Retrieval augmented text-to-SQL generation for epidemiological question answering using electronic health records

ngelo Ziletti*
Bayer AG
angelo.ziletti@bayer.com

Leonardo D' mbrosi
Bayer AG

bstract

Electronic health records (EHR) and claims data are rich sources of real-world data that reflect patient health status and healthcare utilization. Querying these databases to answer epidemiological questions is challenging due to the intricacy of medical terminology and the need for complex SQL queries. Here, we introduce an end-to-end methodology that combines text-to-SQL generation with retrieval augmented generation (RAG) to answer epidemiological questions using EHR and claims data. We show that our approach, which integrates a medical coding step into the text-to-SQL process, significantly improves the performance over simple prompting. Our findings indicate that although current language models are not yet sufficiently accurate for unsupervised use, RAG offers a promising direction for improving their capabilities, as shown in a realistic industry setting.

1 Introduction

Real-world data (RWD) are data routinely gathered from various sources that capture aspects of patient health status and the provision of health care. This encompasses electronic health records (EHR), medical claims data, disease registries, and emerging sources like digital health technologies. By investigating epidemiological quantities like patients' counts and demographics, disease incidence and prevalence, natural history of diseases, and treatment patterns in real-world clinical practice, researchers and healthcare organizations can identify for example target patient populations with unmet needs, discover unknown benefits of available drugs, evaluate potential for market entry, and estimate the potential enrolment of clinical trials.

Problem Statement. Addressing epidemiological questions using RWD databases is complex, as it requires not only an understanding of the data's characteristics, including biases, confounders, and limitations, but also involves interpreting medical terminology across various ontologies, formulating precise SQL queries, executing these queries, and accurately synthesizing the results.

Contributions. With this paper, we present a straightforward and effective end-to-end approach to answer epidemiological questions based on data queried from EHR/Claims databases.

- We release a dataset of manually annotated question-SQL pairs designed for epidemiological research, and adhering to the widely-adopted Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) (OMOP-CDM, 2023).
- We integrate a medical coding step into the text-to-SQL process, enhancing data retrieval and clinical context comprehension.
- We show that retrieval augmented generation (RAG) significantly improves performance compared with static instruction prompting, as confirmed by extensive benchmarking with top-tier large language models (LLMs).
- We share our dataset, code, and prompts[1] to foster reproducibility and catalyse a community-driven effort towards advancing this research area.

The presented approach is currently deployed at Bayer in experimental mode. Epidemiologists and data analysts are using the system to explore and evaluate its capabilities, ensuring that its use is carefully monitored and supervised.

2 The Dataset

Our dataset was created through a manual curation process, engaging specialists in epidemiological

[1] https://github.com/Bayer Group/text to sql epi ehr naacl2024

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 47–53

Quantity	Value
# of question/SQL pairs (all)	306
# of different tables used (all)	13
# of different columns used (all)	44
# logical conditions/query	6.4 (6.7)
# nesting levels/query	1.5 (1.1)
# tables/query	2.7 (0.9)
# columns/query	6.3 (4.7)
# medical entities/query	2.0 (4.1)
Question length [char]/query	91.7 (81.2)
SQL query length [char]/query	796.4 (448.5)

Table 1: Summary statistics of the dataset. For sample statistics, average and standard deviation (in brackets) are reported.

studies to contribute typical questions from their work. Despite its modest size, the dataset offers a realistic selection of epidemiological questions within industry practice, and exhibits a high degree of complexity. 53 samples require more than two level of nesting, and 19 more than three levels. Correctly answering questions often require multiple logical steps: selection of population(s) of interest, relationship between events within a specific time frame, aggregation statistics, and basic mathematical operations (e.g., ratios). The dataset features questions in their natural, free-form language and it is augmented with two paraphrased versions per question-SQL pair, increasing volume while also offering validated labels for retrieval algorithms. Statistics on the dataset are shown in Table 1. Due to budget limits, we will use one version per question for subsequent evaluations.

pplicability across RWD databases. To address the challenge of data retrieval variability across databases with differing data models, we leverage the OMOP-CDM. This model, underpinned by standardized vocabularies (Reich et al., 2024), harmonizes observational healthcare data and it is widely recognized as the standard for RWD analysis, with data from over 2.1 billion patient records across 34 countries (Voss et al., 2023; Reich et al., 2024).

