ClinQueryAgent: A Conversational Agent for Population Health Management

Joseph S. Boyle[1,2,3], **Anthony Dranfield**[3],
Mike O'Neil[3], **Maria Liakata**[2,5], **Alison Q. Smithard**[1,3,4]

[1]Canon Medical Research Europe [2]Queen Mary University of London
[3]Nottingham and Nottinghamshire ICB, NHS England
[4]University of Edinburgh [5]The Alan Turing Institute

Abstract

In this paper we introduce CLINQUERYAGENT, a system for translating natural language population health questions into executable database queries using agents with access to both local and external knowledge bases. Our novel architecture enables the use of powerful cloud-based language models whilst ensuring that no patient data leaves the secure environment. To combat inaccuracies over the course of longer dialogues due to context rot, information retrieval is delegated to a sub-agent. We deploy the system via a chat window embedded within an existing population health management platform where it has been used by 128 staff from 15 healthcare practices covering a total of 148,319 patients in the UK's National Health Service (NHS). We evaluate the system's capacity to autonomously handle a range of health informatics tasks on three datasets and via a beta-testing phase. Our results show that both analysts and clinicians are able to easily generate actionable information from patient health records using natural language requests requiring no programming expertise to verify. A public demo of the system is available to try[1]

1 Introduction

Healthcare data is produced at an astonishing rate. In 2014 there were 2.3B SNOMED CT codes recorded in English healthcare practices. By 2024, this number had grown almost threefold to 6.1B codes: approximately 100 codes per person each year. This data is stored in SQL warehouses which require programming skills to query, preventing most clinical staff from directly using this data to improve care delivery through analytics. `ClinQueryAgent` enables these care professionals to rapidly query healthcare databases in a transparent and secure way.

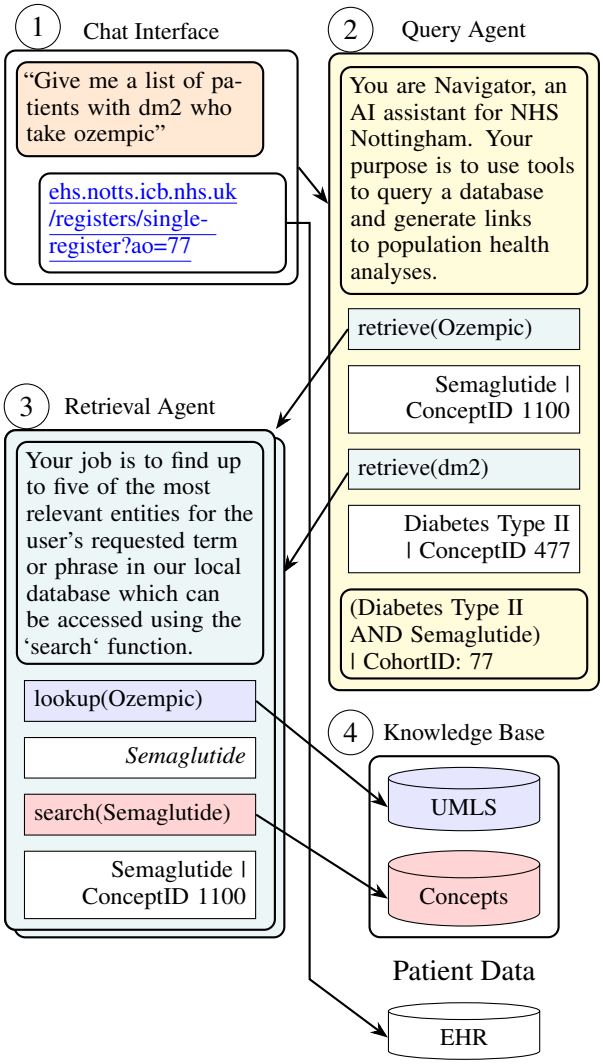


Figure 1: ClinQueryAgent parses natural language questions into database queries in an agentic loop. The Query Agent delegates the task of finding concepts to a Retrieval Agent while the Query Agent focuses on creating the database query. The Retrieval Agent has access to a local database of medical concepts, and the open-access ontology UMLS, which is used to help the retrieval agent parse unusual acronyms and synonyms. Diagram shows ClinQueryAgent's response to a user request about patients with type II diabetes mellitus ("dm2"). The resulting query is executed locally.

