Generative Terminology Mapping

Scaling Medication Text String to RxNorm Conversion in Billion-scale EHR Data

December 2023

Table of Contents

Table of Contents 2
Introduction: Atropos Health 3
Problem Definition 4
  Research Questions 5
Methods 6
  Preparing Medication Source Terms 6
  Approach 1: Named Entity Recognition by Healthcare-Specific NLP Model 7
  Approach 2: Unstructured Text to RxNorm via UMLS API 7
  Approach 3: Generative Terminology Mapping - UMLS API Plus GPT-4 7
Human Expert Review and Ground Truth 9
Results 11
  Summary of Metrics for Different Approaches 11
    Ingredient Correctness Mapping Outcomes 11
    Other Mapping Outcomes 11
  Approach 1: Janus-Derived Maps 13
    Janus AUROC Curves for Different Mapping Outcomes 14
  Approach 2: UMLS API Alone 15
    UMLS API Alone Correctness Metrics 15
  Approach 3: Generative Terminology Mapping 17
   Generative Terminology Mapping Results: 90% Reduction in Errors, 91% Coverage, 98% Reduced Cost 18
    Generative Terminology Mapping (UMLS API Plus GPT-4) Correctness Metrics 18
    Generative Terminology Mapping vs. Ground Truth Results and Confusion Matrix 18
  Data Engineering Challenges 19
  Estimated Cost and Comparisons 20
Conclusion 21
  Generative Terminology Mapping Produces Research-Grade RxNorm-Mapped Medication Data at Scale 21
  Caveats and Future Directions 23
Appendix 24

©Copyright Atropos Health, 2023. All rights reserved.

Generative Terminology Mapping: Scaling Medication Text String to RxNorm Conversion in Billion-scale EHR Data

December 19, 2023

Philip Ballentine, MSc, Senior Director, Data Engineering, Atropos Health

C. William Pike, MD, Medical Director, Atropos Health

Introduction: Atropos Health

Atropos Health[1] is the developer of GENEVA OS™ (Generative Evidence Acceleration Operating System), the operating system for rapid healthcare evidence across a vast network of Real-World Data (RWD). Health systems and life science companies work with Atropos to close evidence gaps from bench to bedside, improving individual patient outcomes with data-driven care, expediting research that advances the field of medicine, and more. Our solution offerings are based on many peer-reviewed publications, thousands of active users over the past decade, and on-staff clinical expertise. We aim to transform healthcare with timely, relevant Real-World Evidence (RWE). To do this, Atropos' Data Engineering team invests significantly in approaches to mapping raw EHR data–often text strings— into standardized terminologies. Applying a pragmatic engineering approach to this work led us to investigating different methods to accelerate and improve these mappings, and to rigorously evaluate these methods for accuracy and reliability.

---

[1] https://www.atroposhealth.com/who-we-are

©Copyright Atropos Health, 2023. All rights reserved.

Problem Definition

As part of GENEVA OS™, Atropos Health utilizes a variety of datasets derived from EHR data, claims data, and other sources. One of these datasets (termed Eos for the purposes of this white paper) consists of de-identified structured data from more than 130 million patients from across the United States collected. As of mid-2023, Eos contained 4.9 billion records of medication orders and administrations data, only 35% of which had any medication terminology code attached (principally National Drug Codes, or NDCs). The remaining 65% of records identify medications with variable text strings such as "TRAZODONE HCL TABLET."

To facilitate research use cases, we need to map this data to a standard terminology, and Atropos Health selected RxNorm as the best choice, since it is 1) free to use; 2) covers medication products dispensed in the United States; and 3) has a built in knowledge graph that allows the relation of RxNorm concepts to other concepts programmatically. For example, RxNorm allows a user to relate the concept for a particular ingredient to all concepts for medications that contain that ingredient. RxNorm is frequently updated and has significant adoption in observational health data research.

Out of ~3.1 billion rows with no mapped medication term in *Eos*, we identified 95,543 distinct medication source terms–strings that represented a medication ordered for or administered to patients– to map. We estimated that a skilled terminologist might be able to map 150 of these terms per hour. Therefore, to fully map this data by hand, we could predict at least 633 hours of work time, which might cost approximately $200,000.

Furthermore, *Eos* is updated quarterly, and these updates add new terms as more source systems are added to the dataset and new medication records are created. Terms are also added when *Eos* or upstream sources apply their own mappings to the data, which can cause many previously established maps to unpredictably "break" and introduce high volumes of new terms. Finally, new RxNorm terms are created and old terms deprecated, which may necessitate remapping of the source term to a new RxNorm code.

