

Interoperability of standardised electronic healthcare records facilitates transfer learning

Elizabeth Remfry¹[0000-0003-3940-8540] and Rafael Henkin¹[0000-0002-5511-5230]✉

Queen Mary University of London, London, United Kingdom
`{e.a.remfry,r.henkin}@qmul.ac.uk`

Abstract. Electronic healthcare records (EHR) use codes from different vocabularies to describe medical occurrences, often varying by type of care and country. Common data models (CDM) such as the Observational Medical Outcomes Partnership (OMOP) have been developed to enable the combination and comparison of heterogeneous datasets. We use the OMOP Standard Vocabularies to standardise two English EHR datasets and assess their interoperability through a BERT-based transformer model via pretraining and fine-tuning. Our results show the potential for standardisation to empower transfer learning, with trade-offs related to data loss.

Keywords: Electronic health records · Data Standardisation · Artificial Intelligence · Transfer learning.

1 Introduction

Electronic healthcare records (EHR) use codes from different vocabularies, often tailored to different countries' healthcare realities, to describe diseases, medications and other kinds of occurrences in clinical practice. Within the same country or even care system, there is considerable heterogeneity between coding systems used in primary and secondary care, for drug prescriptions and surgical procedures. To navigate these ontological challenges and standardise clinical data, common data models (CDM) such as the Observational Medical Outcomes Partnership (OMOP) have been developed [8]. CDMs include data structures and vocabularies to facilitate combining and comparing datasets to unlock health data analysis on a global level [7].

The standardisation of clinical ontologies has also opened doors for transfer learning, where knowledge gained from one dataset can be applied to another without requiring significant effort to align vocabularies or retrain models [4]. Previous work has successfully trained foundation models on OMOP CDM data from one hospital and applied these models to other out-of-domain datasets across different prediction tasks [2, 9, 6].

In this study, we map the medical codes from two UK datasets of routinely collected data from Clinical Practice Research Datalink (CPRD), GOLD and Aurum, to standard concepts in the OMOP Standardized Vocabularies (SV).

To assess interoperability, we pretrain a transformer model on CPRD Aurum, and fine-tune on both CPRD Aurum and GOLD. We report on the quality of the mapping, model performance and impact of transfer learning.

2 Methodology

2.1 Data

We used EHRs from two datasets, CPRD GOLD and Aurum, and include: individuals ≥ 16 years, registered with an English general practice between January 1, 2010, and December 31, 2020, eligible for linkage to secondary care data and diagnosed with at least 2 long-term conditions (LTC) defined by prevalidated codelists [1]. In Aurum, only general practices based in London were included. The primary care datasets were linked to the Hospital Episode Statistics (HES) for secondary care data, Office for National Statistics (ONS) mortality data and Index of Multiple Deprivation (IMD) data.

We identified individuals within these larger datasets with complex multimorbidity (CMM), defined as the diagnosis of at least three LTCs that impact three separate body systems [3] using the same code lists [1]. We extracted all data from the date of CMM, which formed our baseline, until deregistration, death or study end (31/12/2020). This resulted in 185,798 (13%) cases and 1,174,173 (87%) controls in CPRD GOLD and 57,328 (8%) cases and 659,691 (92%) controls in CPRD Aurum. This data was used for the pre-training and fine-tuning of the transformer model. See **Appendix** for more details.

2.2 Mapping pipeline

We took a pragmatic approach to mapping codes. First, we downloaded the standard set of vocabularies from the Athena library¹ along with additional vocabularies required for the CPRD datasets. We then mapped internal CPRD codes *medcodid* to external coding systems: SNOMED CT, OPCS, ICD10, Read and BNF, which are the *source* vocabularies. Next we mapped *source* codes, which are referred to as *concept codes* in OMOP, to *concept IDs* in the downloaded vocabularies. If codes are found in the SVs these can be classified as non-standard or standard. Unique *concept IDs* can map non-standard concepts to standard concepts using the OMOP Concept Relationship table. For example, 271649006 is a *concept code* from SNOMED, which is mapped to the standard *concept ID* 4152194. The use of standard concepts also enables consolidation of codes. For example, multiple SNOMED codes to record blood pressure can be mapped to a single blood pressure standard concept.