[1]https://clinqueryagent.josephsboyle.com/

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 197–207
July 2-7, 2026 ©2026 Association for Computational Linguistics

Workflow description	Real-world Example User Request
List each patient in a cohort (returns a list)	[General Practitioner] **"List the new hypertension diagnoses in the last 12 months"** *Intention: Clinical audit*
Summarise age and sex statistics across a patient cohort (returns the count, mean age, and sex)	[Pharmacist] **"number of patients on semaglutide"** *Intention: Check financial spend [Follow-up: Check if budget is targeting desired patient population e.g. high deprivation patients who cannot afford medication]*
Show distribution of a variable for a cohort (returns a histogram)	[Mental Health Commissioning Manager] **"Can you find out the weights of the SMI QOF register patients excluding remissions?"** *Intention: Profile population of patients with mental illness to understand determinants of health e.g. obesity*
Show the prevalence of a concept across practices / areas / districts Optionally, compute the prevalence within a subpopulation, e.g. those with lung disease (returns a percentage)	[Analyst] **"smoking statistics by district"** *Intention: Assess regional variation, plan smoking cessation services*
Show a To-Do list of patients for whom an intervention is required (returns a list)	[General Practice Manager] **"aged 75 or older with 15+ medications without a structured medication review"** *Intention: Schedule medication reviews*

Table 1: Five categories (workflows) of data analysis that may be handled and displayed by eHealthScope, with corresponding real-world user request examples. The agent implicitly classifies the user request into a workflow description, which it typically describes at the beginning of its answer process: *'To answer this question I will first search for hypertension and then create a cohort from those patients, and then link you to a register of those patients'*. 'SMI QOF register' refers to a specific concept for patients with one or more Serious Mental Illness.

To answer population-level questions, it is necessary to identify the patient cohort(s) of interest, in a similar manner to identifying a target cohort in a research study (Duke et al., 2018). In Figure 1, the required cohort comprises all patients with diabetes mellitus type II who are prescribed Ozempic. Patient cohorts can be identified by defining one or more "concepts" (sets of medical codes) and writing rules to describe eligible assignment patterns in the patient records (usually boolean logic, sometimes in combination with time constraints), then selecting all patients in which this assignment pattern is present. Many open source concept/cohort definition libraries exist such as Thayer et al. (2024). In this case, NHS Nottingham and Nottinghamshire ICB has defined 16,452 local concepts tailored to their local data format and analytical requirements; the above cohort could be identified by finding patients with the local concepts Diabetes II *and* Semaglutide concepts (see Appendix A).

In this work, we introduce CLINQUERYAGENT – an agent-based system for querying clinical databases through a task-oriented dialogue with a clinical user using generative language models (Bengio et al., 2003; Sutskever et al., 2014; Radford et al., 2018). Our system automates the creation of the population health queries, based on the observation that queries are easy to verify but labour intensive to produce. Given a natural language request, the system must express the user's goal as a logical query in terms of clinical concepts. Unlike SQL, a logical representation is easily understood by clinicians, empowering them to audit and co-author the query with the agent. This is subsequently compiled into an executable SQL query and run on patient data by the agent's host system.

We make the following contributions:

1. We demonstrate that conversational agents can provide fast and accurate natural language interfaces to clinical databases by evaluating ClinQueryAgent in a real world system used for population health management (here NHS Nottingham and Nottinghamshire ICB's population health analysis platform eHealthScope).
2. Our system design is privacy-preserving since the agent is isolated from the patient data, showing how to securely deploy systems in environments where the underlying data is sensitive whilst taking advantage of frontier generative models deployed in the cloud.
3. We observe accuracy degrades and cost in-

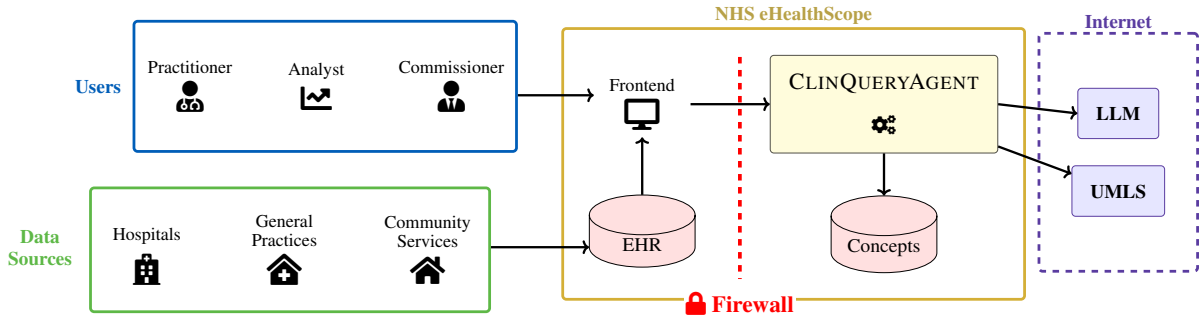


Figure 2: ClinQueryAgent (CQA) is situated within the eHealthScope application which serves a variety of users via the secure NHS intranet. Patient data is aggregated from healthcare centres into a central data warehouse (EHR) which is inaccessible to CQA by design, represented by the firewall between them.

creases during multi-turn conversations due to 'context rot' from large numbers of database results, and propose to reduce the main agent's context via delegation of database search to a subordinate agent.

4. We show that providing access to a knowledge base improves accuracy for queries involving specialist medical terminology.

2 Related Work

Our proposed system sits at the intersection of two prevailing research topics: task-oriented dialogue and text-to-query systems.