Due to these factors, we concluded that an automated approach has many advantages to make this data fit for research use, and explored several candidate approaches. By comparing their outputs to that of an expert informaticist mapping the terms by hand, we were able to draw conclusions about their quality and reliability. Our proposed solution achieved >99% accuracy on the metric of "ingredient correctness" using an open-source, free, and public API in combination with a Large Language Mode (LLM) at 98% lower cost and shorter timeline than a human expert. We believe that this approach, which combines existing tools and knowledge graphs with an off-the-shelf use of a Large Language Model (LLM) such as GPT-4—which we propose as "generative terminology mapping"—has applications across many datasets.

©Copyright Atropos Health, 2023. All rights reserved.

Research Questions

We sought to answer the following questions through this project, regarding mapping of medication terms to RxNorm:

1. Are any of our automated mapping approaches (*Janus*, UMLS API, Generative AI), alone or in combination, accurate enough to identify optimal RxNorm mappings in an unsupervised medication mapping process?
2. What is the expected error rate of these approaches? How significant are the errors made by these automated mapping methods?
3. Can we create an efficient, cost-effective, automated approach to mapping text strings in *Eos* to RxNorm that approximates the accuracy of a human expert informaticist mapper and does not introduce unacceptable errors into our data?

©Copyright Atropos Health, 2023. All rights reserved.

Methods

Preparing Medication Source Terms

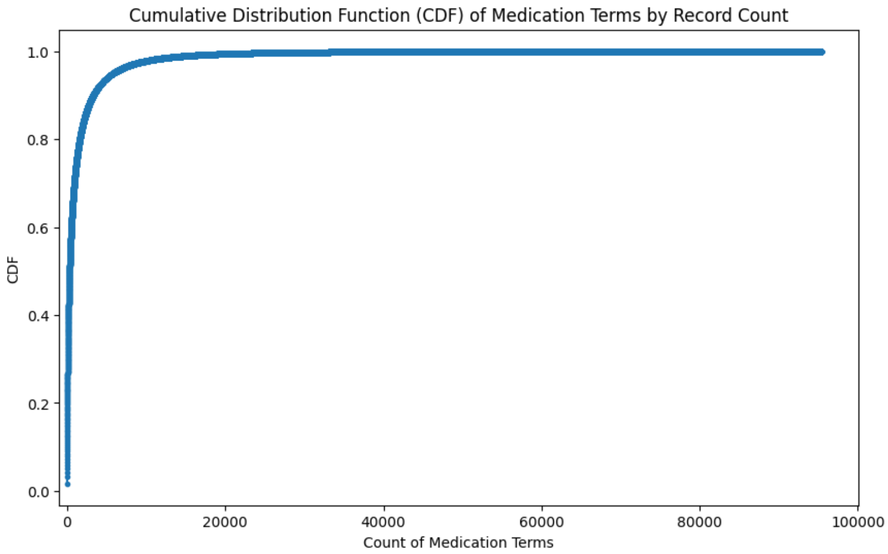
To facilitate mapping this data to RxNorm, we extracted distinct medication from the *Eos* data terms with the following:

```
SELECT CONCAT_WS(" ",COALESCE(COALESCE(GEN_NAME,
MED_NAME),""),COALESCE(MED_STRENGTH,""),COALESCE(UNIT,""),COALESC
E(MED_FORM,""),COALESCE(MED_ROUTE,""))
```

Concatenating columns (some of which can be NULL) was necessary here due to the structure of the Eos data; since it comes from many sources across the United States, there is no guaranteed single column that contains the relevant information 100% of the time.

Roughly 35% of the medication records in *Eos* contained a valid National Drug Code (NDC), which can be deterministically mapped to RxNorm through existing knowledge graphs. This left roughly 65% of the medication data–3.1 billion records–with an unmapped medication term in the form of a text string. This approach generated ~95,000 unique source medication terms (string representations of medications). These exhibited an extreme skewed density plot where a small number of terms comprised the majority of the rows analyzed, which is expected in cases like this. In fact, the top 1,000 terms by incidence count comprised ~51% of the rows in the Eos' unmapped medication data–roughly 1.6 billion records.

This high skew made it feasible to map 50% of the unmapped data in approximately 10 hours from our human expert mapper, providing a "gold standard" for each source term. To evaluate different approaches, we used this gold standard to evaluate several distinct approaches to mapping.



©Copyright Atropos Health, 2023. All rights reserved.