We assessed the total volume of data and calculated the total number of *concept codes* with no mapping to SVs, non-standard concepts and standard concepts. We report on mapping of the larger datasets, prior to the selection of our CMM cohort.

¹ <https://athena.ohdsi.org/>

2.3 Model

We pretrained and fine-tuned a BERT-based model from scratch on our clinical corpus consisting of standard *concept IDs* from our CMM cohort. We used *concept IDs* as they are unique. We built chronologically ordered sequences of concept IDs for each patient, e.g. 4152194 4334559 4086275, keeping only patients with sequences longer than 3 concepts within the study period.

We trained a domain-specific WordPiece tokenizer from scratch on the whole CMM cohort data from CPRD Aurum. The vocabulary size was set to 20,000 and only codes that appeared over 3 times in the dataset were kept, resulting in 57% coverage of the Aurum vocabulary. We pretrained a 107M parameter BERT model on a masked language modelling task, where 15% of tokens are randomly masked. The data was split 80/10/10 for train, test and validation datasets. All sequences were truncated from left, with a maximum context length of 512. We fine-tuned the pretrained models separately on Aurum and GOLD, on a downstream clinical prediction task: 5-year all-cause mortality in our CMM cohort. For fine-tuning, we kept only data from the date of CMM and up to 5 years before a death date for cases. For controls, we kept all data until end of study or deregistration.

We report area under the precision-recall curve (AUPRC) and area under the receiver operating characteristic curve (AUROC). All experiments were conducted on 4 A100 GPUs in the Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT [5].

3 Results

3.1 Concept Mapping

In both GOLD and Aurum, the majority of codes and rows in the data were successfully mapped to standard concepts. In GOLD the majority of concepts were mapped via other vocabularies, particularly as Read and ICD10 codes are not SVs in OMOP. In GOLD, the greatest loss of data was in prescription data, where 5.5% of the data could not be mapped to OMOP *concept IDs* (unmapped codes), and Read codes where 3.8% of the data was mapped to a non-standard *concept ID* which was not matched to a standard *concept ID*. In Aurum, the largest loss of data was again through unmapped SNOMED source codes (11.2%) and non-standard concepts (8.7%) in the primary care data. See **Appendix** for more details for the complete mapping results.

3.2 Model performance

We evaluated our pretrained model by fine-tuning on both same-database scenarios AURUM $\rightarrow$ AURUM and cross-database transfer scenarios AURUM $\rightarrow$ GOLD. Model performance was higher within the same-database, but demonstrated good performance between different datasets with a 0.10 difference between models (see table 1). The results may reflect the contents of the tokenizer

as we trained our custom tokenizer on *concept IDs* from CPRD Aurum only. The tokenizer covered 68% of the *concept IDs* in the Aurum fine-tuning dataset and only 37% in GOLD. We report on the AUROC and AUPRC plots in the **Appendix**.

Table 1. Performance metrics comparing Aurum and GOLD dataset using bootstrap resampling on the test set using 1000 iterations and a 95% confidence interval.

Dataset	AUPRC	AUROC
AURUM → AURUM	0.79 (0.78 – 0.80)	0.97 (0.97 – 0.97)
AURUM → GOLD	0.81 (0.81 – 0.82)	0.96 (0.95 – 0.96)

4 Conclusion and future work

Standardisation enabled us to reuse models and lower the time and computational expense required to train multiple models from scratch. Mapping to OMOP also helped reduce the size of the datasets, as multiple source codes (from SNOMED, Read, etc.) can map to one OMOP *concept ID*.

Standardisation is a step towards interoperability, but does not guarantee it as the SVs include coding systems from different countries. For example, the EHRSHOT model [9] was trained on SV-compatible data, but predominantly includes codes from the LOINC system for test measurements – these tend to be recorded as standard concepts in SNOMED in UK datasets.