Task-Oriented Dialogue In task-oriented dialogues (TODs), an agent and user collaborate to enact a user's goal (Yi et al., 2025).

LLM agents are augmented with tools which enable them to take actions within an environment e.g. retrieving information from Wikipedia in order to obtain an answer. An early example was SimpleTOD, a GPT-2 based agent which could perform all aspects of task completion including database queries (Hosseini-Asl et al., 2020). Subsequent systems have continued to function via autoregressive language modelling, interleaving reasoning traces with task-specific actions. The so-called 'ReAct' paradigm introduced by Yao et al. (2023) added task delegation to external tools by issuing *tool calls*. Within this paradigm, coding agents like Claude Code, Gemini-CLI, and OpenAI Codex are capable of autonomously performing complex programming tasks (Kwa et al., 2025).

Medical Text-to-Query is the task of automatically mapping a clinical question to an executable query. Early medical text-to-query approaches used bespoke grammars to perform this task (Woodyard and Hamel, 1981) and contemporary approaches use neural language models. Recent applications include: querying synthetic records (Lee et al., 2022; Sivasubramaniam et al., 2024; Jiang et al., 2025); generating epidemiological analyses in real-world databases (Ziletti and D'Ambrosi, 2025; Möller-Grell et al., 2025); and drafting implementations of inclusion criteria for medical studies (Yuan et al., 2019; Park et al., 2024). These approaches leverage techniques like in-context learning and vector-based retrieval augmented generation (Brown et al., 2020; Lewis et al., 2021) to improve accuracy.

3 Method

ClinQueryAgent is an autonomous task-oriented dialogue LLM agent. To enact a user's request the agent must: 1) understand their intent; 2) retrieve the relevant entities from a database of ∼16,500 local concepts; 3) combine these concepts into a logical query that the user can execute. This section describes the prompt design process, tool use capabilities, and sub-agent delegation mechanism.

3.1 Prompt design

To guide our agent on how to answer different types of questions, we provide templates in the ClinQueryAgent's system prompt to demonstrate which tools to call in order to handle a user's request. This paradigm is inspired by in-context learning which is a standard method to improve the task-specific performance of LLMs, by providing training examples in the prompt. However, unlike in-context learning examples, templates do not reference specific entities, which avoids biasing the generation process towards including these entities (in our case concepts) even when they are irrelevant to the specific task at hand (Schulhoff et al., 2024).

In this case, our agent is serving the eHealthScope application and so we create templates for five categories ("workflows") of data analysis that may be handled by eHealthScope (see Table 1).

3.2 Tools

ClinQueryAgent is provided with tools to perform:

1. **lexical concept search** within the NHS database; this tool fetches concepts which contain (sub)strings specified by the agent
2. **patient cohort creation** from query logic; this tool does not return any value to the agent, but converts the query logic provided by the agent to executable SQL, then runs the SQL and caches the resulting patient cohort in the eHealthScope database
3. **url creation** to results display; this tool returns a url which the user may click to see the specified patient cohort displayed in the eHealthScope UI

3.3 Accessing external knowledge

To help the agent retrieve concepts not well defined within its pre-trained knowledge, there is the option to provide additional tools to retrieve information from knowledge bases. We experiment with providing access to the OHDSI standardised vocabularies, containing over 10 million codes (Reich et al., 2024), via a **lookup** tool. The vocabularies contain a rich mapping of chemical compounds to drug names, which enables the agent to link e.g. 'ozempic' to 'semaglutide', and extend data from the Unified Medical Language System (UMLS).

3.4 Sub-Agent delegation of information retrieval

Identifying appropriate medical concepts in a real-world clinical database is a challenging task, involving elements of semantic parsing and disambiguation, handling of polysemy, synonymy, metonymy, and abbreviations, as well as local knowledge on naming conventions. The lexical concept search tool may return a large number of concepts ('diabetes' returns 407 results, for instance, since there are many variant concepts relating to diabetes) which the agent must parse to determine which concepts are relevant to the user's query. Over the course of a multi-turn conversation the number of concepts in the context grows and a degradation in query accuracy is observed. We attribute this degradation to the long bloated context, as noted in prior work (Liu et al., 2024; Laban et al., 2025) and colloquially referred to as 'context rot'.

