center for Observational Research, Munich, Germany (3) Outcomes Insights Inc, Calabasas, United States

CONCLUSIONS

These early results indicate that LLMs may be useful in identifying and excluding large numbers of irrelevant codes with very low risk of false negatives. This makes a manual code list review to improve accuracy much more tractable.

Additional work is ongoing to further improve accuracy of the LLM. Efforts include prompt engineering and comparisons of performance of different LLMs. Further work is ongoing to assess LLM performance in other vocabularies.

INTRODUCTION

- Accurate medical code lists are vital for studies using real-world data. Creating code lists in a consistent, transparent and reproducible way is a challenge for researchers[1].
- Comprehensive code lists creation requires reviewing each code description of the full medical vocabulary. This is extremely time-consuming and resource intensive, particularly due to the high number of irrelevant codes that need to be reviewed.
- Large-language models (LLMs) have been shown to be successful in medical reasoning, assisting clinical decision making[2] and to accurately derive diagnosis codes from medical notes[3, 4].
- Here we explore if LLMs can be used to support in code list generation.

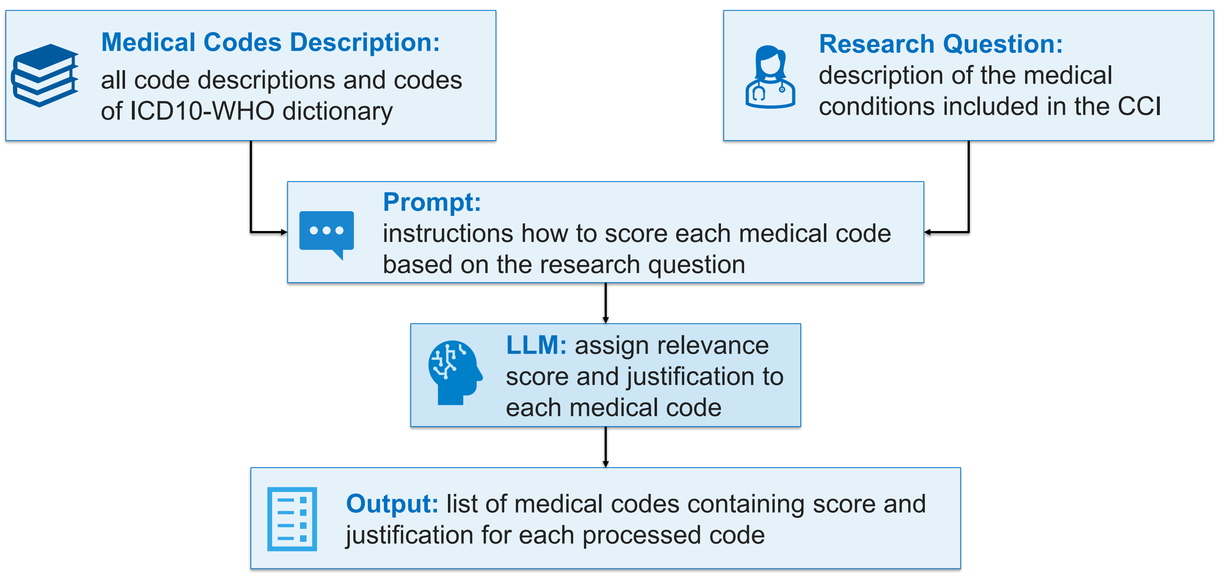
OBJECTIVES

- Evaluate the accuracy of LLM-generated code lists by comparing against those published for conditions of the Charlson Comorbidity Index (CCI).

METHODS

- We used the validated code lists of the CCI's 17 conditions as published by Quan et al[5] as reference standard against which to evaluate the LLM performance.
- For each condition, we prompted the LLM (ChatGPT, o3 mini) to assign a score and brief justification to each of the 16,287 medical code description of the ICD-10-WHO code dictionary (Figure 1).
- The score indicates how relevant a medical code is to the condition of interest and ranged from 0 indicating not relevant at all to 100 indicating an exact match.
- This full list of scored medical codes can be used to generate code lists by setting relevance score thresholds.

Figure 1: LLM code list scoring process flow



- We used performance metrics of sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) to assess LLM ability to classify codes for each of the CCI conditions.

RESULTS

- Figure 2 shows results overall and for each of the 17 CCI condition when setting the threshold to include all codes scored 1 or higher into the LLM-generated code list.

Figure 2: Classification metrics by condition of the Charlson Comorbidity Index

Condition	Sensitivity	Specificity	NPV	PPV
Overall	99.9%	93.5%	100%	6.1%
Congestive Heart Failure	100%	96.4%	100%	3.4%
Peripheral Vascular Disease	100%	93.0%	100%	2.3%
Cerebrovascular Disease	100%	94.4%	100%	9.4%
Dementia	100%	95.6%	100%	3.9%
Chronic Pulmonary Disease	98.3%	93.2%	100%	5.0%
Rheumatic Disease	100%	67.7%	100%	3.1%
Peptic Ulcer Disease	100%	98.7%	100%	16.1%
Mild Liver Disease	100%	95.8%	100%	5.5%
Diabetes w/o ch Complication	100%	98.7%	100%	10.6%
Diabetes w/ ch Complication	100%	95.2%	100%	3.1%
Hemiplegia or Paraplegia	100%	92.2%	100%	1.6%
Renal Disease	100%	92.0%	100%	1.2%
Any Malignancy	100%	95.5%	100%	42.4%
Moderate or Severe Liver Disease	100%	96.1%	100%	1.7%
Metastatic Solid Tumor	100%	92.9%	100%	2.8%
AIDS/ HIV	100%	95.1%	100%	3.2%
Myocardial Infarction	100%	97.0%	100%	2.6%

ch: chronic, NPV: negative predictive value, PPV positive predictive value w/: with, w/o: without

- >99.9% sensitivity across all conditions when excluding codes with relevance score of 0. The only false negative was observed for Chronic Pulmonary Disease (J66.2 Cannabinosis).
- High specificity (93.5%) and negative predictive value (>99.9%), show strong ability to exclude irrelevant codes from code lists.
- Positive predictive value was modest (6.1%) overall, with some variation between conditions, indicating that manual review remains essential to refine final lists.

REFERENCES

[1] Williams et al. (2019) PLOS One; 14(2).
[2] Chang et al. (2023) *Nat Med*; 29:1930-40.
[3] Klang et al. (2024) *medRxiv*.
[4] Cabral et al. (2024) *JAMA Int Med*; 184:581-583.
[5] Quan et al. (2005) Med Care; 43:1130-1139.

CONTACT INFORMATION

Andreas Ochs, PhD; Center for Observational Research, Amgen Ltd,
4 Sanderson Road, Uxbridge, UH8 1DH, United Kingdom. Email: aochs@amgen.com