Among the limitations of our work, our pragmatic download strategy of standardised vocabularies led to some codes from Aurum and GOLD not being mapped because the correct vocabulary was not downloaded. We also did not calculate or explore the amount of data lost per patient; future work should understand where and for whom data is lost in the conversion to OMOP. Future research should also assess the capacity for zero or few-shot learning in both the same and across datasets.

We present an initial exploration of the possibilities of interoperability between two EHR datasets that rely on different clinical ontologies. Whilst the mapping steps were largely successful, data loss was experienced across both datasets. Models performed better when fine-tuning on the same dataset, but there was evidence of some transfer learning across databases, demonstrating potential to enable the sharing of large healthcare models.

Acknowledgments. ER is funded by the Wellcome Trust Health Data in Practice (HDiP) Programme (218584/Z/19/Z). This work uses data provided by patients and collected by the NHS as part of their care and support. The data used in this study was under licence.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Eto, F., Samuel, M., Henkin, R., Mahesh, M., Ahmad, T., Angdembe, A., Hamish McAllister-Williams, R., Missier, P., J. Reynolds, N., R. Barnes, M., Hull, S., Finer, S., Mathur, R.: Ethnic differences in early onset multimorbidity and associations with health service use, long-term prescribing, years of life lost, and mortality: A cross-sectional study using clustering in the UK clinical practice research datalink **20**(10), e1004300. <https://doi.org/10.1371/journal.pmed.1004300>, <https://dx.plos.org/10.1371/journal.pmed.1004300>
2. Guo, L.L., Fries, J., Steinberg, E., Fleming, S.L., Morse, K., Aftandilian, C., Posada, J., Shah, N., Sung, L.: A multi-center study on the adaptability of a shared foundation model for electronic health records **7**(1), 1–9. <https://doi.org/10.1038/s41746-024-01166-w>, <https://www.nature.com/articles/s41746-024-01166-w>, publisher: Nature Publishing Group
3. Harrison, C., Britt, H., Miller, G., Henderson, J.: Examining different measures of multimorbidity, using a large prospective cross-sectional study in australian general practice **4**(7), e004694. <https://doi.org/10.1136/bmjopen-2013-004694>, <https://bmjopen.bmj.com/lookup/doi/10.1136/bmjopen-2013-004694>
4. Hur, K., Oh, J., Kim, J., Kim, J., Lee, M.J., Cho, E., Moon, S.E., Kim, Y.H., Choi, E.: UniHPF : Universal healthcare predictive framework with zero domain knowledge. <https://doi.org/10.48550/arXiv.2211.08082>, <http://arxiv.org/abs/2211.08082>
5. King, T., Butcher, S., Zalewski, L.: Apocrita - high performance computing cluster for queen mary university of london . <https://doi.org/10.5281/zenodo.438045>, <https://zenodo.org/records/438045>, publisher: Zenodo
6. Kirchler, M., Ferro, M., Lorenzini, V., Lippert, C., Ganna, A.: Large language models improve transferability of electronic health record-based predictions across countries and coding systems p. 2025.02.03.25321597. <https://doi.org/10.1101/2025.02.03.25321597>, <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11838679/>
7. Observational Health Data Sciences and Informatics: The book of OHDSI, <https://ohdsi.github.io/TheBookOfOhdsi/>
8. OHDSI: OMOP common data model, <https://ohdsi.github.io/CommonDataModel>
9. Wornow, M., Thapa, R., Steinberg, E., Fries, J.A., Shah, N.H.: EHRSHOT: An EHR benchmark for few-shot evaluation of foundation models. <https://doi.org/10.48550/arXiv.2307.02028>, <http://arxiv.org/abs/2307.02028>

Appendix

Data Processing

CPRD GOLD and Aurum contain four primary coding systems: primary care (Read or SNOMED), prescriptions (dm+d), hospitalisations (ICD10) and procedures (OPCS4). For the purpose of this study, we did not map the CPRD data structures to the OMOP CDM and limited our work to code standardisation. In CPRD GOLD, we combined the tables of clinical findings, referrals and test results to represent primary care data.

Before mapping datasets to OMOP CDM, we removed all rows without valid dates, those that were empty, occurred before birth or after death of the patient, or occurred after the end of the study.