Approach 1: Named Entity Recognition by Healthcare-Specific NLP Model

We engaged the services of an industry-leading healthcare-specific NLP company (pseudonym *Janus*) to use a NLP model specifically meant to resolve unstructured medication name text to RxNorm codes. When the model generated multiple candidate mappings for a given source term, we selected the term with the highest confidence score generated by the model.

Approach 2: Unstructured Text to RxNorm via UMLS API

We used the U.S. National Library of Medicine's (NLM) Unified Medical Language System (UMLS) API, which takes string as input and outputs a matching code from one or more specified standardized health vocabularies. For this approach, the API was configured to output only RxNorm codes. The version of UMLS that we used was 2023AB.[2]

Approach 3: Generative Terminology Mapping - UMLS API Plus GPT-4

After a cursory review of the outputs from Approach 2, we found a high rate of incorrect mappings. In this approach, we used UMLS API as a "first pass" at mapping all of the terms, and then we prompted GPT-4 to act as an informaticist to review the map. As a non-medical and non-informatics specific model, GPT-4 has no special knowledge of RxNorm codes, so we asked it to reason only using the text names of the source term and the mapped RxNorm code. There was no augmentation, addition of Retrieval-Augmented Generation (RAG), or other techniques to inform GPT-4 about medications or RxNorm codes.

The prompt used is as follows, with a specific source and mapped term from UMLS:

```
As an expert in medication term mapping with RxNorm and experienced
in clinical informatics and curation, you have a task: You're given
a SOURCE_TERM from a medication text, which has been mapped to a
MAPPED_TERM in RxNorm. Your objective is to evaluate the accuracy of
this mapping. The primary criterion is that the ingredients of the
mapped term must match those in the source term. If the mapped term
introduces an ingredient not found in the source term, it's deemed
incorrect.
```

---

[2] Atropos Health has no affiliation or relationship with NLM or the UMLS and none of the organizations whose vocabulary sources are included in the UMLS has endorsed Atropos Health or any of its products.

©Copyright Atropos Health, 2023. All rights reserved.

Provide your evaluation in a JSON format with attributes: ingredient_verdict, confidence, and reasoning.

> ingredient_verdict: Can be ingredient_correct or ingredient_incorrect. ingredient_correct means that SOURCE_TERM and MAPPED_TERM details align perfectly. ingredient_incorrect means that there is a mismatch between SOURCE_TERM and MAPPED_TERM in terms of ingredients
>
> confidence: Reflects your confidence level in the verdict. Options are high, moderate, or low.
>
> reasoning: Concise explanation, no more than 80 words.

For the given terms:

SOURCE_TERM: IPRATROPIUM BROMIDE SOLUTION INHALATION

MAPPED_TERM: ipratropium bromide 0.2 MG/ML Inhalation Solution

Evaluate the mapping accordingly.

An example of a response (parsed out of JSON format) is:

> ingedient_verdict: ingredient_correct
>
> confidence: high
>
> reasoning: The MAPPED_TERM 'ipratropium bromide 0.2 MG/ML Inhalation Solution' maintains the same active ingredient 'Ipratropium Bromide' as in the SOURCE_TERM 'IPRATROPIUM BROMIDE SOLUTION'. Therefore, the mapping is correct.

©Copyright Atropos Health, 2023. All rights reserved.

Human Expert Review and Ground Truth

To grade the quality of the maps, we created different levels or dimensions of correctness that the human expert informaticist categorized for each of the source term to mapped term pairs. These were binary evaluations and were assessed separately for each map, so a given map would have values for all of these dimensions.

- A binary indication (“map_correct_flag”) or “All Aspects Correct” of whether the mapped RxNorm code semantically subsumes the drug concept represented by the source term (one concept “subsumes” another if either the two concepts are synonymous, or the former is more general but semantically contains the other. Thus, “sandwich” subsumes both “sandwich” and “ham sandwich”).
- A binary indication (“map_optimal_flag”) of whether the mapped RxNorm code was the semantically closest code available to the source term. A mapping could be correct without being optimal, if a more specific mapping could have been provided.
- A binary indication (“ingredient(s)_correct_flag”) of whether the mapped RxNorm code correctly reflects the ingredient(s) in the source term.[3] We used this to indicate “Ingredient Correctness.”

Based on the expert terminologist review, automated medication term-RxNorm pairs were classifiable into several categories, listed here with examples:

A. “Optimal”: the RxNorm code was the most specific available code that subsumed the source term.

B. “All Aspects Correct”: the RxNorm code subsumed the source term in all respects, including unit dose, etc. This does not imply optimality — there might be a more specific code available.