To reduce irrelevant context, we experiment with using a sub-agent to perform the search process. In this configuration (see Fig. 3), the primary agent

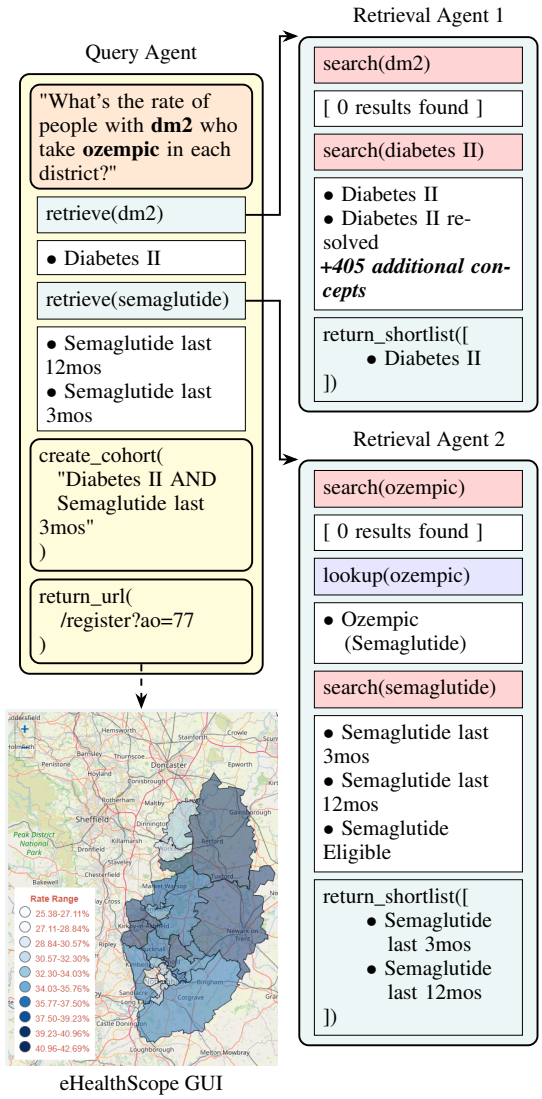


Figure 3: An example of computing statistics across regions and visualising them in the eHealthScope GUI. Linking to the most appropriate concept can be a complex multi-step task involving lookups, and filtering large quantities of results: for example the diabetes II search returns 400+ concepts. The entity linking task is delegated to sub-agents, which are created with a fresh context for each `retrieve(...)` call. The QueryAgent determines which of the retrieved concepts to include in the query; here, patients with 'Semaglutide last 3 mos" have the most recent evidence of taking semaglutide.

(Query Agent) instantiates a new sub-agent (Retrieval Agent) via a delegation tool, prompting it with the search term. The retrieval agent performs its own iteration loop, exploring the concept database, eventually returning a small number of concepts and explaining their relevance, e.g. *"Returned the concept for 'Semaglutide' as it is the generic form of the requested term 'Ozempic' "*.

The sub-agent has two tools: **search** for searching for concepts, and **return_shortlist**, for return-

ing a small number of concepts and an explanation as to their relevance to the user's question. Calling **return_shortlist** ends the sub-agent's routine, and Query Agent continues with the newly retrieved information.

Use of a sub-agent keeps the Query Agent's context window shorter, allowing it to reason over a smaller and more relevant set of concepts. Additionally, each concept search triggers the instantiation of a new sub-agent, allowing each sub-agent's context to remain focussed on a single task.

4 Experimental Setup

Model We run a real world evaluation in a dialogue based deployment of ClinQueryAgent. Our framework is agnostic to the choice of LLM. We choose to evaluate using Gemini 2.5 Flash (Comanici et al., 2025), based on its frontier price-performance characteristics in agentic benchmarks (Kapoor et al., 2025). The 'thinking' token budget is configured to 0, as we found it to slow down response times substantially without improving quality. Greedy decoding is used (temperature=0).

Datasets First, we adapted a benchmark of epidemiological cohort questions from Ziletti and D'Ambrosi (2025)'s *EpiCoho* dataset to the context of our application, resulting in 39 questions. Second, to test the agent's ability to leverage an external knowledge base, we created a synthetic dataset of 100 questions in the form "List patients who take X" where X is a brand name sampled at random from the UK medicines agency[2]. These questions require matching of the brand name to its generic name concept e.g. Ozempic to semaglutide. Third, we manually reviewed each user request from our real-world beta-deployment and constructed a dataset comprising 82 questions.

In sum there were 39 + 100 + 82 = 221 unique test samples across the three datasets. Appendix B gives further details about dataset creation.

Tasks We evaluated the agent's ability to answer a) Single Questions, where one agent is asked a single question, and b) Chained Questions, where all questions are asked sequentially within one chat.

Metrics The same question may be correctly answered with different queries (sometimes using different concepts), thus we do not evaluate the generated query verbatim against the reference query but assess similarity as follows: *recall* is the proportion of the patients in the reference cohort who are in the retrieved cohort, *precision* is the proportion of patients in the retrieved cohort who are in the reference cohort, and *F1-score* is their harmonic mean. We further characterise performance using the number of *tokens* consumed, trajectory *time*, and *tool calls*. Each of these metrics is computed per trajectory and averaged.

[2] www.gov.uk/government/publications/category-lists-following-implementation-of-the-windsor-framework

5 Results

Table 2 shows performance on the three datasets. We compare the performance of a naive single agent versus extension with the retrieval agent, with or without an external UMLS resource.