3 Methods

Our methodology, outlined in Fig. 1, employs LLM prompting to translate natural language questions into SQL queries. It advances EHR text-to-SQL methods beyond the constraints of exact or string-based matching to fully encompass the semantic complexities of clinical terminology (Wang et al., 2020; Lee et al., 2022). To achieve this, we introduce a step where an LLM generates SQL with placeholders for medical entities (e.g., [condition@disphagia] in Fig. 1d), which are then mapped to precise clinical ontology terms (Sec. 4.1, Fig. 1d-e). This yields executable queries that accurately retrieve database information. Building on the success of RAG in enhancing LLMs for complex NLP tasks (Lewis et al., 2020), we use our dataset (Sec. 2) as an external knowledge base. Relevant question-SQL pairs are extracted and incorporated into the prompt, refining SQL generation. The completed SQL queries, embedded with medical codes, are run on an OMOP CDM-compliant database (Fig. 1f) to facilitate data retrieval. If needed, an answer can be articulated from the retrieved data through further LLM prompting (Fig. 1g).

4 Evaluation

4.1 Experimental setup

Large language models. We employ several leading LLMs as of February 2024: OpenAI's GPT-3.5 Turbo (Brown et al., 2020) and GPT-4 Turbo (OpenAI, 2023), Google's GeminiPro 1.0 (Gemini Team, 2023), Anthropic's Claude 2.1 (Anthropic AI, 2023), and Mistral AI's Mixtral 8x7B and Mixtral Medium (Mistral AI, 2023), with Mixtral 8x7B being the only open-source model (Jiang et al., 2024). We use one simple and one advanced prompt. The simple prompt provides essential instructions for creating queries that adhere to the conventions of the pipeline (Fig. 1). The advanced prompt adds detailed directives on concept IDs, race analysis, geographical analysis, date filters, column naming, patient count, age calculation, and additional instructions on SQL query validity review. Following Pourreza and Rafiei (2023), we allow LLMs up to three attempts to self-correct non-executable SQL queries using the compiler's error feedback.

Retrieval augmented generation. For similarity computation in RAG, we apply entity masking to substitute medical entities with generic labels (e.g., <DRUG>). We utilize the BGE-LARGE-EN-V1.5 embedding model from Hugging Face (Wolf et al., 2020), which has been fine-tuned for retrieval augmentation of LLMs (Zhang et al., 2023). We opt for masked question selection rather than utilizing the query because it eliminates the need for an initial LLM call to generate SQL for retrieval, while maintaining a comparable accuracy (Gao et al., 2023).

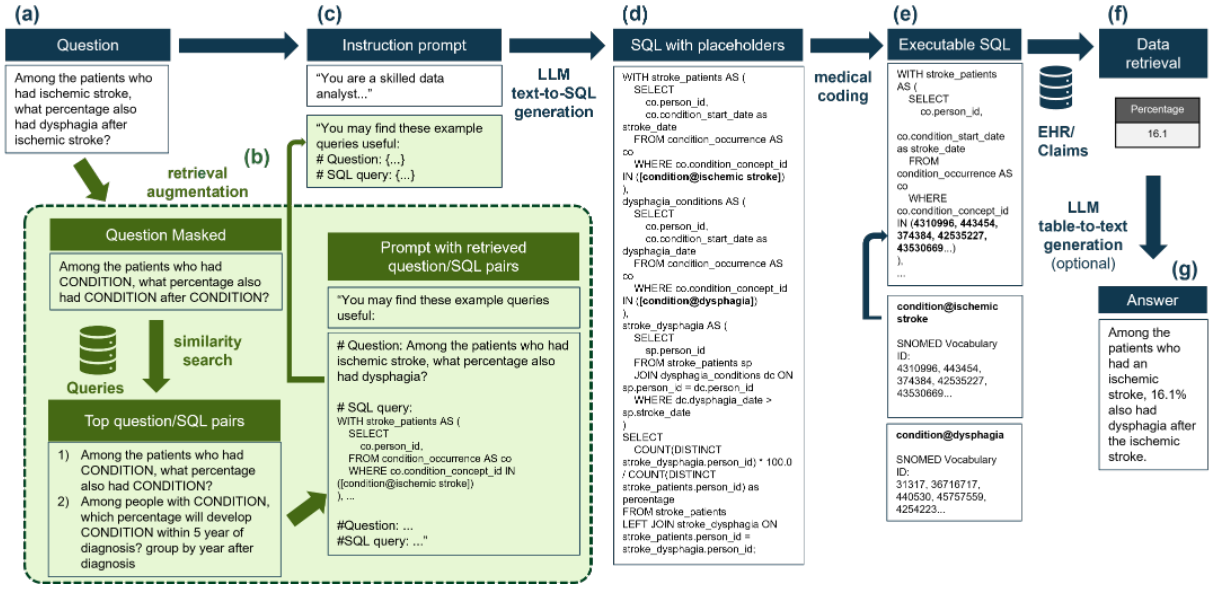

Figure 1: From a question in natural language to an answer in natural language using electronic health record or claims databases: end-to-end workflow.

