

AI-Powered Drug Classification and Indication Mapping for Pharmacoepidemiologic Studies

Benjamin Alexander Ogorek, Thomas Patrick Rhoads, Eric Scott Finkelman,
Isaac Rogelio Rodriguez-Chavez

Submitted to: JMIR AI
on: August 16, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 34

 Figures 35

 Figure 1..... 36

 Figure 2..... 37

 Figure 3..... 38

 TOC/Feature image for homepages 39

 TOC/Feature image for homepage 0..... 40

AI-Powered Drug Classification and Indication Mapping for Pharmacoepidemiologic Studies

Benjamin Alexander Ogorek¹ PhD; Thomas Patrick Rhoads¹ MBA; Eric Scott Finkelman¹ BS, MS; Isaac Rogelio Rodriguez-Chavez^{1,2} MS, MHS, PhD

¹Spencer Health Solutions, Inc. Morrisville US

²Biosolutions Consulting Rockville US

Corresponding Author:

Benjamin Alexander Ogorek PhD

Spencer Health Solutions, Inc.

2501 Aerial Center Pkwy, Ste 100

Morrisville

US

Abstract

Background: Pharmacoepidemiologic studies require Anatomic Therapeutic Chemical (ATC) drug classification from real-world data sources. These studies enable standardized analysis of drug utilization patterns and safety monitoring, ultimately promoting rational drug use and improving health outcomes. Proprietary tools for this purpose are expensive while free tools lack generalizability. Large language models (LLMs), like GPT-4o, offer a cost-effective alternative as they can produce explanations about a drug's ATC code and return the output in a structured fashion.

Objective: This paper seeks to establish LLMs as an assisting technology in the drug classification task, a prerequisite to good pharmacoepidemiologic research. This requires developing AI prompts and data processing procedures and showing that the resulting accuracy, efficiency and effectiveness is as good or better than established methods.

Methods: Patients residing in the US and Canada with medication scheduled through a smart medication dispenser called "spencer SmartHub" (Spencer Health Solutions, Inc., Morrisville, NC) were included in this study if they had a scheduled medication refill in 2024 and consented to the use of their data for research. An AI prompt requesting best and next-best 2nd level ATC codes from de-identified daily-dose strings was generated iteratively with expert guidance on clinical research, digital medicine, and regulatory affairs. An initial prompt was created that ensured aspirin at various doses would be classified as either an analgesic or antithrombotic. Upon success, the prompt was used in a pilot sample of 20 daily dose strings and graded by the expert. While there was more than one incorrect response, the prompt was revised. The prompt was then applied to an inference sample of n=200 daily dose strings, taken without replacement. Finite population inference on the proportions of correct and approximately correct ATC drug classification was carried out. All errors made by the algorithm were reviewed.

Results: There were 3,371 de-identified patients who met the inclusion criteria, 2908 (86%) residing in Canada and 463 (14%) residing in the United States. This resulted in 12,294 daily dose strings. The initial prompt with few-shot learning and concise output was unable to distinguish between aspirin's analgesic vs antithrombotic therapeutic uses. A revised prompt using chain-of-thought reasoning succeeded and achieved 100% correctness on the pilot sample of n=20. In the inferential sample, a proportion of 0.96 (80% CI 0.943-0.978), were deemed correct by the expert, with the approximately correct designation never being used. The top mistakes were incorrectly classifying dietary supplements as medications, mistaking the identity of a drug, and incorrectly following delimiter instructions.

Conclusions: GPT-4o offers an accurate, efficient and effective drug classification approach to augment real-world drug databases with ATC drug classes, giving all research teams access to a powerful tool to satisfy a key prerequisite of pharmacoepidemiologic analysis using real-world data from across the globe.

(JMIR Preprints 16/08/2024:65481)

DOI: <https://doi.org/10.2196/preprints.65481>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <http://www.jmir.org/>

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

AI-Powered Drug Classification and Indication Mapping for Pharmacoepidemiologic Studies

Abstract

Background: Pharmacoepidemiologic studies, which promote rational drug use and improve health outcomes, often require Anatomic Therapeutic Chemical (ATC) drug classification within real-world data (RWD) sources. Existing classification tools are expensive, brittle, or have restrictive terms of service, and lack context that may inform classification itself.

Objective: This paper seeks to establish large language models (LLMs) as an assisting technology in the drug classification task. This includes developing AI prompts that reason about drugs using RWD and showing that the resulting accuracy, efficiency and effectiveness are favorable to alternative methods.

Methods: A prompt was constructed to classify aspirin as either an analgesic or antithrombotic and evaluated within 12,294 anonymized daily dose strings from a polychronic population residing in the US and Canada. The patients used a smart medication dispenser called “spencer” (Spencer Health Solutions, Inc., Morrisville, NC) and consented to the use of their data for research. The LLM prompt requested that a best and next best second-level ATC code be returned and grading was performed on a three-point scale. After success in a pilot sample of 20, an inference sample of 200 was taken without replacement. Finite population inference was carried out on the proportion of outputs receiving one of the top two grades. As a benchmark, Google’s Custom Search API was used to query the drug name plus “ATC code” followed by regex-based extraction of ATC codes. All imperfect results were reviewed.

Results: The population consisted of 12,294 daily dose strings from 2,908 (86%) patients residing in Canada and 463 (14%) residing in the US. A prompt using chain-of-thought reasoning was able to distinguish between aspirin’s analgesic vs antithrombotic therapeutic uses and performed well in the pilot sample. In the inferential sample, 175 (n=200, 87.5%) were graded as perfect, 10 (n=200, 5.0%) had a minor issue, and 15 (n=200, 7.5%) had a major issue. The estimate of the proportion of at least mostly correct classification was 92.5% (185/200, 80% CI 90.1-94.9). For the search-based algorithm, 163 (n=200, 82.5%) were deemed acceptable. The chain-of-thought reasoning was most helpful with supplements (e.g., folic acid) when high doses indicated antianemic preparations. The problem formulation of daily dose inputs and multiple ATC outputs was sometimes incompatible with the drug (e.g., pregabalin, calcitriol, methotrexate).

Conclusions: GPT-4o offers cost effective drug classification from RWD without violating any terms of service. Using a chain-of-thought prompting technique, GPT-4o can reason about drug dosages that affect the class. The wide accessibility of LLMs gives every research team the ability to classify drugs at scale, a key prerequisite of pharmacoepidemiologic research.

Keywords: Generative language model; artificial intelligence; AI; LLMs; NLP; Natural Language Processing; drug classification; ATC; spencer device; smart hub

Background

Introduction

Pharmacoepidemiology is a "bridge science" between epidemiology and clinical pharmacology that has, since its formative years, encompassed wide-ranging fields such as epidemiology, pharmacology, medicine, biostatistics, and social sciences [1]. It is positioned to benefit from advances in an even wider array of disciplines, including bioinformatics, data science, machine learning, artificial intelligence, natural language processing, large language models, systems pharmacology, pharmacogenomics, pharmacometabolomics, and health informatics [2–8]. These

advances are integrated into pharmacoepidemiologic research to discover ineffective or unsafe drugs, encourage better prescription decisions, and reduce unnecessary healthcare costs [9].

Large Language Models (LLMs) are assistive tools that process broad ranges of factual knowledge, opening new avenues for synthesizing information across diverse fields. Their incorporation into pharmacoepidemiology holds great potential for redefining how we can understand the safety, efficacy, and utilization patterns of prescribed medications in large patient populations and reduce unnecessary costs.

Pharmacoepidemiology and the Anatomical Therapeutic Chemical (ATC) Classification System

Pharmacoepidemiologic research depends on a concept of *drug class*, a classification of drugs in terms of mechanism of action, therapeutic intent, or chemical structure, with the Anatomical Therapeutic Chemical (ATC) classification system considered mandatory [9–12]. Applications of the ATC methodology according to the WHO are included in Table 1.

Table 1. Applications of the ATC Methodology according to the WHO.

Application	How ATC Codes are used
National Standard for Medicinal Products	National standard for classification of medicinal product for various countries
International Classification	A common language for describing drugs
Health Policy	Drug utilization statistics using ATC codes are used to improve the quality of drug use in a population
Pharmacoepidemiology/Drug Utilization Research	Study trends and patterns in drug use
Pharmacovigilance	Linking Adverse Drug Reactions (ADRs) to ATC classes
Assisting procurement agencies and payer organizations	To gain an overview of the availability of drugs and reduce the risk of drug shortages.

The ATC system classifies active substances (i.e., drugs), including combinations of active ingredients, into groups at five different levels. This five-level structure of the ATC system enables a hierarchical classification, which supplies broad and specific categorization of drugs, and helps in different levels of analysis in pharmacoepidemiologic studies. Using the example of lisinopril and hydrochlorothiazide, the first level ATC code is “C” for Cardiovascular, the organ or system acted upon, the second-level ATC code is “C09” for agents acting on the renin–angiotensin system, and the fifth level ATC code is “C09BA03”: lisinopril and diuretics [13,14]. While there is the basic principle of one ATC code for each drug [13], here the fifth level ATC code is still ambiguous due to the presence of an unspecified diuretic.

The ATC system is not suitable for guiding decisions about reimbursement, pricing, and therapeutic substitution [15], and thus real-world data typically lack ATC codes, instead using drug codes that support day-to-day operation and logistical functions within healthcare and pharmaceutical systems. This discrepancy between research needs and real-world data structures presents a significant challenge for pharmacoepidemiologists.

In the United States, most pharmacy management systems and electronic health record systems use National Drugs Codes (NDCs) which were designed for inventory management and reimbursement and do not directly identify a drug class [10,16]. In Canada, the analogous system is the Drug

Identification Number (DIN) [17]. In the Netherlands, it is the Z-Index [18]. To conduct pharmacoepidemiologic research with these data sources, accurate ATC classification of drugs becomes a task essential to the research process. This necessity for accurate drug classification underscores the potential value of advanced computational methods, such as machine learning and natural language processing, in bridging the gap between diverse drug coding systems and the standardized ATC classification required for robust pharmacoepidemiologic analysis.

Tools to Map Prescribed Drugs to ATC Classes

The Center's ATC/DDD Index web utility includes a query engine for drug name strings [19]. While there is no published Application Programming Interface (API) for the search tool, adding `?name=<drug>` to the end of the URL will execute the search and could set up a web scraping task. Fuzzy matching is limited to the options of "containing query" and "starting with query," which is restrictive in practice. For example, in either case, the drug string "apo-domperidone" returns no results while "domperidone" returns three.

ATC codes appear in Wikipedia infoboxes which suggests the use of DBpedia [20]. Querying DBpedia, however, requires exact resource names, for instance, "dbr:Metformin." The DBpedia organization does provide a lookup service [21] for text queries, like the ATC/DDD Index, but it is also thrown off by variations in the drug's name.