All configurations achieve similar F1-scores in the single question setup for the EpiCoho-M and Real World datasets. Conversely in the chained setup a large degradation in accuracy is observed when QueryAgent is used without a retrieval agent, with the mean F1-score dropping from 0.64 to 0.43 on EpiCoho-M and from 0.66 to 0.12 on the Real World dataset. Using the Retrieval Agent reduces the number of tokens consumed from 855,427 to 56,277 (↓93%). Surprisingly, F1-score on the Real World dataset improves from 0.68 on single questions to 0.75 on chained questions, suggesting that the context of previously answered questions may improve accuracy.

Despite the accuracy benefits of sub-agent use, in many configurations the answer took longer to generate. In the UK medicine brand names dataset, the UMLS agent took twice as long to construct queries (10.79s vs 4.68s) but achieved a superior average F1-score: 0.85 vs 0.26. Inspection showed that this additional latency was a result of the Retrieval Agent + UMLS making multiple knowledge base lookups in distinct tool-call steps. Instructing the agent to perform tool calls in parallel rather than in series would likely reduce latency substantially.

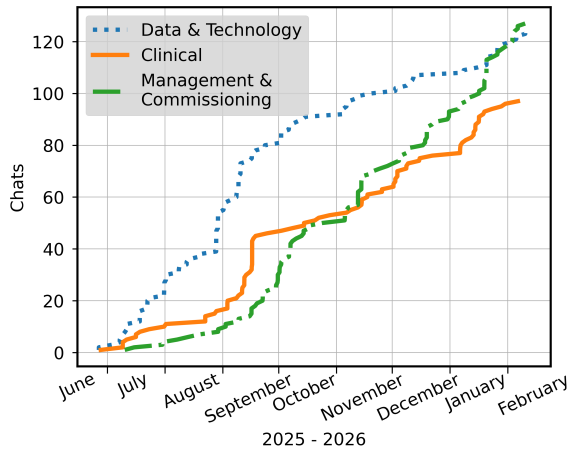
5.1 Real world usage

Figure 4 shows use of the real-world system during the beta-testing phase, with steady uptake during this period. The tool is most popular with technology staff which we ascribe to its use by population health analysts, whose core work involves querying population-level digital data. Overall, there were 128 unique users who engaged in 354 chats. User feedback is analysed in Appendix C.

		Patient Overlap			Mean Per Question		
Type	Method	R	P	F1	Tokens	Time (s)	Tools
EpiCoho-M Dataset							
Single	QueryAgent	0.72	**0.68**	0.64	**35,299**	**15.14**	3.39
	+ Retrieval Agent	0.76	0.63	0.63	38,693	23.18	**3.36**
	+ Retrieval Agent + UMLS	**0.80**	**0.68**	**0.68**	38,509	18.97	3.47
Chained	QueryAgent	0.45	0.53	0.43	855,427	64.82	**2.13**
	+ Retrieval Agent	0.70	**0.72**	0.61	56,277	**19.53**	10.17
	+ Retrieval Agent + UMLS	**0.73**	0.71	**0.65**	**42,438**	23.72	9.91
UK Medicine Brands Dataset							
Single	QueryAgent	0.31	0.25	0.26	**10,239**	**4.68**	**1.91**
	+ Retrieval Agent	0.58	0.50	0.51	18,696	13.44	2.94
	+ Retrieval Agent + UMLS	**0.92**	**0.85**	**0.85**	14,415	10.79	2.95
Real World Dataset							
Single	QueryAgent	0.71	**0.71**	0.66	37,096	**9.92**	2.94
	+ Retrieval Agent	**0.74**	**0.71**	**0.68**	**25,794**	13.86	2.89
	+ Retrieval Agent + UMLS	**0.74**	**0.71**	**0.68**	28,758	13.98	**2.85**
Chained	QueryAgent	0.12	0.14	0.12	181,568	**4.07**	**2.11**
	+ Retrieval Agent	**0.76**	0.8	0.73	**46,821**	12.99	5.80
	+ Retrieval Agent + UMLS	**0.76**	**0.85**	**0.75**	58,031	15.51	6.30

Table 2: Performance metrics on the *EpiCoho-M* epidemiological cohort questions, the Real World beta-test questions and the UK brands datasets, with/without the concept retrieval sub-agent, and with/without access to UMLS. We assess accuracy according to how closely the retrieved patient cohort matches the expected patient cohort in a real-world database, using recall (R), precision (P) and F1-score (F1). We also show the mean number of tokens used, time, and number of tool calls per question.

Figure 4: Cumulative message count for ClinQueryAgent in its beta deployment phase from 2025-06-26 – 2026-02-13. Clinical staff include GPs, care-coordinators, and pharmacists; Management and Commissioning includes operations staff, practice managers; Data & Technology staff include population health analysts, software developers, and data management staff.



6 Discussion

In their overview of Agentic AI in healthcare, Karunanayake (2025) identify privacy and explainability as key challenges for successful integration into real world clinical practice. We address these via *structural privacy* and *verifiable logic*.