Medical coding. LLMs extract medical entities and integrate them into SQL as placeholders (Fig. 1d), effectively recasting the medical coding task into medical entity normalization (Portelli et al., 2022; Ziletti et al., 2022; Zhang et al., 2022; Limsopatham and Collier, 2016). To perform entity normalization, we first compute the cosine similarity of each entity's SapBERT embeddings (Liu et al., 2021) with SNOMED ontology terms, and select the top-50 matches. Then, similarly to Yang et al. (2022), we prompt GPT-4 Turbo to verify whether a given code should be assigned to the input entity, refining the list.

Database and evaluation. The evaluation data reported are obtained querying the DE-SynPUF dataset (SynPUF, 2010), which is a synthetic dataset that emulates the structure of actual claims data. It includes 6.8 million beneficiary records, 112 million claims records, and 111 million prescription drug events records (Gonzales et al., 2023). The same analysis could be applied to any database conforming to the OMOP-CDM, thus potentially allowing access to 2.1 billion patient records (Reich et al., 2024). For evaluation, we manually developed a dataset of question-SQL pairs, as detailed in Sec. 2. These are then executed against the DE-SynPUF dataset, and the retrieved data from both reference and generated queries are compared to assess performance. This process reflects the practical use of SQL queries on healthcare databases. To ensure a realistic evaluation setup, the actual question being evaluated is removed from the RAG procedure. A generated query is marked as correct if it retrieves data enabling an answer that aligns with the reference query's answer (within a 10% tolerance), and incorrect otherwise. The tolerance compensates for variations from GPT-4 Turbo-based medical coding, maintaining the focus on text-to-SQL evaluation accuracy.

4.2 Experimental results

Results are shown in Table 2, and outlined below.

Enhanced performance with detailed prompting. Advanced prompting typically increases execution scores across models (except GPT-3.5 Turbo), but its impact on accuracy varies: Claude 2.1, Mistral-m, and GPT-4 Turbo show marked accuracy improvements with the advanced prompt, whereas Mixtral, GeminiPro, and GPT-3.5 Turbo see no such gains, suggesting that the additional details in the prompt may not benefit smaller or less sophisticated models. Overall performance is quite poor with either prompting methods.

Performance gains with contextual information. The inclusion of relevant examples via RAG significantly and consistently improves performance (Table 2, cf. RAG-top1/2/5 vs Prompt(advanced)). Notably, Mistral-m and GPT-4 Turbo exhibit marked improvements, suggesting they may possess a more advanced few-shot learning ability relative to the other models. Models outperform zero-shot

	Mixtral		GeminiPro		Claude 2.1		Mistral-m		GPT-3.5 Turbo		GPT-4 Turbo	
	Acc	Exec	Acc	Exec	Acc	Exec	Acc	Exec	Acc	Exec	Acc	Exec
Prompt (simple)	2.0	7.8	6.9	29.4	20.6	53.5	8.8	32.4	20.2	67.0	**28.4**	**77.5**
Prompt (advanced)	2.9	18.6	6.9	34.7	25.5	78.4	17.6	44.1	15.8	63.4	**38.2**	**91.2**
RAG-random1	19.6	46.1	11.8	35.3	33.3	76.5	38.2	68.3	29.0	84.0	**50.0**	**97.1**
RAG-top1	33.3	52.0	38.2	59.8	29.4	73.5	50.0	69.6	59.8	90.2	**72.5**	**97.1**
RAG-top2	20.6	40.2	37.3	56.9	38.6	75.2	46.1	73.5	61.8	94.1	**77.5**	**98.0**
RAG-top5	22.5	44.1	35.0	62.0	34.3	71.6	51.0	73.5	52.0	95.1	**77.5**	**97.1**
RAG-top1-oracle	52.0	62.7	67.6	73.5	58.8	83.3	56.9	74.5	**91.1**	**99.0**	82.8	95.0

Table 2: Comparative evaluation of LLMs' performance on text-to-SQL generation for epidemiological question answering. Accuracy (Acc) and executability (Exec) percentages are presented across different models and prompting conditions. Best results are in bold, while second best are underlined. RAG-top1/2/5 indicates the use of the top 1, 2, or 5 most similar questions to augment generation. RAG-random1 and RAG-top1-oracle scenarios provide models with a random dataset sample and the correct SQL query, respectively, for context.

prompting also when given a random dataset sample (RAG-random1), indicating that exposure to dataset structure and domain-specific language is helpful, even without query-specific context.