When operational drug codes (e.g., NDC or DIN) are in the database, a different suite of tools becomes available. In the US, the National Library of Medicine offers a service around RxNorm, a standardized nomenclature for clinical drugs [22], which links NDC to fourth-level ATC code [23]. Health Canada's Drug Product Database [24] allows joining on the Drug Identification Number (DIN) directly. For countries with languages spoken by fewer people, the task is more difficult. Researchers in the Netherlands, for instance, researchers had to resort to multiple data sources and extensive data cleaning to build their Dutch ATC ontology and still requires manual tasks [25].

DrugBank offers full-text search built on top of AWS Elasticsearch that can map drug names in various forms to standard terminology [26,27]. A DrugBank search result includes the full list of appropriate ATC codes, synonyms, indications, and more. While DrugBank is offered to the public and academia for free, commercial use requires a paid license [28]. The reliance on such commercial products can limit the accessibility of crucial drug classification tools, potentially creating disparities in research capabilities between well-funded and resource-limited institutions.

A commercial option that is inexpensive, yet controversial, is Google's Custom Search API [29]. The results contain the data essential to displaying search results on a website and, while they do not include the entire content of each site, they contain a "snippet" which often includes key information of interest. While recent US court decisions indicate a trend towards more leniency in web scraping of publicly accessible data, such as *hiQ Labs v. LinkedIn* [30], Google's terms of service explicitly prohibit storing information obtained from the service in "any non-transitory manner" [31]. This in principle makes possible a claim against the user for breach of contract, but the actual enforceability of these terms is questionable [32].

ATC Drug Classification Challenges in a Multi-Regional Smart Medication Adherence System

Integration into the medication database of the ATC classification is critical to ensuring robust pharmacoepidemiologic research across a wide array of healthcare systems. This section demonstrates the challenges and limitations to the implementation of the ATC classification in the context of a real-world smart medication dispenser.

Spencer Health Solutions, Inc. (SHS) is a health technology startup that manufactures "spencer" [33], a smart medication dispenser that dispenses strip-packaged oral solid tablets. At the time of drug refill creation, records are inserted into a database (maintained by SHS) which include the

scheduled time of each pouch and information on the drugs in each pouch. A separate set of drug records contains drug name, drug code (NDC for US drugs, DIN for Canadian, and Z-Index for Dutch drugs), and strength description.



Figure 1. Key components of the spencer smart medication dispenser.

The spencer device shows the potential of smart medication dispensers to generate real-world data that are useful in pharmacoepidemiologic studies, but there have been difficulties with relying on operational drugs codes to obtain second-level ATC codes. First, many of the drugs in the SHS drug database do not contain valid NDC or DIN codes, especially supplements such as “*advanced 4-strain probiotic*” or “*webber naturals womens 50 plus most*” (full name truncated in the database). Additionally, when the daily dose will inform the therapeutic subgroup, NDC and DIN codes are insufficient since patients often take multiple pills per day. For example, aspirin at low daily doses (75-100mg) is a match to second-level ATC code B01, i.e., antithrombotic agents, whereas aspirin at doses above 325 mg maps to N02, i.e., analgesics. Finally, since the original drug classification exercise, patients began using spencer in the Netherlands, with Z-index values populating the drug code field in the European database. Though out of the scope of this paper, the existing classification pipeline based on two region-specific methodologies is unhelpful for this new drug source. These challenges indicate both the complexity of drug classification in multi-regional systems and that more sophisticated, context-aware approaches would be required, accounting for differences in dosage, formulation, and regional coding systems.

The limitations seen in the real-world application set a clear case for the need for a more flexible, comprehensive, and automated approach to drug classification. Advanced computational methods, such as machine learning, natural language processing, and artificial intelligence, can help address these challenges by providing more accurate, adaptable classification systems dealing with diversified drug information across different regions and contexts of healthcare.

Drug Classification Using GPT-4o

In recent years, large language models (LLMs) have shown the ability to perform a broad array of tasks with minimal task-specific training using textual prompts with “few-shot” and “chain-of-thought” prompting techniques [8], enabling the automation of tasks that are currently difficult and manual. GPT-4o [34], OpenAI’s most advanced model at the time of writing the initial manuscript, contains a wealth of information about pharmaceutical drugs as well as the ATC classification system. In interactive exploration, GPT-4o was good at reasoning about second-level ATC level codes from drug inputs that had a variety of input formats. This feature of flexibility in handling various formats of input is valuable as sources of drug information are heterogeneous among different health-care systems and across countries. Additionally, a GPT-4o prompt can include additional information in an efficient and optimized manner about the daily dose that a patient is taking, which is useful when the dose may influence the final ATC classification.

The application of LLMs to drug classification presents a novel approach that could potentially address many of the limitations with traditional methods of drug classification. In this context, a large pharmacoepidemiologic knowledge base that utilizes LLMs (e.g., GPT-4o) is robust in its handling of regional variations, dose-dependent classification, and data harmonization across different healthcare systems. Therefore, LLMs may improve multi-regional pharmacoepidemiologic research capabilities related to drug classification challenges.

Objective

This paper seeks to establish LLMs as an assisting technology in the drug classification task. This includes developing AI prompts that reason about drugs using RWD and showing that the resulting accuracy, efficiency and effectiveness are comparable to alternative methods. The developed prompts should available immediately to classify drugs under a wide range of research budgets.

Methods

Patient Population

Drug records of patients were included in this study if they met the criteria described in Textbox 1.

Textbox 1. Inclusion criteria for drug records in the study.

1. The patients resided in the US and Canada and belonged to a value-based care management organization.
2. The patient signed the care organization's consent form and agreed to the Spencer Health Solutions, Inc. end-user license agreement, permitting their de-identified data to be used for research purposes.
3. A refill was created with a pouch scheduled to be dispensed on or after 2024-01-01, as queried on 2024-06-01.

No raw drug record was sent outside the corporate network as part of this research. As will be described in the next section, the data submitted to GPT-4o was an irreversibly anonymized set of daily drug dose strings.

Anonymization Procedure for Daily Drug Doses

This section describes the anonymized extraction of drug dosing data from the database. A drugs table in the application database included the drug name, the strength as a text string, and the quantity of pills (including fractions). The drug names were sometimes combinations of a manufacturer and a generic (e.g., "apo-rosvastatin") and other times a brand name (e.g., "prolopa 50-12.5"). Other times, the drug name would be as generic as "acetaminophen." Dosage strengths were unstandardized strings such as "50mg" or "50 MG" and could include multiple active ingredients ("50/200MG" for carbidopa levodopa), a unit of time ("300.0 MG/24HR"), and other variations. (Note that while lower case "mg" is more appropriate than upper case "MG" to represent milligrams, the two will be used interchangeably in this paper.)

To construct a set of inputs suitable for OpenAI's chat completion API, a string was constructed for each unique drug and daily dose combination, without any other patient information. The drug name was processed by only a lower-case transformation. Drug strength was processed by imputing missing strengths as "UNK" and otherwise left unchanged (strengths of "0" were allowed to pass through). The daily dose processing was described in Textbox 2.

Textbox 2. The processing of daily dose.

1. Pill quantity was summed by PatientId, date, drug name, and drug strength.
2. A pill quantity string was defined within the previous data set using the formula:
 - a. "{quantity} pill" if quantity = 1,
 - b. "{quantity} pills" otherwise
3. The previous data set was aggregated over drug strength within patient, date, and drug name by creating a comma-delimited list of multiple strengths for each drug, taking the form "{pill quantity string} of {drug strength}."
4. The patient identifier was discarded, and a distinct operation performed on the drug name and drug strength string.

For example, consider a patient prescribed the drug with name "pms-quetiapine" (generic quetiapine fumarate from manufacturer Pharmascience, Inc.) twice per day at 8 AM and at 7 PM. At each dispense, one pill of strength 200 mg and two pills of strength 25 mg are scheduled. Following the steps described in Textbox 2, "pms-quetiapine|2 pills of 200 mg,4 pills of 25 mg" would be the

resulting string input. An example with combination ingredients is 1 pill of “carbidopa/levodopa er” of strength “25/100 MG” scheduled twice daily, or “carbidopa/levodopa er|2 pills of 25/100 MG.” These are anonymized records that cannot be linked back to an individual.

Developing the Classification AI Prompt

The iterative development of the AI classification prompt used a single motivating example, the case of aspirin, and the goal of iterative prompt development was to guide the LLM to classify low doses as ATC code B01 for antithrombotic use and higher doses as ATC code N02 for pain relief. Aspirin is a well-known example of a drug where context matters, and any bias from this specific motivating example will be apparent in the results. The creation of the prompt was an iterative exercise that existed in two phases: the initial prompt creation and revision using the pilot sample. In the inference stage, the prompt was fixed, and no further changes were made.

During the initial prompt creation phase, two techniques that were available to be employed were few-shot learning and chain-of-thought prompting. After achieving a prompt that worked as desired for aspirin at different daily dosing levels, 20 drug names and drug strength strings were randomly sampled from all possibilities observed in the data set and served as a pilot sample for the initial prompt. These were evaluated by expert review by co-author Dr. Isaac Rodriguez-Chavez, with expertise in clinical research, digital medicine, and regulatory affairs.