BERT model

Pretraining was conducted using the AdamW optimizer, with a learning rate of 5e-5 and warm-up ratio of 0.1. We trained the entire train dataset of CPRD Aurum for 5 epochs with early stopping.

The model was fine tuned for 3 epochs with early stopping with a learning rate of 2e-5. We used weighted cross entropy due to class imbalance. To assess variations in model performance we employed a bootstrap resampling method on our test set using 1000 iterations and report a 95% confidence interval.

Mapping results

Tables 2 and 3 show the mapping results for the four source vocabularies in GOLD and the four source vocabularies in Aurum. Table 4 summarises the distributions of cases and controls between GOLD and Aurum, as well as characteristics of input sequences to the model. The Aurum dataset is slightly more imbalanced than the GOLD dataset. Tables 5 and 6 show the top 30 concepts that, after standardisation, are common between cases and controls in GOLD and Aurum, respectively. Although many concepts have similar distributions, Aurum controls have an overwhelming amount of administrative codes including SMS messages and general labels such as "History of event".

Table 2. CPRD GOLD data mapping results. NS-MV: Non-standard concepts or missing vocabulary, S-D: standard concepts directly mapped via source vocabulary, S-V: standard concepts mapped via another vocabulary

	GOLD							
	Prim. Care		Presc.		Hosp.		Proc.	
	Read codes	%	dm+d	% ICD10	% ICD10	% OPCS4	%	%
Input	1,230.1M	—	716.3M	—	98.5M	—	35.0M	—
Unmapped	4.8M	0.4%	39.1M	5.5%	1.1M	1.1%	10.5K	0%
NS/MV	46.2M	3.8%	2.3M	0.3%	1.1M	1.2%	43.3K	0%
S-D	—	0.0%	24.8M	3.5%	—	0.0%	7.9M	22.5%
S-V	1,179.1M	95.9%	650.1M	90.8%	96.2M	97.7%	27.1M	77.3%
Mapped rows	1,179.1M	95.9%	674.9M	94.2%	96.2M	97.7%	34.9M	99.8%

Table 3. CPRD Aurum data mapping results. NS-MV: Non-standard concepts or missing vocabulary, S-D: standard concepts directly mapped via source vocabulary, S-V: standard concepts mapped via another vocabulary

	Aurum							
	Prim. Care		Presc.		Hosp.		Proc.	
	SNOMED	%	dm+d	%	ICD10	%	OPCS4	%
Input	1,311.3M	—	459.3M	—	64.6M	—	22.4M	—
Unmapped	146.2M	11.2%	—	0.0%	555.3K	0.9%	8.5K	0.04%
NS/MV	113.8M	8.7%	2.5M	0.6%	643.3K	1.0%	30.9K	0.1%
S-D	677.4M	51.7%	17.2M	3.8%	—	0.0%	4.5M	20.1%
S-V	373.9M	28.5%	429.4M	93.5%	63.4M	98.1%	17.9M	79.8%
Mapped rows	1,051.3M	80.2%	446.7M	97.3%	63.4M	98.1%	22.4M	99.8%

Table 4. Characteristics of model input data

	GOLD		Aurum	
	Cases	Controls	Cases	Controls
Number of patients	185,798 (14%)	1,174,173 (76%)	57,328 (7%)	659,961 (93%)
Avg. number of unique concepts per patient	172.97	141.22	160.67	140.17
Avg. number of events per patient	900.19	566.45	871.52	655.56

Table 5. Shared concepts between GOLD and Aurum cases datasets after mapping. Values show the number of concepts per patient in each dataset.