Structural privacy

By strictly separating patient data (execution layer) from medical concepts and queries (reasoning layer), we achieve a *structural privacy* exceeding standard GDPR/HIPAA compliant cloud providers such as Microsoft Azure, simultaneously enabling the use of frontier models via the internet and protecting patient data. One drawback of this architecture is that the agent's tools must be crafted with the restriction of access to patient data in mind, which is more costly than the alternative of providing the agent with a raw database connection.

Verifiable Logic

The conversational interface to ClinQueryAgent allows a collaborative style of query creation in which the user helps to refine the draft and ask follow-up questions. By generating the queries in logical form, e.g. "Diabetes 2 AND Semaglutide", we enable users to check the generated query. For clinicians and population health managerial staff who are unfamiliar with SQL, this changes the question from 'Do I trust this LLM?' to 'Is this logic a correct implementation of my original request?'

7 Conclusion

We demonstrate ClinQueryAgent, a novel task-oriented dialogue system for performing population health analysis. Ablation experiments on three datasets (Table 2) found that deploying a sub-agent for the task of information retrieval helps the system to maintain coherence over the course of long conversations and that integrating external medical vocabulary information augments performance on queries requiring knowledge of medicine names.

ClinQueryAgent is integrated into NHS Nottingham and Nottinghamshire ICB's population health platform and, as of February 2026, has been used by 128 clinicians from 15 healthcare practices covering a total of 148,319 patients.

End-to-end query generation in a real-world database of 16,500 clinical concepts takes less than 15 seconds on average, and usage statistics suggest that most 'first-draft' queries acceptably operationalise the user's intent without the need for multi-turn refinement. We propose this design as an example for future AI-enabled text-to-query systems where data protection and interpretability are priorities.

8 Limitations

The system was implemented in a proprietary database used for real world clinical care and so it was necessary to modify the EpiCoho dataset, meaning that ClinQueryAgent is not directly comparable with other methods on this dataset. The system's multi-agent task decomposition; logical query language; and robust isolation of patient data can be readily applied to other systems. For example, the online demo accompanying this paper replaces both the database and UI layers with alternatives whilst maintaining the core system design principles listed above.

ClinQueryAgent has some limitations in terms of unknown concepts or knowledge. First, only pre-existing local concepts can be used meaning that some questions cannot be answered, e.g. those relating to a particularly novel disease for which there is no concept yet. Future work could tackle the generation of new concepts at inference time, enabling more varied and precise queries. Second, the agent lacks awareness of some local terms, for example a user asked where to find a 'diabetes quick' (a type of document), which the agent misunderstood. Future work could seek to integrate organisational document bases and wikis.

There are limitations to how reliably the system can answer questions like "Find me patients who should be on the End Of Life register". This question requires clinical judgement, which may or may not have already been codified into accessible concept(s).

Inspection of user feedback showed occasional requests for out-of-scope administrative tasks such as 'adding new users' (Appendix Table 5). We added support for these use cases through the addition of a new tool which would link the user to the documentation if it was asked this type of question.

9 Ethics Statement

This study incorporating a service evaluation uses anonymised, routinely collected data from National Health Service (NHS) Nottingham and Nottinghamshire services and does not require NHS Research Ethics Committee (REC) approval according to NHS Health Research Authority (HRA) guidelines[3]. The need for consent to participate was deemed unnecessary according to the aforementioned HRA guidelines.

JB and AS had honorary contracts with the NHS with de-identified access to the data in order to conduct this work, as part of the regular delivery of patient care under the remit of NHS Nottingham and Nottinghamshire Strategic Analysis and Intelligence Unit (SAIU).

Acknowledgments

We thank the NHS Nottingham and Nottinghamshire ICB SAIU for their expertise and contributions to this project, particularly to Ben Clarke for their contributions to the project and to Carl Davis for organising the collaboration. We thank Patrick Schrempf for reviewing this manuscript and the project's source code.

This work was supported by the Engineering and Physical Sciences Research Council [grant number EP/Y009800/1], through funding from Responsible AI UK (KP0016) as a Keystone project lead by Maria Liakata, and through funding for the Centre for Doctoral Training in Data-Centric Engineering [grant number EP/V519935/1].

[3] www.hra-decisiontools.org.uk/research/

References

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research*, 3(Feb):1137–1155.

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, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint*. ArXiv:2507.06261 [cs].

Jon Duke, Chris Knoll, and Nigam Shah. 2018. OHDSI Cohort Definition and Phenotyping.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A Simple Language Model for Task-Oriented Dialogue. In *Advances in Neural Information Processing Systems*, volume 33, pages 20179–20191. Curran Associates, Inc.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. 2025. MedAgentBench: A Realistic Virtual EHR Environment to Benchmark Medical LLM Agents. *arXiv preprint*. ArXiv:2501.14654 [cs].