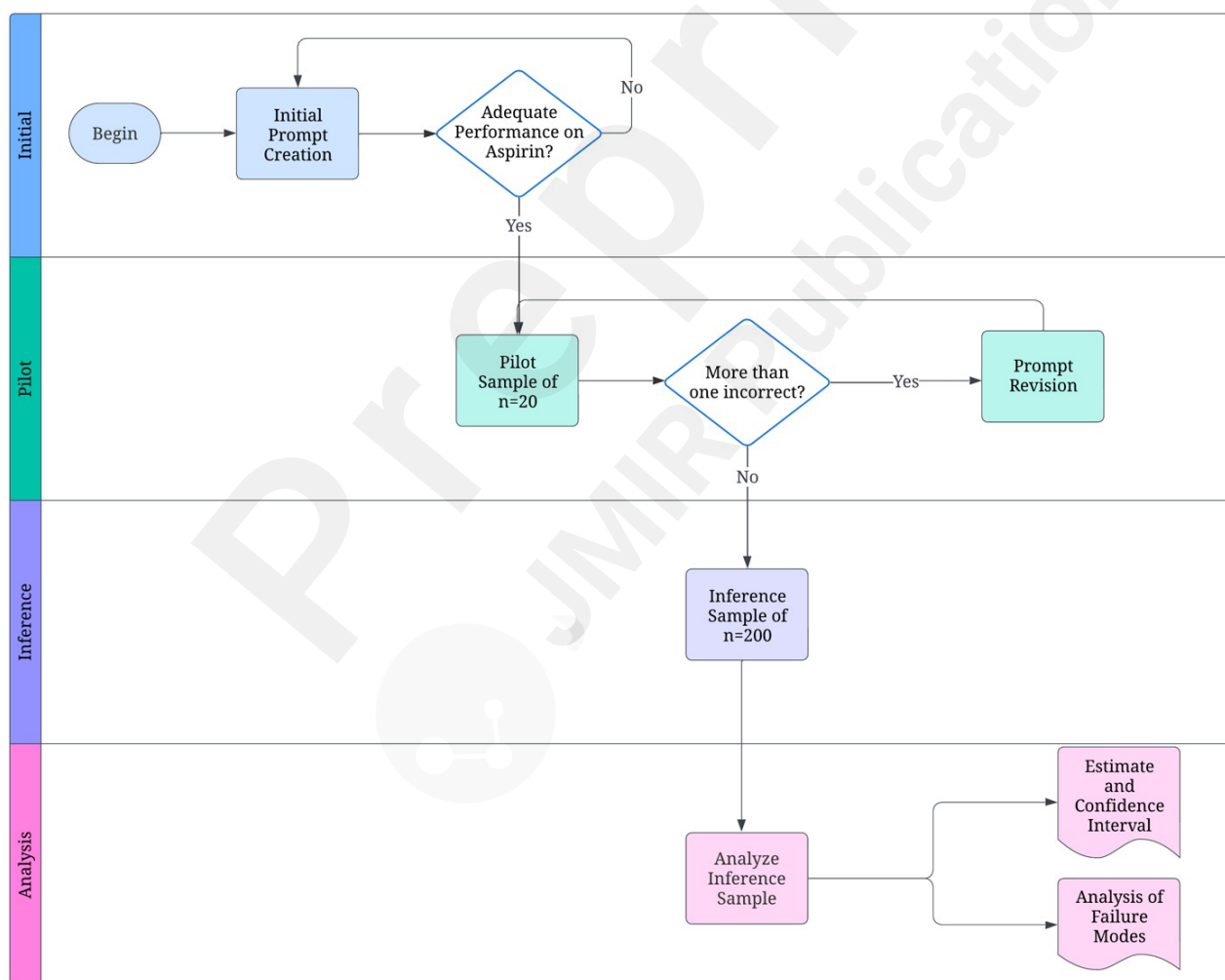


Figure 2. The methodology used to develop and test an ATC classification AI prompt.

The pilot sample was sent to GPT-4o via the OpenAI Batch API [35] with all parameters set at their defaults except for the following. Temperature is a parameter that can vary between 0 and 2, where low values (e.g., 0.2) result in more consistent outputs whereas higher values result in “more

creative” results (e.g., 1.0) [36]. This research used a temperature of 0, as consistency was a priority. The other non-default parameter was “max_tokens,” set to 1000, which represents the maximum amount of tokens that can be generated in the chat completion [37]. A token is on average $\frac{3}{4}$ of a word (100 tokens is about 75 words) [38], and max_tokens was set below the highest allowed value of 4096 to avoid the longest explanations but to still allow for long responses if necessary.

If more than one classification was deemed incorrect, the prompt would be revised before proceeding. The development and validation of this AI-driven classification system was thus optimized through an iterative refinement process guided by expert feedback (see Figure 2).

Finite Population Inference

This section details the statistical methodology that was used in estimating the performance of the GPT-4o model and custom prompt in second-level ATC classification. The methodology focused on the finite population of daily dose strings in the SHS database. Finite population formulas, while not materially different in this case from their infinite population counterparts, were used to emphasize that the inference is made to the specific set of SHS drug prescriptions and not to any larger population. This approach emphasizes the study's focus on the internal validity of the findings. The finite population sampling approach is model free and has well understood properties in repeated sampling [39].

The values that were sampled were numerical grades of one, two, and three, generated in the following manner. The output from GPT-4o was graded collaboratively amongst the authors on a three-point scale, where a grade of one (1) is flawless, a grade of two (2) is more correct than incorrect, and three (3) is more incorrect than correct. While the counts of each grade will be provided, formal inference to the full population of 12,294 daily dose strings focused on p , the proportion of at least mostly correct grades (i.e., grades of one or two) in this population. This focused the presentation of performance on a meaningful criterion while reducing inferential complexity.

The finite population sampling estimate of p is $\hat{p}$, the sample proportion of outputs that are at least mostly correct in the chosen sample of n . The finite population confidence interval is

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} \cdot \frac{N-n}{N-1}},$$

where N is the number of daily drug dose strings in the full population, n is the sample size, and α is the proportion corresponding to the approximate $(1-\alpha) \cdot 100\%$ confidence interval for population parameter p . This formula accounts for the finiteness of the population, giving a better estimate of the confidence interval compared with formulas for infinite populations.

The confidence interval formula was also used to arrive at the sample size. A margin of 0.02 combined with a population proportion of 0.95 and an 80% confidence leads to a sample size of 189, which was rounded up to $n=200$. The lower level of confidence than the typical 95% was chosen here to lead to a somewhat smaller sample, and there is precedent in making this tradeoff in medical research contexts where “reasonable certainty” is acceptable [40].

Each imperfect output (grades of two or three) was examined qualitatively and presented to the reader in tabular form. The errors were then coded into categories and the frequencies displayed visually. For mistakes that were difficult to explain, the OpenAI interactive “playground” was consulted for reproducibility.

The Benchmark of the Google Custom Search API

When researching a single drug, it is generally easy to find ATC codes by searching for the drug name in Google followed by “ATC code.” Despite the issues with automation of this approach at scale, the drugs from the 200 daily dose strings were sent to the Google Custom Search API in the form “<drug name> ATC Code.” For each drug, a Python program looped through the top 10 results

(the default number of items returned) and searched for a strings matching the regular expression '[A-Z]\d{2}[A-Z]{2}\d{2}'. The first three characters of these strings were added to a list, and the highest frequency second-level ATC code was the choice made by the search-based algorithm.

The grading of the search-based algorithm was simplified. While the LLM-based algorithm had to find the best and next best second-level ATC codes, the search-based algorithm only had to find the single best ATC code. Furthermore, there were only two grades applied: “appropriate” and “not appropriate,” where appropriateness is for a database of oral solids. If there is a tie between two second-level ATC codes and both are appropriate, then the output is appropriate. If there is a tie and at least one code is inappropriate, then that output is not accepted. This sets up the fraction of appropriate grades as a benchmark, acknowledging that the search-based algorithm is not encouraged to pick ATC codes relating to oral solids.

The results of the search-based algorithm are adjunctive to the results of the LLM and the decision was made not to pursue formal inference. Comparing the success rates was inevitable but not the goal of the paper. It was however informative to see cases where the Google search fails and the LLM succeeds, and vice versa. A table of cross frequencies was thus compiled, and all cases where the search-based algorithm’s output was graded “not appropriate” was added to the table of errors.

Ethical Considerations

This study used operational data collected from a commercial medication dispensing system used in routine patient care and was not subject to IRB review requirements, so approval was not obtained. Users of the spencer device provided consent for data collection through the End User License Agreement (EULA), which covers the collection of medication adherence data and responses to quality of life and patient-reported outcome PRO surveys as part of the system’s standard operation. No additional compensation was provided to users beyond the normal terms of their device usage agreement. All data analyzed in this study were de-identified prior to analysis. Spencer Health Solutions has achieved both ISO27001 and Data Privacy Framework certifications, and the system employs uses industry-standard encryption and security measures.

This research analyzed data collected during standard clinical care and device usage. All results are presented as anonymous aggregate statistics. The original data collection occurred as part of routine clinical practice, with patients providing consent for research use through the device terms of service and care management agreement. Under Canadian TCPS 2 Article 2.4, research ethics board. Research Ethics Board review is not required for research that relies exclusively on secondary use of anonymous information where the process does not generate identifiable information. Under US regulation 45 CFR 46.104(d)(4)(ii), IRB review is not required when information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects, the investigator does not contact the subjects, and the investigator will not re-identify subjects.

Results

The Initial AI Prompt

The initial prompt constructed appears in Textbox 3. With a goal of classifying aspirin in a dose-dependent manner with instructions and examples, this prompt requested output consisting of only ATC codes in a pipe-delimited list.

Textbox 3. The initial AI prompt: few-shot learning with concise output.

```
# Instructions
You are a classifier of drug prescriptions into Anatomical
Therapeutic Chemical (ATC) second level subgroups. As a drug and
dose combination may contain multiple ATC second level subgroups,
```

your job is to return the closest ATC second level category, followed by the next closest. If there is one and only one category, then return "NA" for the second category. If the drug cannot be classified or is ambiguous, return "UNKNOWN" for both ATC fields. Return exactly two fields are separated by a pipe ("|").

Input and Output format

- Input: <drug name>|<daily drug dose>
- Output: <Best level 2 ATC Code>|<Next Best level 2 ATC Code (if applicable)>

Additional Information

All drugs are oral solids. The drug name as a string. Handle different capitalizations, common misspellings, and concatenations with manufacturer names. If the daily drug dose is missing or unintelligible, then do your best with the drug name alone.

Examples

* Input: metformin|2 pills of 500MG
* Output: A10|NA

* Input: apo-dexamethasone|0.5 pills of 4.0 MG
* Output: H02|S01

* Input: XYZ-1234|1 pill of 1g
* Output: UNKNOWN|UNKNOWN

* Input: webber naturals womens 50 plus most
* Output: A11|NA

When tested interactively, the initial prompt repeatedly failed to classify high doses of aspirin with N02 as the most likely second-level ATC code. The following inputs all consistently produced the response "B01|N02": "aspirin|1 pill of 81mg," "aspirin|2 pills of 325mg," "aspirin|12 pills of 325mg," and "aspirin|3 pills of 1000mg." Sentences were added such as, "Prioritize the dose" and "Think about what conditions the total daily dose would be most likely to treat," but these were not effective. Results from testing this first AI prompt, with a focus on aspirin classification performance, emphasized challenges faced in the development of a robust classification system that may deal with dose-dependent categorizations. This motivated the creation of a second prompt with chain-of-thought reasoning techniques.

The Revised AI Prompt Using Chain of Thought Reasoning

Based on the results of the initial AI prompt, the preference for concise output was dropped in favor of a chain-of-thought prompting technique. Here, it is necessary to return the reasoning before outputting the ATC classes so that the LLM can use the text generated by the reasoning. Few-shot learning examples were used here as well, and the entire revised AI prompt is shown in Textbox 4. Textbox 4. Revised AI prompt with chain-of-thought prompting techniques.

Instructions

You are a classifier of oral drug prescriptions into Anatomical Therapeutic Chemical (ATC) level 2 therapeutic subgroups. Handle different capitalizations, common misspellings, and concatenations

with manufacturer names. As a drug and dose combination may be linked to several ATC level 2 subgroups, your job is to return the most likely and next most likely level 2 subgroups. The user's input should take the form: <Oral solid drug name>|<Daily drug dose>, whereby you should output a 3-column, pipe-delimited message taking the form: <Reasoning>|<Most likely level 2 ATC Code>|<Next most likely level 2 ATC Code (or "NA")>. Reason step by step as demonstrated in the examples, taking into consideration the most likely indication given the daily dose.