Concept name	Vocab.	GOLD	Aurum
aspirin 75 MG Oral Tablet	RxNorm	17.55	14.43
bendroflumethiazide 2.5 MG Oral Tablet	RxNorm	12.85	9.66
Creatinine measurement, serum	SNOMED	8.67	8.12
Serum potassium measurement	SNOMED	7.95	7.30
Serum sodium level	SNOMED	7.93	7.33
acetaminophen 500 MG Oral Tablet	RxNorm	7.38	7.11
simvastatin 40 MG Oral Tablet	RxNorm	7.41	6.86
Haemoglobin estimation	SNOMED	6.41	5.48
Albumin measurement, serum	SNOMED	6.15	5.71
Total white cell count	SNOMED	6.19	5.33
Platelet count	SNOMED	6.17	5.29
Serum alkaline phosphatase measurement	SNOMED	6.04	5.42
MCV - Mean corpuscular volume	SNOMED	6.03	5.15
Serum urea measurement	SNOMED	5.90	5.53
atenolol 50 MG Oral Tablet	RxNorm	6.21	4.21
Simvastatin 20 MG Delayed Release Oral Tablet	RxNorm Extension	5.48	6.55
Serum cholesterol studies	SNOMED	5.48	6.24
Neutrophil count	SNOMED	5.80	5.01
Haematocrit	SNOMED	5.79	4.95
Lymphocyte count	SNOMED	5.76	4.96
Red blood cell count	SNOMED	5.76	4.88
Monocyte count	SNOMED	5.73	4.96
Eosinophil count	SNOMED	5.69	4.94
furosemide 40 MG Oral Tablet	RxNorm	5.73	4.65
omeprazole 20 MG Delayed Release Oral Capsule	RxNorm	5.49	5.34
MCH - Mean corpuscular haemoglobin	SNOMED	5.56	5.00
Basophil count	SNOMED	5.33	4.47
Telephone encounter	SNOMED	5.86	1.99
metformin hydrochloride 500 MG Oral Tablet	RxNorm	4.43	6.05
Weight finding	SNOMED	6.27	0.01

Table 6. Shared concepts between GOLD and Aurum controls datasets after mapping. Values show the number of concepts per patient in each dataset.

Concept name	Vocab.	GOLD	Aurum
SMS (short message service) text message sent to patient	SNOMED	14.75	302.01
History of event	OMOP Extension	19.53	142.25
Creatinine measurement, serum	SNOMED	35.05	65.31
Serum potassium measurement	SNOMED	32.02	61.58
Serum sodium level	SNOMED	31.99	61.53
aspirin 75 MG Oral Tablet	RxNorm	31.94	51.38
Haemoglobin estimation	SNOMED	29.67	54.82
Total white cell count	SNOMED	29.11	54.55
Platelet count	SNOMED	28.91	53.81
Albumin measurement, serum	SNOMED	28.08	56.23
MCV - Mean corpuscular volume	SNOMED	28.65	53.09
Neutrophil count	SNOMED	28.06	52.71
Lymphocyte count	SNOMED	27.89	52.35
Monocyte count	SNOMED	27.86	52.18
Haematocrit	SNOMED	28.04	51.12
Eosinophil count	SNOMED	27.74	51.81
MCH - Mean corpuscular haemoglobin	SNOMED	26.72	52.34
Red blood cell count	SNOMED	27.51	49.40
bendroflumethiazide 2.5 MG Oral Tablet	RxNorm	30.03	37.21
omeprazole 20 MG Delayed Release Oral Capsule	RxNorm	26.67	47.42
Basophil count	SNOMED	26.08	47.84
Serum alkaline phosphatase measurement	SNOMED	25.06	50.99
Serum urea measurement	SNOMED	24.56	47.70
simvastatin 40 MG Oral Tablet	RxNorm	27.05	39.38
Telephone encounter	SNOMED	32.64	19.78
Serum alanine aminotransferase level	SNOMED	23.04	44.20
Serum cholesterol studies	SNOMED	19.78	48.33
MCHC - Mean corpuscular haemoglobin concentration	SNOMED	21.25	41.14
GFR (glomerular filtration rate) calculated by abbreviated Modification of Diet in Renal Disease Study Group calculation	SNOMED	18.41	48.65
Serum HDL cholesterol measurement	SNOMED	17.95	43.42