Sayash Kapoor, Benedikt Stroebl, Peter Kirgis, Nitya Nadgir, Zachary S. Siegel, Boyi Wei, Tianci Xue, Ziru Chen, Felix Chen, Saiteja Utpala, Franck Ndzomga, Dheeraj Oruganty, Sophie Luskin, Kangheng Liu, Botao Yu, Amit Arora, Dongyoon Hahm, Harsh Trivedi, Huan Sun, and 12 others. 2025. Holistic Agent Leaderboard: The Missing Infrastructure for AI Agent Evaluation. *arXiv preprint*. ArXiv:2510.11977 [cs].

Nalan Karunanayake. 2025. Next-generation agentic AI for transforming healthcare. *Informatics and Health*, 2(2):73–83.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, and 6 others. 2025. Measuring AI Ability to Complete Long Tasks. *arXiv preprint*. ArXiv:2503.14499 [cs].

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. LLMs Get Lost In Multi-Turn Conversation. *arXiv preprint*. ArXiv:2505.06120 [cs].

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. *Advances in Neural Information Processing Systems*, 35:15589–15601.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *arXiv preprint*. ArXiv:2005.11401 [cs].

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173. Place: Cambridge, MA.

Niko Möller-Grell, Shihao Shenzhang, Zhangshu Joshua Jiang, and Richard Dobson. 2025. Agentic conversation on OMOP CDM: the OMCP-A2A foundation library. *OHDSI Global Symposium 2025*, 1.

Jimyung Park, Yilu Fang, Casey Ta, Gongbo Zhang, Betina Idnay, Fangyi Chen, David Feng, Rebecca Shyu, Emily R. Gordon, Matthew Spotnitz, and Chunhua Weng. 2024. Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation. *Journal of Biomedical Informatics*, 154:104649.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

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 American Medical Informatics Association*, 31(3):583–590.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and 12 others. 2024. The Prompt Report: A Systematic Survey of Prompting Techniques. *arXiv preprint*. ArXiv:2406.06608 [cs] version: 1.

Sithursan Sivasubramaniam, Cedric Osei-Akoto, Yi Zhang, Kurt Stockinger, and Jonathan Fuerst. 2024. SM3-Text-to-Query: Synthetic Multi-Model

Medical Text-to-Query Benchmark. *arXiv preprint*. ArXiv:2411.05521 [cs].

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

Daniel S Thayer, Shahzad Mumtaz, Muhammad A Elmessary, Ieuan Scanlon, Artur Zinnurov, Alex-Ioan Coldea, Jack Scanlon, Martin Chapman, Vasa Curcin, Ann John, Marcos DelPozo-Banos, Hannah Davies, Andreas Karwath, Georgios V Gkoutos, Natalie K Fitzpatrick, Jennifer K Quint, Susheel Varma, Chris Milner, Carla Oliveira, and 4 others. 2024. Creating a next-generation phenotype library: the health data research UK Phenotype Library. *JAMIA Open*, 7(2):ooae049.

Mary Woodyard and Baruch Hamel. 1981. A natural language interface to a clinical data base management system. *Computers and Biomedical Research*, 14(1):41–62.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. REAC T: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS.

Zihao Yi, Jiarui Ouyang, Zhe Xu, Yuwen Liu, Tianhao Liao, Haohao Luo, and Ying Shen. 2025. A Survey on Recent Advances in LLM-Based Multi-turn Dialogue Systems. *arXiv preprint*. ArXiv:2402.18013 [cs].

Chi Yuan, Patrick B Ryan, Casey Ta, Yixuan Guo, Ziran Li, Jill Hardin, Rupa Makadia, Peng Jin, Ning Shang, Tian Kang, and Chunhua Weng. 2019. Criteria2Query: a natural language interface to clinical databases for cohort definition. *Journal of the American Medical Informatics Association*, 26(4):294–305.

Angelo Ziletti and Leonardo D'Ambrosi. 2025. Generating patient cohorts from electronic health records using two-step retrieval-augmented text-to-SQL generation. *arXiv preprint*. ArXiv:2502.21107 [cs].

A Example Concept Definitions

Concept	Code ID	Code Description	Pseudocode
Diabetes type II	44054006 472969004 443694000 ...	Type II Diabetes History of diabetes type 2 Uncontrolled type 2 diabetes *+160 more*	`SELECT pat_id` `FROM Conditions` `WHERE code IN (` `  44054006, 472969004, ...)`
Semaglutide - last 3 months	777514008 782102009 ...	Semaglutide only product Semaglutide 1.34 mg/mL solution for injection *+10 more*	`SELECT pat_id` `FROM Medications` `WHERE code IN (777514008, ...)` `AND date ≥ DATE_SUB(NOW, 3M)`

Table 3: Example NHS Nottingham and Nottinghamshire ICB concept definitions. The concept codesets are described in columns 2 & 3, and the final "Pseudocode" column illustrates the rules to combine them; these are usually boolean logic, sometimes including time constraints. The corresponding patient cohort comprises all patients with this code assignment pattern present in their record.

B Dataset Creation

Dataset 1: EpiCoho-M The original EpiCohoKB dataset contains 108 questions designed for US data (Ziletti and D'Ambrosi, 2025). We required to make a number of adaptations as follows.