Examples

- * Input: metformin|2 pills of 500MG
- * Output: 2 pills of 500MG is 1000MG per day. The typical effective dose range for metformin in managing type 2 diabetes is between 500 mg to 2000 mg per day, and 1000MG is right within this range. The code A10 refers to "Drugs used in diabetes."|A10|NA
- * Input: apo-dexamethasone|0.5 pills of 4.0 MG
- * Output: 0.5 pills of 4.0 MG is 2.0 MG per day. Dexamethasone at this dose is typically used for mild to moderate inflammatory and autoimmune conditions. The primary code H02 refers to "Corticosteroids for systemic use, plain."|H02|NA
- * Input: XYZ-1234|1 pill of 1g
- * Output: XYZ-1234 is not an oral solid drug I am familiar with.|NA|NA
- * Input: webber naturals womens 50 plus most|1 pill of UNK
- * Output: The dose is unintelligible, but this is a multivitamin and is classified by ATC Code A11: Vitamins.|A11|NA

As a side note, dexamethasone from the initial prompt examples was omitted from the revised prompt. It's secondary classification of S01, an ophthalmological, would not be dispensed through spencer as it is a liquid eye drop.

The revised prompt was able to classify aspirin in the desired manner. For instance, in the OpenAI playground, the assistant responded to the input "aspirin|1 pill of 81mg" with the following output. Textbox 5. Output for aspirin at 81 mg per day.

"1 pill of 81mg is 81mg per day. Aspirin at this low dose is typically used for its antiplatelet effects to prevent cardiovascular events such as heart attacks and strokes. The primary ATC code for antithrombotic agents is B01.

Reasoning: Aspirin at 81mg is commonly used for cardiovascular protection rather than for pain relief or anti-inflammatory purposes, which would require higher doses.

|B01|NA"

Despite the carriage returns, the output was pipe-delimited as requested, and the reasoning correctly specifies that aspirin that this dose is used for its antiplatelet effects rather than its pain relief effects. Note that it did not include N02 as the next most likely drug classification, and this was deemed acceptable since a patient is unlikely to be taking 81mg of aspirin daily for pain. An input of "aspirin|2 pills of 325mg" led to the output in Textbox 6.

Textbox 6. Output of aspirin at 650 mg per day.

"2 pills of 325mg is 650mg per day. Aspirin at this dose is typically used for pain relief, anti-inflammatory purposes, and sometimes for cardiovascular protection. The primary ATC code for aspirin when used for pain and inflammation is N02: "Analgesics." When used for cardiovascular protection, the ATC code is B01: "Antithrombotic agents." Given the dosage, it is more likely to be used for pain relief.

Reasoning: 2 pills of 325mg is 650mg per day. Aspirin at this dose is typically used for pain relief and anti-inflammatory purposes. The primary code N02 refers to "Analgesics."

|N02|B01"

Thus, the revised prompt with chain-of-thought reasoning returned N02 as the best ATC code for high dose aspirin and B01 as the next best second-level ATC. This was deemed acceptable as aspirin could conceivably be taken for a dual purpose, and the revised prompt was promoted to the pilot sample phase.

A Finite Population of Daily Drug Doses

To this point, all AI prompt testing was done via hypothetical examples. For testing on prescription drugs in a real population, 3,371 de-identified patients met the inclusion criteria described in Textbox 1, 2,908 (86%) residing in Canada and 463 (14%) residing in the United States. These patients collectively had 4.76 million doses scheduled after 2024-01-01 as of the database query date of 2024-06-01. Within these doses, there were 2,077 distinct drug names, 517 that were prescribed to patients residing in Canada and 646 prescribed to patients residing in the US. Only 86 drug names prescribed to patients residing in both regions. When combined with daily pill quantity and strength information, there were 12,294 daily drug prescription strings (e.g., "pms-quetiapine|1 pill of 100.0 MG, 1 pill of 200.0 MG"), the finite population of interest.

The Pilot Sample for Testing the AI Prompt

The random sample of 20 daily dosage strings was used to determine if the revised prompt would continue to be studied. The 20 daily dose strings appear in Table 2 alongside the output from the OpenAI Batch API.

Table 2. The pilot sample for testing the revised AI prompt.

Input String for User Role	Best second-level ATC	Next best second-level ATC
carbamazepin tab 2 pills of 200MG	N03	NA
calcium-ng-vitd - jam 1 pill of UNK	A12	NA
teva-entacapone 7 pills of 200.0 MG	N04	NA
apo-midodrine 8 pills of 2.5 MG	C01	NA
hydroxyzine hydrochloride 4 pills of 10.0 MG	N05	R06
apo-domperidone 2 pills of 10.0 MG	A03	NA
aventyl 4 pills of 25.0 MG	N06	NA
sandoz perindopril erbumine/indapamide hd 2 pills of 0.0	C09	C03
vitamin d2 1 pill of 1.25 mg	A11	NA
synthroid 2 pills of 125.0 MCG	H03	NA
apo-lamotrigine 0.5 pills of 25.0 MG	N03	N05
jamp-azithromycin 2 pills of 250.0 MG	J01	NA
metoprolol-l 3 pills of 50.0 MG	C07	NA
sandoz irbesartan 1 pill of 75.0 MG	C09	NA
apo-pregabalin 4 pills of 75.0 MG	N03	N06
apo-ramipril 1 pill of 10.0 MG, 1 pill of 5.0	C09	NA

MG		
odan bupropion sr 2 pills of 100.0 MG	N06	N07
valacyclovir tab 1 pill of 500MG	J05	NA
olanzapine 3 pills of 10 MG	N05	NA

This result was deemed adequate for proceeding, despite two issues. The string “apo-lamotrigine|0.5 pills of 25.0 MG” returned a correct best second-level ATC code of N03 but an incorrect next-best second-level ATC code of N05. For the string “apo-pregabalin|4 pills of 75.0 MG,” the first of several pregabalin doses to come, the antiepileptic nature is emphasized at the expense of the neuropathic pain aspect that the drug has come to be most associated with. An N02 best second-level ATC code would have arguably been superior.

The Inference Sample for Estimating Accuracy of AI Drug Classification

With the 20 records from the pilot sample excluded, a final sample of 200 daily drug dose strings were taken without replacement. These strings were sent to the OpenAI Batch API using GPT-4o with the settings previously described on 2024-07-10, for a total cost of \$0.33. Of the 200 daily dose string inputs, 175 (87.5%) were graded as perfect, 10 (5.0%) had a minor issue, and 15 (7.5%) had a major issue. For inference to our population of 12,294 daily drug prescription strings, the estimate of mostly correct outputs was $\hat{p} = 92.5\%$ (185/200, 80% CI 90.1-94.9).

Despite not being tuned for an oral solid database, the Google search pipeline did well. Out of the 200 drug names submitted to the algorithm, 82.5% (165/200) were deemed acceptable for use in the oral solid database while 17.5% (35/200) were not. Table 4 shows the breakout of grades from both algorithms.

Table 4. GPT-4o with prompt vs Custom Search API with regular expressions.

		Custom Search Score	
	Acceptable	Not Acceptable	Margin
LLM Score			
1 (Perfect)	148	27	175
2 (Minor Issue)	8	2	10
3 (Major Issue)	9	6	15
Margin	165	35	200

All imperfect gradings from either algorithm are presented in Table 5.

Table 5. Imperfect classifications made by either the LLM-based algorithm or the search-based algorithm

Id	LLM Input string	Search Input string	LLM best second-level ATC code	LLM next best second-level ATC code	Custom Search API extracted codes	LLM Score	Custom Search Outcome	Comments
25	sandoz folic acid 1 pill of 5.0 MG	sandoz folic acid ATC code	B03	NA	{'V03': 2, 'B03': 1, 'N02': 1, 'N01': 1}	1	Not Acceptable	This is a therapeutic dose of folic acid for

					1, 'L01': 1, 'N05': 1}			treating anemia, so B03 is correct. V03 is incorrect as returned by the search algorithm.
34	kadian sr 1 pill of 10.0 MG	Kadian sr ATC code	N02	NA	{'N02': 2, 'A07': 2}	1	Not Acceptable	Regarding A07 returned by the search algorithm, morphine is a treatment for severe diarrhea but is not relevant for this drug database.
36	jamp-mycophenolate 4 pills of 500.0 MG	jamp-mycophenolate ATC code	L04	NA	{'J01': 2, 'L04': 1, 'S01': 1, 'R03': 1, 'L01': 1, 'N06': 1}	1	Not Acceptable	The search algorithm produced J01, which is incorrect.
37	reminder for probiotics 1 pill of UNK	reminder for probiotics ATC code	A07	NA	{'A02': 1}	2	Not Acceptable	The LLM algorithm should have a next best ATC code for probiotics as there are many to choose from. A02 is not relevant for probiotics.
38	calcium mag zinc + vit d3 1 pill of UNK	calcium mag zinc + vit d3 ATC code	A12	A11	{'M05': 1, 'A11': 1}	1	Not Acceptable	Search algorithm failed to reach a consensus, does not include A12.
46	allopurinol 4 pills of 100MG	allopurinol ATC code	M04	NA	{'C02': 2, 'R01': 1, 'G01': 1, 'N06': 1}	1	Not Acceptable	The search algorithm does not include M04.
47	prednisone 2 pills of 1 mg	Prednisone ATC code	H02	NA	{'H02': 3, 'S02': 1, 'A10': 1}	2	Acceptable	LLM output should have included