Diminishing returns with increased context. Providing a single example (RAG-top-1) leads to substantial improvements in performance, but adding more top results (RAG-top2 and RAG-top5) does not result in a similar increase. Some models exhibit a performance peak or a minor decline with additional context, indicating a limit to the beneficial amount of context.

Superiority of GPT-4 Turbo. GPT-4 Turbo is the best model overall by a large margin, followed by GPT-3.5 Turbo. Mistral-m outperforms both Claude 2.1 and GeminiPro. The open-source Mixtral model lags behind proprietary models in both accuracy and executability across all scenarios.

Model-specific approach to oracle context. In the RAG-top1-oracle scenario, where the prompt includes the correct SQL query, GPT-3.5 Turbo unexpectedly surpasses GPT-4 Turbo by closely mirroring the provided context, favouring direct replication. In contrast, GPT-4 Turbo and other models take a "deliberative" approach, often modifying the input, which, while useful for complex reasoning, hinders tasks that require exact copying.

5 Related Work

Text-to-SQL datasets for EHRs. The MIMIC-SQL dataset (Wang et al., 2020) comprises 10 000 template-generated questions for the MIMIC-III (Johnson et al., 2016) database. It contains both question designed to retrieve patient-specific information, and questions on patients counts with logical and basic mathematical operations. Tarbell et al. (2023) noted limited diversity in MIMIC-SQL's queries, possibly affecting its utility for testing text-to-SQL model generalizability. emrKBQA (Raghavan et al., 2021) contains 1 million patient-specific questions, also based on MIMIC-III. EHRSQL(Lee et al., 2022) is a dataset created by extracting templates from clinical questions posed by hospital staff, which are then used to generate a comprehensive set of queries for MIMIC-III and eICU (Pollard et al., 2018). It relies on an earlier, less performing text-to-text model for query generation (Raffel et al., 2020). All these datasets do not adhere with OMOP-CDM, and they opt for direct string matching for concept retrieval. The closest dataset to ours is the OMOP query library (OHDSI, 2019; OMOP-CDM-Query-Library, 2019), which is a collection of queries in OMOP-CDM. We adapted and included fifteen SQL queries from this library pertinent to epidemiological research into our dataset. Park et al. (2023) use rule-based methods and GPT-4 to translate clinical trial eligibility criteria into SQL queries for OMOP-CDM.

Text-to-SQL with LLMs and in-domain demonstrations. Prompting LLMs has proven effective, often outperforming specialized fine-tuned models in text-to-SQL task (Pourreza and Rafiei, 2023). Both in-domain (Chang and Fosler-Lussier, 2023a) and out-of-domain (Chang and Fosler-Lussier, 2023b) demonstrations improve LLMs' performance. Gao et al. (2023) explores retrieval scenarios for in-domain demonstration selection. To the best of our knowledge, the exploration of these text-to-SQL methods within EHR (or biomedical) research has not yet extended to small datasets that are critical for industry applications.

6 Conclusion

In this work, we presented the task of answering epidemiological questions using RWD. We demonstrated that RAG is effective in improving performance on all tested scenarios. Our study extends the demonstrated efficacy of RAG from general text-to-SQL benchmarks (Gao et al., 2023; Chang and Fosler-Lussier, 2023b) to include to small, domain-specific biomedical datasets, underlining its utility in data-scarce industry settings. The primary limitation is the dataset's limited size and specialized focus on epidemiological questions, suggesting further research should broaden its scope and scale.

7 cknowledgments

We would like to acknowledge Zuzanna Krawczyk-Borysiak for creating 27 text-SQL pairs; the remaining text-SQL pairs within the dataset were curated by A.Z. We also thank Jared Worful, Melanie Hackl, and Zuzanna Krawczyk-Borysiak for helpful discussions.

References

Anthropic AI. 2023. Model card and evaluations for claude models. https://www files.anthropic.com/production/images/Model Card Claude 2.pdf. Accessed: February 15, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *dvances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shuaichen Chang and Eric Fosler-Lussier. 2023a. How to prompt llms for text-to-sql: A study in zero-shot, single-domain, and cross-domain settings.

Shuaichen Chang and Eric Fosler-Lussier. 2023b. Selective demonstrations for cross-domain text-to-SQL. In *Findings of the ssociation for Computational Linguistics: EMNLP 2023*, pages 14174–14189, Singapore. Association for Computational Linguistics.

Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. 2023. Text-to-sql empowered by large language models: A benchmark evaluation.