- To adapt these questions for our UK data, we replaced expressions with locally appropriate counterparts, e.g. 'African American' with 'Black British'.
- To adapt questions for our data which is formed from local concepts rather than codes, questions referring to specific codes (e.g. CPT[4] procedure code '92960') were removed.
- On the eHealthScope platform, time information is only available where it is pre-programmed into specific concepts, therefore we removed any chronological aspect to questions.
- To simplify our chained question setup, we kept only the first example where there were multiple instances of questions with little variation e.g. 'which patients > 17 yo have atopic dermatitis?' and 'which patients are > 17 yo and have atopic dermatitis'.

After these modifications, the EpiCoho-M(odified) dataset contained 39 questions.

We also required to convert the ground truth queries from SQL expressions referring to clinical codes, to boolean logic referring to concepts. For each of the 39 questions, we manually created a single reference answer. During the generation process, answers which did not match these reference answers were reviewed manually; any which were valid implementations of user intent were added to the set of allowable gold-standard queries.

Dataset 2: UK Medicine Brand Names We downloaded a dataset of 266 brand names from the UK medicines agency [5], and mapped them to reference concepts based on their constituent compound names. For example, brand names of the compound 'Tamsulosin Hydrochloride' were mapped to the reference concept '[prescribed] Tamsulosin Hydrochloride, last 3 months'. A simple dataset was then created by inserting the brand name into a question template of the form "List patients who take `{brand_name}`".

Dataset 3: Real World Dataset During the real world beta testing phase, there were 237 conversations up until 26th November 2025. For each of these conversations, the initial message was inspected to determine if it contained a clear population health request which should be answered with a query (e.g. 'patients on tirzepide') or not (e.g. 'what can you do', 'Add a new user'). The latter were filtered out, giving a total of 82 requests.

For each request, we then created a suitable response query and executed this to discover the corresponding real world patient cohort, which was designated as the reference cohort.

[4] AMA Current Procedural Terminology Codes: https://www.ama-assn.org/practice-management/cpt

[5] www.gov.uk/government/publications/category-lists-following-implementation-of-the-windsor-framework

C User Feedback from Beta Testing Phase

We deploy our system in the real world and provide a feedback mechanism for users. Figure 4 shows that there were 128 users who engaged in 354 chats over the 8-month period 2025-06-26 to 2026-02-13. For each request, we asked users to judge whether the system's response was 'useful'. Table 4 shows the results. Whilst users only chose to provide feedback 13% of the time (47/354), they were positive about the system's response for 62% of these (29/47). Additionally, they chose to access the query results 65% of the time (230/354), indicating that the response was considered potentially correct.

Feedback	count	url clicked
positive	29	24/29
negative	18	9/18
no response	307	197/307
Total	354	230/354

Table 4: User labels annotated during the beta testing deployment. Users were prompted with the message 'Was this useful?' followed by thumbs-up and thumbs-down options to give positive and negative feedback.

We further analyse the 14 responses which received negative feedback. Table 5 shows that for 6 responses, user dissatisfaction arose from making requests that are out of scope for eHealthScope. For 2 responses, the issues have since been fixed. For the remaining 6 requests, issues range from the agent throwing an error to the response not being user-friendly to the agent making errors in not finding the correct concept or combination of concepts.

User Role	Intent	Analysis
Practice Manager 1	"how do I add a new user"	This action is not supported by the agent
Practice Manager 2	"edit a user"	This action is not supported by the agent
Practice Manager 2	"edit a user"	This action is not supported by the agent
Practice Manager 3	"how do I modify user permissions"	This action is not supported by the agent
Practice Manager 4	"permissions log"	This action is not supported by the agent
Receptionist	"speak to a human being"	This action is not supported by the agent
Data Analyst	"Crude prevalence of personality disorders [...] by PCN. Ages 15+ only"	The agent was unable to find a concept representing 15+ and incorrectly presumed that none existed, leading it to provide an approximate answer
Developer 1	"how do I profile populations"	The agent repeats the section of its prompt describing profiling populations and the user gives the free-text feedback: "[the resulting explanation] is not very friendly for a new user"
Developer 1	User: "diabetes quick" Agent: "I'm not sure what you mean by that"	The user is asking for a type of document and gives the free-text feedback "A 'quick' is a type of document"
Pharmacist 1	User: "how many patients are prescribed PERT"	The agent counted patients taking 'Pertuzumab', which is a different entity to 'Pancreatic Enzyme Replacement Therapy (PERT)'
Principal Analyst	Agent claimed to have returned the result but had only created a cohort.	In alpha testing the agent would occasionally do this. This issue was resolved by improving the system and tool prompts.
Developer 1	Agent responded but did not generate a url link to view the results.	This issue was resolved by refining the agent's prompt.

Table 5: Error analysis for each instance of negative feedback received in initial four months beta deployment. The system was updated to support the requested actions related to user management in general practices.