					1, 'A07': 1}			A07 as a next best ATC.
54	xylac 4 pills of 10.0 MG, 2 pills of 25.0 MG	xylac ATC code	R06	NA	{'N05': 4}	3	Acceptable	The LLM algorithm incorrectly classified xylac as a treatment for allergies rather than schizophren ia.
57	hydrochlorothia zide 0.5 pills of 25.0 MG	Hydrochlorothiazide ATC Code	C03	NA	{'C03': 3, 'C09': 3}	1	Not Acceptable	The search algorithm failed to reach a consensus. C09 is a combination product with a diuretic.
59	jamp-vitamin d 1 pill of 50000.0 UNIT	jamp-vitamin d ATC code	A11	NA	{'M05': 1, 'A11': 1}	1	Not Acceptable	For the search algorithm, vitamin D fails to get an A11 majority.
60	jam-chewable vitamin c 1 pill of UNK	jam-chewable vitamin c ATC code	A11	NA	{'A11': 1, 'D03': 1, 'G04': 1}	1	Not Acceptable	For the search algorithm, vitamin C fails to get an A11 majority.
64	pms-terazosin 1 pill of 2.0 MG	pms-terazosin ATC code	C02	G04	{'G04': 2, 'N04': 2}	1	Not Acceptable	The search algorithm was unable to arrive at a consensus.
68	pms-pregabalin 2 pills of 25.0 MG	pms-pregabalin	N03	N06	{'N02': 4}	2	Acceptable	The case of pregabalin.
72	apo-naltrexone 1.5 pills of 50.0 MG	apo-naltrexone ATC code	N07	NA	{'N02': 3, 'N07': 1}	1	Not Acceptable	N02 is a combination product.
73	reminder for repatha 1 pill of UNK	reminder for repatha ATC code	NA	NA	{'C10': 1, 'A02': 1}	1	Not Acceptable	The search algorithm couldn't get a majority. The LLM output indicates correctly that this is not an oral solid,

								however, something is lost by not classifying the drug.
74	jamp famotidine 1 pill of 40.0 MG	jamp famotidine ATC code	A02	NA	{'A02': 1, 'L04': 1, 'J05': 1}	1	Not Acceptable	Search-based algorithm is unable to obtain a majority vote.
76	novo-ferrogluc 1 pill of 300.0 MG	novo-ferrogluc ATC code	B03	NA	{'B03': 1}	1	Acceptable	B03 is a good choice here given this is an iron supplement and the dose is in a range for treating anemia.
84	turmeric (qunol) capsule 1 pill of 1500.0 mg	turmeric (qunol) capsule ATC code	A16		{}	3	Acceptable	The LLM incorrectly returned ATC codes for a dietary supplement. There was also a delimiter error which resulted in the description's choice, M01, not being displayed and the second most likely code being displayed in the top position.
87	sandoz sodium bicarbonate 4 pills of 500.0 MG	sandoz sodium bicarbonate ATC code	A02	NA	{}	1	Not Acceptable	Search algorithm failed to return results.
89	compounded naltrexone capsule 3 pills of 0.25 mg	compounded naltrexone capsule ATC code	N02		{'N02': 2, 'N07': 1}	3	Not Acceptable	The description within the LLM says that it's N07, but there was a

								delimiter error.
93	euro-fer 1 pill of UNK	euro-fer ATC code	B03	NA	{'J07': 1, 'M05': 1}	1	3	The drug is marketed as an anti-anemia preparation.
95	apo-pregabalin 2 pills of 75.0 MG	apo-pregabalin ATC code	N03	N06	{'N02': 3, 'N03': 1}	2	Not Acceptable	The case of pregabalin.
96	turmeric curcumin x-strength 2 pills of UNK	turmeric curcumin x-strength ATC code	A16	NA	{'L01': 1}	3	Not Acceptable	Dietary supplement classified as drug.
97	keto 600 exogenous 2 pills of UNK	keto 600 exogenous ATC code	A16	NA	{'L02': 1, 'A14': 1, 'G03': 1}	3	Not Acceptable	Dietary supplement classified as drug.
99	vitamin b12 sl 2.5 pills of UNK	vitamin b12 sl ATC code	A11	NA	{'B03': 2}	3	Acceptable	LLM did not include B03, for anti-anemia preparations, which to be consistent is required for vitamin b12, regardless of the dose.
100	align 1 pill of 4.0 MG	Align ATC code	N06	NA	{}	3	Acceptable	The LLM description treats a probiotic as an Alzheimer's drug
112	mar-diltiazem t 1 pill of 240.0 MG	mar-diltiazem t ATC code	C08	NA	{'C08': 1, 'B01': 1}	1	Not Acceptable	No majority in the search algorithm results.
115	naltrexone hcl 0.5 pills of 50.0 MG	naltrexone hcl ATC code	N07	NA	{'V03': 3, 'N07': 2, 'N02': 1}	1	Not Acceptable	V03 is not the correct code
119	calcitriol-odan 1 pill of 0.5 MCG	calcitriol-odan ATC code	A11	NA	{'D05': 2, 'A11': 2, 'A06': 1}	2	Not Acceptable	Decision was made to require both A11 and H05 for the LLM output. D05 is correct for topical formulation and is not ideal for an

								oral solid database.
120	jamp-pregabalin 2 pills of 50.0 MG	jamp-pregabalin	N03	N06	{'N02': 4, 'N03': 1}	2	Acceptable	The case of pregabalin
126	xylac 0.5 pills of 10.0 MG	xylac ATC code	R06	NA	{'N05': 3}	3	Acceptable	Xylac treats psychosis, not allergies.
133	one-a-day womens 50+ advantage 1 pill of 0.0	one-a-day womens 50+ advantage ATC code	A11	NA	{}	1	Not Acceptable	No result from the search-based algorithm.
140	sandoz clonidine 6 pills of 0.025 MG	sandoz clonidine ATC code	C02	NA	{'S01': 3, 'N02': 2, 'C02': 1, 'N06': 1}	1	Not Acceptable	S01 is relevant for eye drops, so not wrong but not ideal for an oral solid database.
141	jamp-magnesium 1 pill of 100.0 MG	jamp-magnesium ATC code	A12	NA	{'A02': 2, 'C03': 1, 'G04': 1, 'A04': 1, 'A12': 1}	1	Not Acceptable	The search algorithm comes back with false positives.
147	sandoz colchicine 1 pill of 0.6 MG	sandoz colchicine ATC code	M04	NA	{'N02': 1, 'J01': 1, 'R03': 1, 'M01': 1, 'M04': 1, 'C10': 1}	1	Not Acceptable	The search algorithm comes back with false positives.
152	sodium chloride 2 pills of UNK	sodium chloride ATC code	A12	NA	{'B05': 3, 'A12': 1, 'R01': 1}	1	Not Acceptable	B05 would relate to infusions, which is not an ideal code for an oral solid database.
157	pregabalin 6 pills of 25.0 MG	Pregabalin ATC code	N03	N06	{'N02': 6, 'N03': 1}	2	Acceptable	The case of pregabalin.
158	d3-50 1 pill of 50000.0 UNIT	d3-50 ATC code	A11	NA	{'M05': 2, 'A11': 2}	1	Not Acceptable	D3 is an oral solid and M05 implies infusions
159	calcitriol 1 pill of .25 ug	Calcitriol ATC code	H05	NA	{'A11': 5, 'D05': 3, 'A06': 1}	2	Acceptable	Decision was made to require both A11 and H05 for the

								LLM output. A11 is accepted from the search-based algorithm.
161	l-methylfolate calcium 1 pill of 7.5 MG	l-methylfolate calcium ATC code	B03	NA	{'V03': 3, 'B03': 1, 'D11': 1, 'N02': 1}	2	Not Acceptable	B03 is correct as this is a form of folate and clearly in the therapeutic dose range.
162	reminder for vitamin b12 1 pill of UNK	reminder for vitamin b12 ATC code	B03	NA	{'M05': 1}	1	Not Acceptable	A decision was made to accept B03 for vitamin b12, regardless of the dose.
169	apo-pregabalin 2 pills of 150.0 MG, 4 pills of 50.0 MG	apo-pregabalin ATC code	N03	N06	{'N04': 7}	2	Acceptable	The case of pregabalin.
172	mya 1 pill of 0.0	mya ATC code	NA	NA	{}	3	Not Acceptable	MYA is a Canadian birth control pill.
178	concerta 2 pills of 54.0 MG	concerta N07	N07		{'N06': 7}	3	Acceptable	There was a delimiter error made in the LLM algorithm. The description does have N06 as the best code.
179	pms-pregabalin 2 pills of 150.0 MG	pms-pregabalin ATC code	N03	N06	{'N02': 4}	2	Acceptable	The case of pregabalin.
187	apo-lithium carbonate 3 pills of 150.0 MG	apo-lithium carbonate ATC code	N05	NA	{'N02': 2, 'C01': 1, 'N05': 1}	1	Not Acceptable	A clear case of treatment for bipolar.
191	lyrica 8 pills of 75.0 MG	Lyrica ATC code	N03	N06	{'N02': 6, 'N03': 2}	2	Acceptable	The case of pregabalin.
194	webber omega 3-6-9 1 pill of 1200.0 mg	webber omega 3-6-9 ATC code	A11	NA	{}	3	Not Acceptable	The best code should have been C10.
195	statex 4 pills of 5.0 MG	statex ATC code	C10	NA	{'N02': 1}	3	Acceptable	This is a brand name for morphine.

201	b12 1 pill of UNK	b12 ATC code	A11	NA	{'B03': 4}	1	Acceptable	The decision was made to require B03 for vitamin b12 regardless of the dose.
205	pms-methotrexate 5 pills of 2.5 MG	pms-methotrexate ATC code	L01	M01	{'L04': 5, 'L01': 1}	3	Acceptable	L04 would be the correct best ATC code for an oral solid, though it's important to note that methotrexate is often given in a single weekly dosage, so this shows a limitation of the prompt formulation.
206	hanzema 1 pill of 30.0 MG	hanzema ATC code	A02	NA	{'D05': 2, 'D11': 2, 'L01': 2}	3	Not Acceptable	The correct answer was D11, as hanzema is indicated for the treatment of severe chronic hand eczema. The search-based algorithm failed to reach a majority.
216	vitamin c timed release 2 pills of UNK	vitamin c timed release ATC code	A11	NA	{'A11': 1, 'N06': 1, 'M03': 1}	1	Not Acceptable	Search-based algorithm failed to reach a majority.
220	calcium magnesium with zinc 3 pills of 0.0	calcium magnesium with zinc ATC code	A12	NA	{'B05': 2, 'A12': 2, 'A02': 2}	1	Not Acceptable	Search-based algorithm failed to reach a majority.