Gemini Team. 2023. Gemini: A family of highly capable multimodal models. Technical report, Google. Accessed: February 15, 2024.

Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):1–16.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. Ehrsql: A practical text-to-sql benchmark for electronic health records. In *dvances in Neural Information Processing Systems*, volume 35, pages 15589–15601. Curran Associates, Inc.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rockt schel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th nnual Meeting of the ssociation for Computational Linguistics (Volume 1: Long Papers)*, pages 1014–1023, Berlin, Germany. Association for Computational Linguistics.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Mistral AI. 2023. Generative endopoints of mistral ai. https://docs.mistral.ai/platform/endpoints/. Accessed: February 15, 2024.

OHDSI. 2019. *The Book of OHDSI*, 1 edition. Observational Health Data Sciences and Informatics, Seoul, Korea. Accessed: 2021-03-30.

OMOP-CDM. 2023. Omop cdm common data model. https://ohdsi.github.io/CommonDataModel/. Accessed: January 23, 2024.

OMOP-CDM-Query-Library. 2019. Omop cdm query library. https://github.com/OHDSI/QueryLibrary. Accessed: January 23, 2024.

OpenAI. 2023. Gpt-4 technical report.

Jimyung Park, Yilu Fang, and Chunhua Weng. 2023. Criteria2Query 3.0 Powered by Generative Large Language Models. Observational Health Data Sciences and Informatics (OHDSI). https://www.ohdsi.org/wp content/uploads/2023/10/423 Park BriefReport.pdf.

Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5(1):180178.

Beatrice Portelli, Simone Scaboro, Enrico Santus, Hooman Sedghamiz, Emmanuele Chersoni, and Giuseppe Serra. 2022. Generalizing over long tail concepts for medical term normalization. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8580–8591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mohammadreza Pourreza and Davood Rafiei. 2023. DIN-SQL: Decomposed in-context learning of text-to-SQL with self-correction. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Preethi Raghavan, Jennifer J Liang, Diwakar Mahajan, Rachita Chandra, and Peter Szolovits. 2021. emrKBQA: A clinical knowledge-base question answering dataset. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 64–73, Online. Association for Computational Linguistics.

Christian Reich, Anna Ostropolets, Patrick Ryan, Peter Rijnbeek, Martijn Schuemie, Alexander Davydov, Dmitry Dymshyts, and George Hripcsak. 2024. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the merican Medical Informatics ssociation*, page ocad247.

SynPUF. 2010. Medicare claims synthetic public use files (synpufs). https://www.cms.gov/data research/statistics trends and reports/medicare claims synthetic public use files. Accessed: January 23, 2024.

Richard Tarbell, Kim-Kwang Raymond Choo, Glenn Dietrich, and Anthony Rios. 2023. Towards understanding the generalization of medical text-to-sql models and datasets. *MI ... nnual Symposium proceedings. MI Symposium*, 2023:669–678.

Erica A Voss, Clair Blacketer, Sebastiaan van Sandijk, Maxim Moinat, Michael Kallfelz, Michel van Speybroeck, Daniel Prieto-Alhambra, Martijn J Schuemie, and Peter R Rijnbeek. 2023. European health data and evidence network—learnings from building out a standardized international health data network. *Journal of the merican Medical Informatics ssociation*, 31(1):209–219.

Ping Wang, Tian Shi, and Chandan K. Reddy. 2020. Text-to-sql generation for question answering on electronic medical records. In *Proceedings of The Web Conference 2020*, WWW '20, page 350–361, New York, NY, USA. Association for Computing Machinery.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Zhichao Yang, Sunjae Kwon, Zonghai Yao, and Hongfeng Yu. 2022. Multi-label few-shot icd coding as autoregressive generation with prompt. *Proceedings of the I Conference on rtificial Intelligence. I Conference on rtificial Intelligence*, 37 4:5366–5374.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models.

Sheng Zhang, Hao Cheng, Shikhar Vashishth, Cliff Wong, Jinfeng Xiao, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Knowledge-rich self-supervision for biomedical entity linking. In *Findings of the ssociation for Computational Linguistics: EMNLP 2022*, pages 868–880, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Angelo Ziletti, Alan Akbik, Christoph Berns, Thomas Herold, Marion Legler, and Martina Viell. 2022. Medical coding with biomedical transformer ensembles and zero/few-shot learning. In *Proceedings of*

the 2022 Conference of the North merican Chapter of the ssociation for Computational Linguistics: Human Language Technologies: Industry Track, pages 176–187, Hybrid: Seattle, Washington + Online. Association for Computational Linguistics.