10 E. Remfry and R. Henkin

AUROC and AUPRC model performance

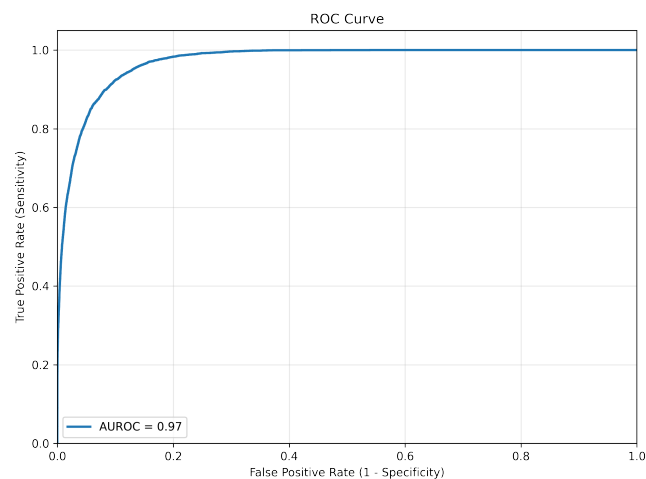


Fig. 1. AUROC for AURUM $\rightarrow$ AURUM

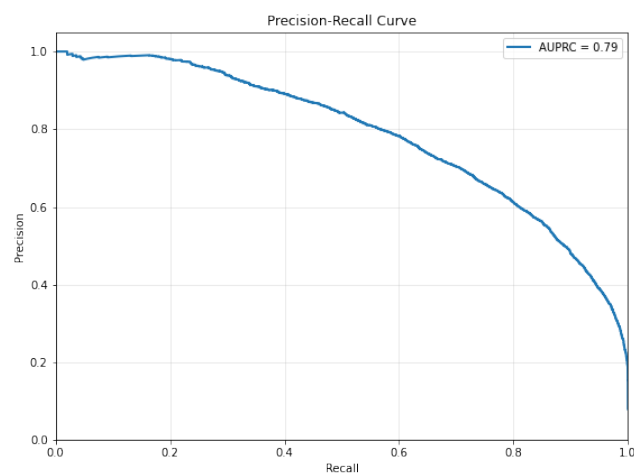


Fig. 2. AUPRC for AURUM $\rightarrow$ AURUM

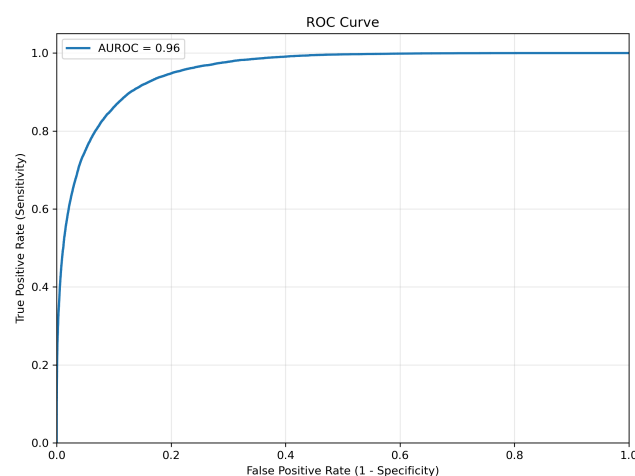


Fig. 3. AUROC for AURUM $\rightarrow$ GOLD

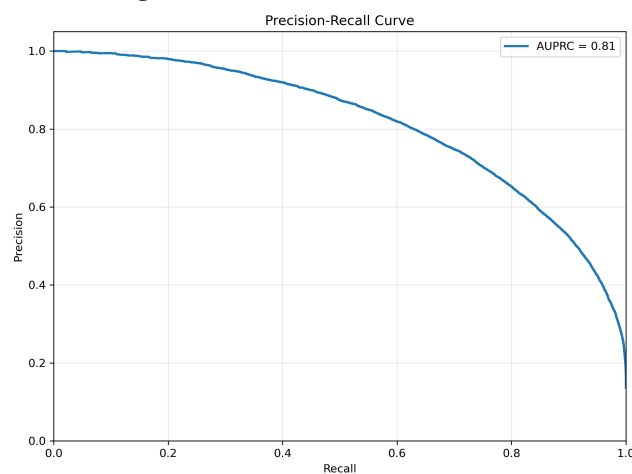


Fig. 4. AUPRC for AURUM $\rightarrow$ GOLD