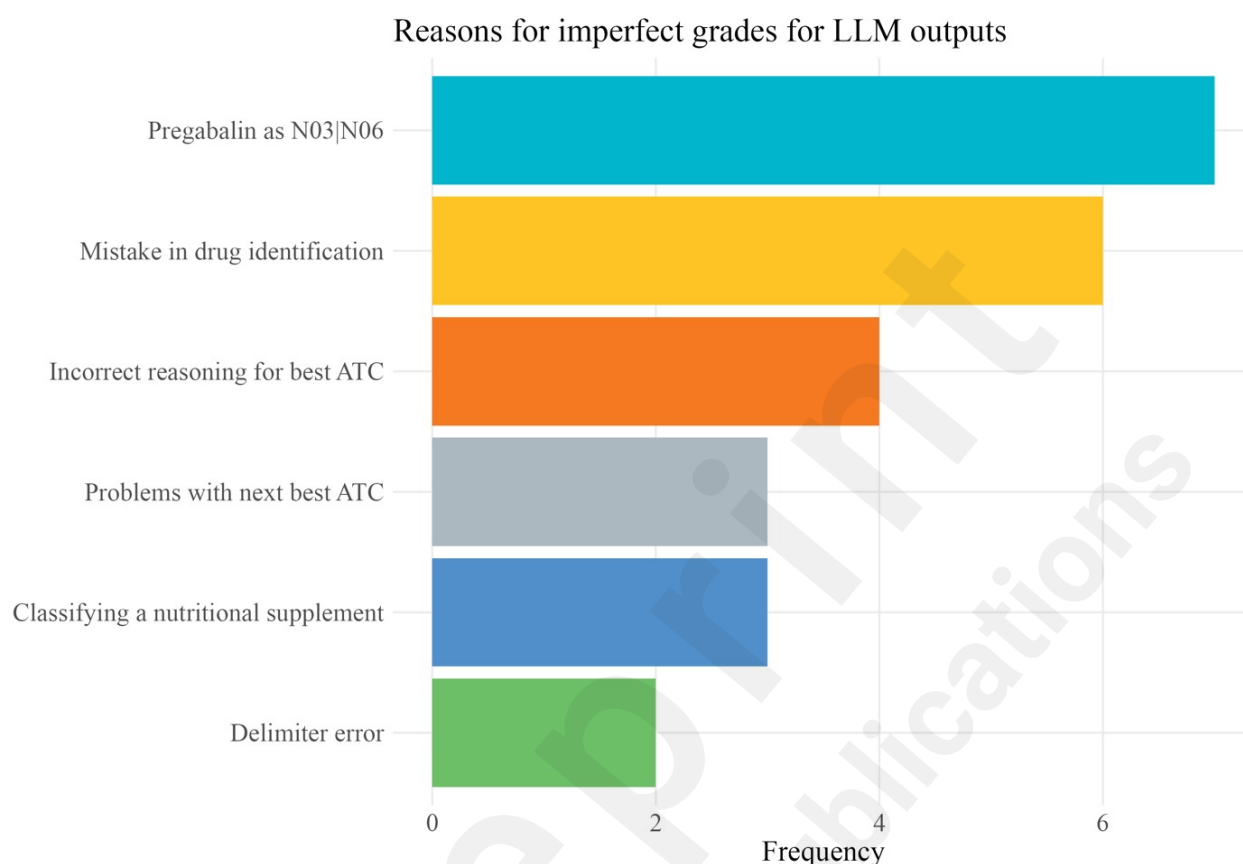


Figure 3. Failure modes observed by AI drug classification within the inferential sample.

When discussing the imperfect grades from the LLM algorithm, pregabalin was the most frequent reason, appearing seven total times in the inference sample under different brand names and different dosing configurations. In each case, the LLM's output returned N03 as a best second-level ATC code followed by N06. The search-based algorithm consistently returned N02, which aligns with pregabalin's most well recognized role of treating neuropathic pain. However, pregabalin was developed as an antiepileptic, is still used in its antiepileptic capacity and well as a treatment for generalized anxiety disorder, so the choice was made to consider the LLM's outputs for pregabalin as more correct than incorrect.

The largest category of serious errors is where the LLM either did not recognize a real drug or reasoned about the wrong drug. An example of the former was MYA, a Canadian birth control oral solid [41] that GPT-4o was unable to retrieve information about. Interactive prompt modifications in the OpenAI playground could not overcome this. In a second case, the Canadian morphine drug STATEX [42] was reasoned about as a statin, presumably due to the lexical similarity. However, this mistake could not be replicated within the OpenAI playground.

Reasoning about vitamins that might be anemia treatments caused two grade 3 errors relating to the best second-level ATC. The decision was made to require B03 as the best second-level ATC code for Vitamin b12 supplements if the dose was high or if the dose was unknown, and in these two cases Vitamin b12 supplements were classified primarily as A11 (Vitamins). A11 was allowed as a next best code if it was provided, but it was not required.

The Vitamin D analogue calcitriol proved to be another noteworthy case in the vitamins domain. While Vitamin D analogs explicitly fall under A11, calcitriol is no ordinary vitamin, and its therapeutic use case better falls under H05 (calcium homeostasis). The decision was made to require

both A11 and H05 for full credit, but to give mostly correct status if either one was present. In the two times that calcitriol appeared, only A11 was present in one case and only H05 in another, so these were graded as more correct than incorrect.

The reasoning in the chain-of-thought responses was not without cost. The extra output costs money, and there's additional likelihood of a delimiter error because the algorithm needs to stop reasoning and add the pipe delimiters and ATC codes. Two times out of the 200, delimiter errors were a primary cause of a mostly incorrect grade.

The LLM respected the oral solid criteria, sometimes to a fault. In a favorable case, azithromycin was classified as J01 when it would have been classified as S01 if packaged in liquid form as eye drops. However, the search-based algorithm also returned J01, without any oral solid prompting. A case that was not ideal was a reminder pouch for Repatha, an injectable, where the LLM only returned NAs even though Repatha itself is easily classified as C10. This was not considered incorrect because the prompt was to specifically asked to return NAs when drugs couldn't be classified as oral solids.

During the grading process, difficulties with the problem formulation of best versus next best ATC code became apparent in ways that were not obvious from the motivating example of aspirin. Terazosin, which can be used as an antihypertensive or a urological in oral solid form and doses overlap, was well served by a C02|G04 output. A calcium, magnesium, zinc plus vitamin d3 was well served by an A12|A11 output. But in other cases, things were less clear. Pregabalin could have at least three ATC codes. For the case of calcitriol, there are only two relevant ATC codes but the order of best versus next best is not obvious.

The drug methotrexate illuminated a limitation of the daily dose formulation of the prompt. The string input was "pms-methotrexate|5 pills of 2.5 MG" which suggests a 12.5 mg daily dose, however this medication is often dosed once per week. This output was graded as mostly incorrect as it returned L01 instead of L04, whereas the former would more typically be associated with high dose infusions for cancer treatment.

After focusing the first part of this research on chain-of-thought reasoning about drug dose, it's natural to question whether this was worth the additional prompt complexity and number of prompts. While the reasoning would often mention "at this dose," or mention that the dose fell within a common range of prescriptions, in the end, the results mostly aligned with the search-based algorithm. The cases where the dose mattered were antianemic preparations such as high dose folic acid (a case that the search-based algorithm missed) and high dose folate. It also helped to give prednisone an H02 systemic use classification as opposed to a higher dosage A07 classification for inflammatory bowel disease. Notably, low dose aspirin did show up in the inferential sample, but the drug name was "aspirin low dose" and the search-based algorithm was easily able reach a majority vote of B01.

Discussion

Principal Findings

Large language models such GPT-4o perform well in the drug classification task. In the hand-picked example of aspirin, GPT-4o was able to distinguish between two therapeutic uses based on the dose, which happened only after incorporating a chain-of-thought prompting technique. This prompt, when applied to a larger sample, was deemed mostly correct a vast majority of the time (n=200, 92.5%).

Google Search, which can also be automated by the Custom Search API, also does well with extracting ATC codes when combined with a simple pipeline using regular expressions and voting. While the proportion of appropriate search-algorithm responses (n=200, 82.5%) was somewhat lower than the at least mostly correct proportion from the LLM, the algorithm was not specifically tuned to the task of coding oral solids. Different search terms could be tried, more pages returned, etcetera.

However, the question is not whether Google Search could be tuned to outperform an LLM, because with enough time and energy, the answer is certainly on the internet somewhere. One place where such answers show up is a publicly accessible version of DrugBank, a proprietary source of drug information, that is indexed by Google Search and often shows up in the summary snippets returned by the Custom Search API. Along with Google's own restriction on persistent storage of results, automatic scraping of such data represents multiple terms of service violations. While the enforceability is questionable, why incur the risk when LLMs perform as well as they do and are meant to extract information for the user to keep for as long as is necessary?

Additionally, Google Search, while quite reproducible in the short term due to caching, has been arguably degrading in quality for years [43] and its long-term reproducibility is highly uncertain. Prediction algorithms built upon Google Search have spectacularly failed when the underlying search infrastructure changed, as in the infamous case of Google Flu [44].

Reproducibility for versioned LLMs with temperature parameters like GPT-4o should in principle be perfect as setting the temperature parameter to 0 should lead to deterministic output (at least when there are no probability ties between tokens). However, this is not true in practice [45]. Differences between the Batch API and OpenAI playground that were sometimes material, and this has been experienced by other users. One user of the OpenAI Batch API commented that batching changed the behavior of an OpenAI LLM, making the results "too similar" [46]. At the time of writing, there is an experimental feature from OpenAI that includes a seed parameter and system fingerprint, which may be helpful in achieving perfect reproducibility.

Despite concerns with reliability in this research, GPT-4o's reliability has been praised in medical contexts as higher than alternatives. In a study on extraction and summarization of Japanese-language clinical research protocols, GPT-4o was said to exhibit "high reproducibility," with 80% and 100% accuracy for research objectives and research designs [47]. In an information extraction task on veterinary electronic health records, GPT-4o demonstrated greater reproducibility than human pairs, with an average Cohen's kappa of 0.98 vs 0.8 for humans [48].

For \$0.33 USD, 200 daily drug prescription strings were classified into best and next-best second-level ATC codes with the reasoning provided. That would correspond to spending just over \$20.00 USD to classify all 12,294 daily drug prescription strings in the SHS drug database. The affordability of GPT-4o as a drug categorization tool signifies a democratization of research instruments in pharmacoepidemiology, as the inexpensive use of GPT-4o, especially via the OpenAI Batch API, means that drug classification can be accommodated by virtually any research budget. Prompts can be shared within the research community, and the richness of drug data sets will be enhanced for teams worldwide. Furthermore, as competition increases, capabilities are likely to improve while prices fall. An example of this is GPT-4o itself, which at the time of release was twice as fast and half the price of GPT-4 Turbo [49]. This cost-effectiveness democratizes access to advanced drug classification tools and, as such, promotes equitable research opportunities.

Limitations and Future Work

This study involved a specific population of daily drug prescriptions from polychronic patients in a value-based care organization residing in the US and Canada. Geographically, the patients were biased towards Canada with some US representation. Future work is needed to replicate drug classification in other settings. Additionally, one of the motivations for moving to LLMs was that they could accommodate drugs in non-English speaking countries that use operational codes other than DIN and NDC. While GPT-4o boasts of "significant improvement on text in non-English languages" [49], such classification was out of scope for this research. Future studies are needed to test the global generalizability of this approach.

The manual grading used in this paper has the potential for bias and errors. The grading process was tedious and represented 220 separate research investigations rather than a lookup. It's likely that contradictions remain, however all imperfect grades were documented in the Results section. In the

future, more formal methods such as the Delphi method [50] could add additional rigor.

The problem formulation of best and next best second-level ATC codes, based on daily dosage strings, was occasionally not ideal for the drugs and doses encountered. There were more than two codes, times where ordering was not clear, questionable relevance of a next best code, and weekly dosing schedules. Other formulations could be considered, such as those that return true or false for a set of codes, potentially not based on dose at all. Ignoring dose would simplify the problem much further. In addition to alleviating the daily vs weekly frequency complexity, the multiplicity of inputs is cut down to a fraction of what was encountered in this research as only the drug name is the input. Complete and total reproducibility of GPT-4o outputs was not possible at the time of writing. New features such as system fingerprints and seeds may address this and is an important topic for future research. One prompting technique to deal with a multiplicity of possible outputs is to sample them from the LLM with temperature set above zero and use a voting process. This is the essence of the “self-consistency” prompting approach [46]. Additionally, with powerful open-source LLMs becoming available such as the Llama family of models [51], there is the question of whether full reproducibility is achievable when the researcher is running the model locally. If so, is there a reduction in the quality of the drug classification when using an open-source model that may pose a tradeoff to reproducibility?

Outside of reproducibility, more research on the benefits of open-source LLMs for drug classification and related tasks is needed. For instance, open-source models run locally have the privacy advantage of not needing to have data leave any internal network, but could open up other privacy risks. The economics of using an open-source LLM versus proprietary LLMs via APIs is also unclear. This research was limited to one LLM, the proprietary GPT-4o (specifically the version “gpt-4o-2024-05-13”).

New large language models are arriving regularly, with new capabilities. Future research will be needed to evaluate the drug classification capabilities of the next generation of the next generation of models. Continuing assessment of the emergence of new AI models will enable the use of most-effective and updated tools for pharmacoepidemiologic research.

Conclusions

This research demonstrates that GPT-4o is a powerful and accessible tool for enhancing pharmacoepidemiologic research by automating drug classification. GPT-4o and large language models in general represent an inexpensive and straightforward method for augmenting real-world drug databases with Anatomical Therapeutic Chemical drug classes. This gives nearly all research teams access to a powerful tool to satisfy a key prerequisite of pharmacoepidemiologic analysis using data from electronic health records, pharmacy management systems, and claims records. It is not just a matter of increased efficiency but also democratizing access to high-quality pharmacoepidemiologic drug classification instruments. Better, more accessible drug information is a precursor to higher quality and quantity of pharmacoepidemiologic data sets and a path towards better drug prescription policy and clinical outcomes for patients across the world.

Acknowledgements

This work was funded by Spencer Health Solutions, Inc., a private corporation headquartered in Morrisville, North Carolina.

Authors' Contributions

All authors participated in the conceptualization, review and revision of the final manuscript draft for submission. IRRC provided expert guidance on the performance grading. BAO and IRRC proposed the methodology, assessed performance of the algorithms, and wrote the original draft. BAO performed the statistical analyses. ESF and TPR supervised the research through all stages and

performed validation activities.

Conflicts of Interest

The authors are all current or former employees or contractors of Spencer Health Solutions, Inc., the developer of the spencer platform.

Abbreviations

A11: Vitamins (a second-level ATC code)

AI: Artificial Intelligence

API: Application Programming Interface

ATC: Anatomic Therapeutic Chemical classification system

B01: Antithrombotic Agents (a second-level ATC code)

DIN: Drug Identification Number

GPT: Generative Pre-trained Transformer, a family of large language models that includes GPT-4o

LLM: Large Language Model

N02: Analgesics (a second-level ATC code)

NDCs: National Drugs Code

MG: Uppercase of “mg” abbreviation, for milligrams

NLP: Natural Language Processing

SHS: Spencer Health Solutions, Inc.

References

1. Bérard A. Pharmacoepidemiology Research-Real-World Evidence for Decision Making. *Front Pharmacol* 2021 Sep 7;12. PMID:34557096
2. Reisinger SJ, Ryan PB, O'Hara DJ, Powell GE, Painter JL, Pattishall EN, Morris JA. Development and evaluation of a common data model enabling active drug safety surveillance using disparate healthcare databases. *Journal of the American Medical Informatics Association* 2010 Nov 1;17(6):652–662. PMID:20962127
3. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005 Apr;58(4):323–337. doi: 10.1016/j.jclinepi.2004.10.012
4. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong ICK, Rijnbeek PR, van der Lei J, Pratt N, Norén GN, Li Y-C, Stang PE, Madigan D, Ryan PB. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015;216:574–8. PMID:26262116
5. Platt R, Wilson M, Chan KA, Benner JS, Marchibroda J, McClellan M. The New Sentinel Network — Improving the Evidence of Medical-Product Safety. *New England Journal of Medicine* 2009 Aug 13;361(7):645–647. PMID:19635947
6. Gini R, Schuemie M, Brown J, Ryan P, Vacchi E, Coppola M, Cazzola W, Coloma P, Berni R, Diallo G, Oliveira JL, Avillach P, Trifirò G, Rijnbeek P, Bellentani M, Van Der Lei J, Klazinga N, Sturkenboom M. Data Extraction And Management In Networks Of Observational Health Care Databases For Scientific Research: A Comparison Among EU-ADR, OMOP, Mini-Sentinel And MATRICE Strategies. *eGEMs (Generating Evidence & Methods to improve patient outcomes)* 2016 Aug 2;4(1):2. PMID:27014709
7. Schneeweiss S, Avorn J. A review of uses of health care utilization databases for epidemiologic research on therapeutics. *J Clin Epidemiol* 2005 Apr;58(4):323–337. PMID:15862718
8. Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, Wu J, Winter C, Hesse C, Chen M, Sigler E, Litwin M, Gray S, Chess B, Clark J, Berner C, McCandlish S, Radford A, Sutskever I, Amodei D. Language Models are Few-Shot Learners. 2020 May 28;

9. Sabaté M, Montané E. Pharmacoepidemiology: An Overview. *J Clin Med* 2023 Nov 10;12(22):7033. PMID:38002647
10. Williams N, Rudolph KE. A drug classification pipeline for Medicaid claims using RxNorm. 2024 Apr 1;
11. Applications of the ATC/DDD methodology. Available from: <https://www.who.int/tools/atc-ddd-toolkit/applications-methodology> [accessed Jun 23, 2024]
12. WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD Methodology: History. 2018. Available from: https://atcddd.fhi.no/atc_ddd_methodology/history/ [accessed Jul 16, 2024]
13. World Health Organization. Anatomical Therapeutic Chemical (ATC) Classification. Available from: <https://www.who.int/tools/atc-ddd-toolkit/atc-classification> [accessed Jul 16, 2024]
14. WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD Index: C09BA03. 2014. Available from: https://atcddd.fhi.no/atc_ddd_index/?code=C09BA03 [accessed Jul 16, 2024]
15. WHO Collaborating Centre for Drug Statistics Methodology. Purpose of the ATC/DDD system. 2018. Available from: https://atcddd.fhi.no/atc_ddd_methodology/purpose_of_the_atc_ddd_system/ [accessed Jul 16, 2024]
16. Simonaitis L, McDonald CJ. Using National Drug Codes and drug knowledge bases to organize prescription records from multiple sources. *American Journal of Health-System Pharmacy* 2009 Oct 1;66(19):1743–1753. PMID:19767382
17. Drug Identification Number (DIN). Available from: <https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/fact-sheets/drug-identification-number.html> [accessed Jun 23, 2024]
18. Z-Index. About Z-Index. Available from: <https://www.z-index.nl/english> [accessed Jul 18, 2024]
19. WHO Collaborating Centre for Drug Statistics Methodology. ATC/DDD Index. Available from: https://atcddd.fhi.no/atc_ddd_index/ [accessed Dec 9, 2024]
20. Auer S, Bizer C, Kobilarov G, Lehmann J, Cyganiak R, Ives Z. DBpedia: A Nucleus for a Web of Open Data. 2007. p. 722–735. doi: 10.1007/978-3-540-76298-0_52
21. DBpedia. DBpedia Lookup - Generic RDF Indexer & Searcher. Available from: <https://lookup.dbpedia.org/> [accessed Dec 9, 2024]
22. Liu SWMRMG and SNelson. RxNorm: prescription for electronic drug information exchange. *IT Prof* 2005;17–23.
23. United States National Library of Medicine. Rxclass. Available from: <https://mor.nlm.nih.gov/RxClass/> [accessed Jun 23, 2024]
24. Drug Product Database: Access the extracts. 2020. Available from: <https://www.canada.ca/en/health-canada/services/drugs-health-products/drug-products/drug-product-database/extracts.html> [accessed Apr 30, 2022]
25. Kellmann AJ, Lanting P, Franke L, van Enckevort EJ, Swertz MA. Semi-automatic translation of medicine usage data (in Dutch, free-text) from Lifelines COVID-19 questionnaires to ATC codes. *Database* 2023 Apr 26;2023. PMID:37114804
26. Croset SRH and DR-Schuhmann. Integration of the anatomical therapeutic chemical classification system and drugbank using owl and text-mining. *GI Workgroup Ontologies in Biomedicine and Life Sciences (OBML)* 2012;
27. Knox C, Wilson M, Klinger CM, Franklin M, Oler E, Wilson A, Pon A, Cox J, Chin NE (Lucy), Strawbridge SA, Garcia-Patino M, Kruger R, Sivakumaran A, Sanford S, Doshi R, Khetarpal N, Fatokun O, Doucet D, Zubkowski A, Rayat DY, Jackson H, Harford K, Anjum A, Zakir M, Wang F, Tian S, Lee B, Liigand J, Peters H, Wang RQ (Rachel), Nguyen T, So D, Sharp M, da Silva R, Gabriel C, Scantlebury J, Jasinski M, Ackerman D, Jewison T, Sajed T, Gautam V, Wishart DS. DrugBank 6.0: the DrugBank Knowledgebase for 2024. *Nucleic Acids Res* 2024 Jan 5;52(D1):D1265–D1275. PMID:37953279
28. DrugBank Online About Page. Available from: <https://go.drugbank.com/about> [accessed Jun 13, 2024]

29. Google. Custom Search JSON API. Available from: <https://developers.google.com/custom-search/v1/overview> [accessed Dec 9, 2024]
30. hiQ Labs, Inc. v. LinkedIn Corp., 273 F. Supp. 3d 1099 (N.D. Cal. 2017).
31. Google. Programmable Search Engine Terms of Service. Available from: Programmable Search Engine Terms of Service [accessed Dec 9, 2024]
32. Liu H-W. Two decades of laws and practice around screen scraping in the common law world and its open banking watershed moment. *Wash Int'l LJ* 2020;30:28.
33. Patel T, Ivo J, Pitre T, Faisal S, Antunes K, Oda K. An In-Home Medication Dispensing System to Support Medication Adherence for Patients With Chronic Conditions in the Community Setting: Prospective Observational Pilot Study. *JMIR Form Res* 2022 May 19;6(5):e34906. PMID:35587371
34. GPT-4o. Available from: <https://platform.openai.com/docs/models/gpt-4o> [accessed Jun 24, 2024]
35. Open AI. Batch API. Available from: <https://platform.openai.com/docs/guides/batch> [accessed Jul 1, 2024]
36. OpenAI. How should I set the temperature parameter? Available from: <https://platform.openai.com/docs/guides/text-generation/how-should-i-set-the-temperature-parameter> [accessed Jul 28, 2024]
37. OpenAI. max_tokens. Available from: https://platform.openai.com/docs/api-reference/chat/create#chat-create-max_tokens [accessed Jul 28, 2024]
38. OpenAI. What are tokens and how to count them? Available from: <https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them> [accessed Jul 28, 2024]
39. Lohr SL. Sampling: design and analysis. Chapman and Hall/CRC; 2021.
40. Cocks K, Torgerson DJ. Sample size calculations for pilot randomized trials: a confidence interval approach. *J Clin Epidemiol* 2013 Feb;66(2):197–201. doi: 10.1016/j.jclinepi.2012.09.002
41. APOTEX INC. Mya - PRODUCT MONOGRAPH. 2020.
42. Paladin Labs Inc. STATEX®. 2019 Aug. Available from: https://www.paladin-pharma.com/our_products/Statex_en.pdf [accessed Aug 5, 2024]
43. Bevendorff J, Wiegmann M, Potthast M, Stein B. Is Google Getting Worse? A Longitudinal Investigation of SEO Spam in Search Engines. 2024. p. 56–71. doi: 10.1007/978-3-031-56063-7_4
44. Lazer D, Kennedy R, King G, Vespignani A. The Parable of Google Flu: Traps in Big Data Analysis. *Science* (1979) 2014 Mar 14;343(6176):1203–1205. doi: 10.1126/science.1248506
45. OpenAI Community. Why the API output is inconsistent even after the temperature is set to 0. Available from: <https://community.openai.com/t/why-the-api-output-is-inconsistent-even-after-the-temperature-is-set-to-0/329541> [accessed Jul 24, 2024]
46. Wang X, Wei J, Schuurmans D, Le Q, Chi E, Narang S, Chowdhery A, Zhou D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. 2022 Mar 21;
47. Fukataki Y, Wakako H, Naoki N, Ito YM. Ethical review of clinical research with generative AI: Evaluating ChatGPT's accuracy and reproducibility. 2024. doi: 10.1101/2024.11.19.24317555
48. Wulcan JM, Jacques KL, Lee MA, Kovacs SL, Dausend N, Prince LE, Wulcan J, Marsilio S, Keller SM. Classification performance and reproducibility of GPT-4 omni for information extraction from veterinary electronic health records. 2024 Sep 9;
49. OpenAI. Hello GPT-4o. 2024. Available from: <https://openai.com/index/hello-gpt-4o/> [accessed Jul 24, 2024]
50. Dalkey N, Helmer O. An Experimental Application of the DELPHI Method to the Use of Experts. *Manage Sci* 1963 Apr;9(3):458–467. doi: 10.1287/mnsc.9.3.458
51. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and Efficient Foundation Language Models. 2023 Feb 27;



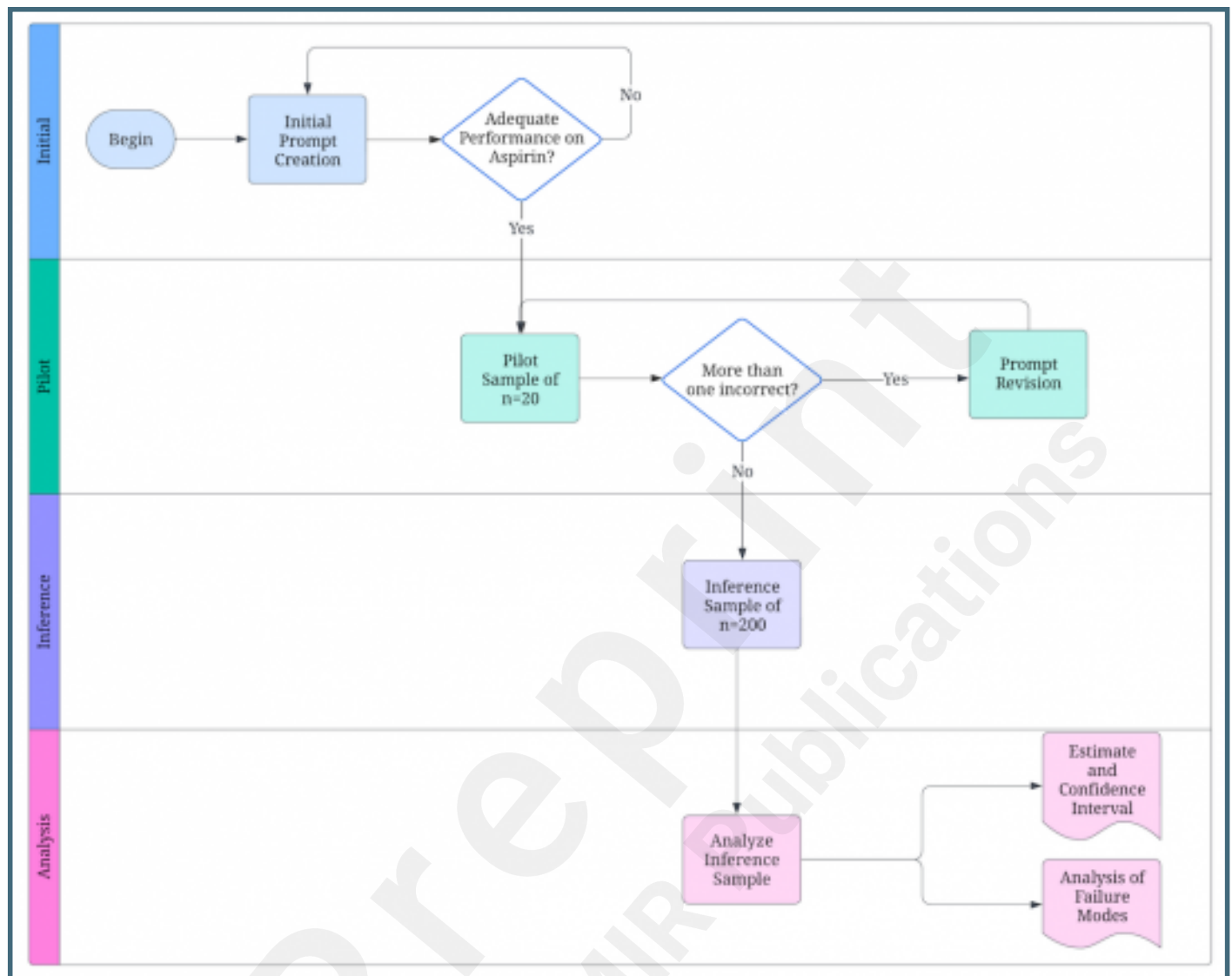
Supplementary Files

Figures

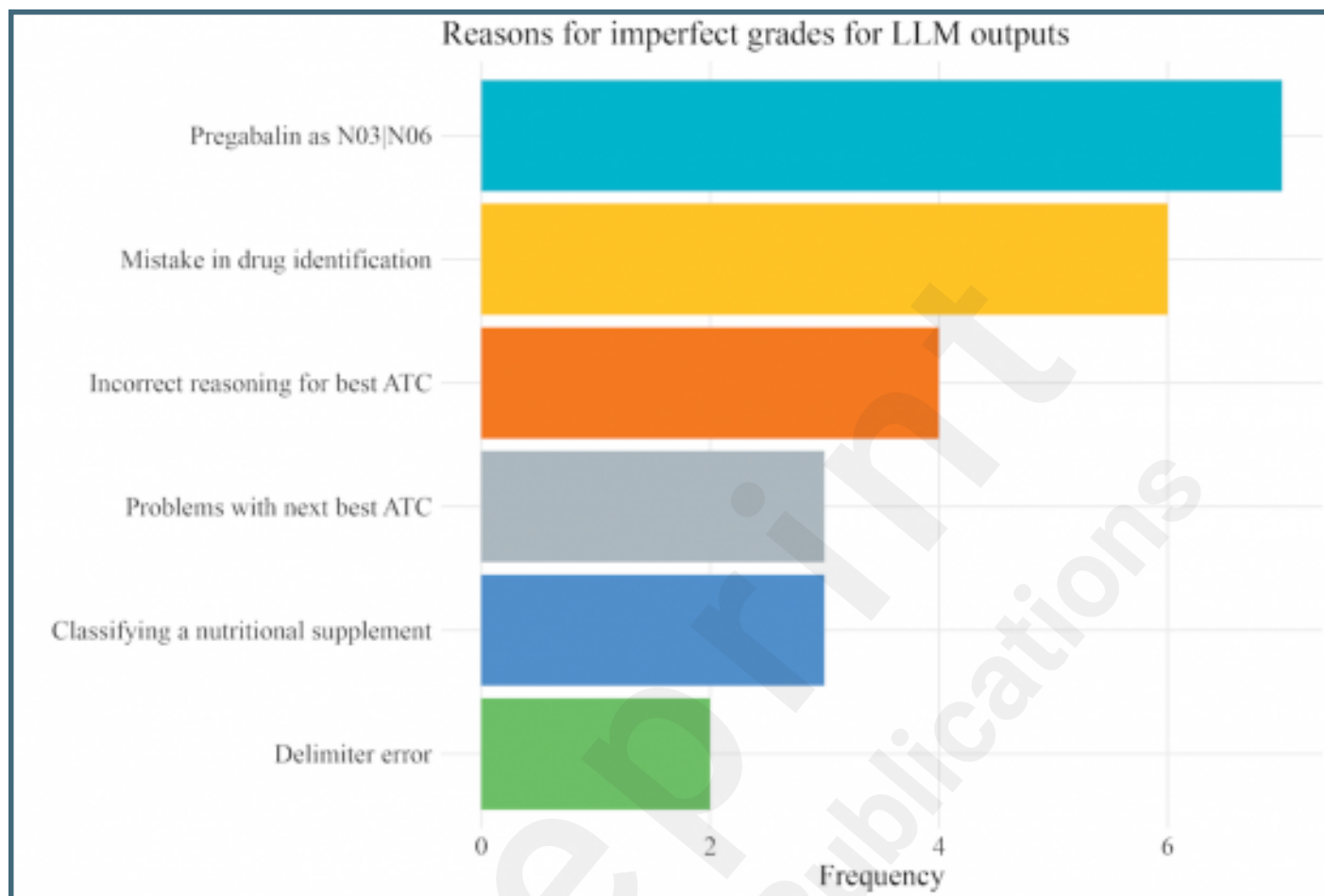
Key components of the spencer smart medication dispenser.



The methodology used to develop and test an ATC classification AI prompt.



Failure modes observed by AI drug classification within the inferential sample.



TOC/Feature image for homepages

AI (DALL-E) generated image meant to depict AI-assisted drug classification. For final publication we would like to replace this with an image made by a human graphic designer.