Source term: “PILOCARPINE NITRATE SOLUTION OPHTHALMIC”
Mapped term: RxNorm 103244, “pilocarpine nitrate”
Optimal map: RxNorm 373454, “pilocarpine ophthalmic solution”

---

[3] If either map referenced any ingredients not included in the other, this column was flagged as incorrect. If the mapped RxNorm code referenced an RxNorm ingredient that corresponded to an RxNorm precise ingredient in the input term (link), that was not considered an error; if the reverse occurred (the mapped RxNorm code referenced an RxNorm precise ingredient corresponding to an RxNorm ingredient in the input term), it was flagged as incorrect, since that constituted assuming information not implicit in the input term.

©Copyright Atropos Health, 2023. All rights reserved.

C. “Ingredient Correct”; the ingredients represented in the term matched the ingredients implicit in the RxNorm code, Some other detail of the RxNorm code may not reflect the contents of the source term, e.g. the code details about dosage form or strength that were not in the source term.

    Source term: TETRACYCLINE HCL OINTMENT OPHTHALMIC
    Mapped term: RxNorm 1164863, “tetracycline Drug Implant Product”
    Optimal map: RxNorm 2648277, “tetracycline hydrochloride ophthalmic ointment”

D. “Ingredient Incorrect”: the ingredients represented in the term did not match or omitted one or more of the ingredients implicit in the RxNorm code.

    Source term: DULAGLUTIDE MG/0.5 ML AUTO OR PEN INJECTOR SUBCUTANEOUS
    Mapped term: RxNorm 1996188, “Sublocade Injectable Product”
    Optimal map: RxNorm 1551293, “dulaglutide injectable product”

©Copyright Atropos Health, 2023. All rights reserved.

Results

Summary of Metrics for Different Approaches

Ingredient Correctness Mapping Outcomes

Approach	Ingredient Correct	Ingredient Incorrect	Number of Final Maps
Janus	832 (83.4%)	166 (16.6%)	1000
UMLS API Alone[4]	923 (92.5%)	75 (7.5%)	998
Generative Terminology Mapping	991 (99.2%)	7 (0.70%)	923

Other Mapping Outcomes

Approach	All Aspects Correct	Map Optimal	Number of Final Maps
Janus	633 (63.4%)	422 (42.3%)	1000
UMLS API Alone	441(44.2%)	323 (32.4%)	998
Generative Terminology Mapping	441 (47.8%)	323 (35.0%)	923

Of the mapping outcomes, the most crucial to get right is “Ingredient Correctness” for our use cases, which typically define cohort inclusion or exclusion criteria or determine outcomes. While there are certainly clinical uses for knowing the nuanced details of a drug exposure (form, dosing, strength, packaging), our platform emphasizes drug exposure - the level of specificity of the ingredient is sufficient in these cases.

The most serious and consequential type of error would be mapping a source term to an incorrect ingredient; for example, mapping a source term such as “ACET TAB 500 MG” to the mapped RxNorm code 1364430 for “apixaban”. This would have dramatic consequences in our platform, as

---

[4] UMLS automated mapping was performed for 998 terms due to a character encoding issue in 2 terms.

©Copyright Atropos Health, 2023. All rights reserved.

mentioned these drug ingredients remain a crucial part of cohort building. Our tolerance for such errors is therefore extremely low, with even a 5% error pushing the upper bounds of acceptable.

It's important to note that our approach – generative terminology mapping – could not improve optimality or all aspect correctness in any way other than to filter out the maps that were "Ingredient Incorrect" from the final setup. It could not add more maps that were correct in all aspects or change non-optimal maps into optimal maps or incorrect maps into correct maps. Its impact on the rate of "ingredient correctness," however, was significant, as outlined in this white paper.

©Copyright Atropos Health, 2023. All rights reserved.

Approach 1: *Janus*-Derived Maps

Janus *Correctness Metrics*

Metric	Count	Percent of Total (n=998)
All Aspects Correct	633	63.4%
Ingredients Correct	834	83.6%
Map Optimal	422	42.3%

Ultimately, we found that the mappings generated by the healthcare-specific NLP model from *Janus* were not sufficient for our use cases. Even looking only at Ingredient Correctness, we found that only 83.4% of the maps had correct ingredients, according to our definition. This was not primarily due to multiple-ingredient drugs, but rather to mappings that were simply incorrect, such as mapping terms such as “MULTIVITAMIN TABLET” to RxNorm 45045 - “avobenzone,” with relatively high confidence. Accepting these maps would have led to more than 173 million records with incorrect ingredients. For overall correctness and optimality, the performance was worse.

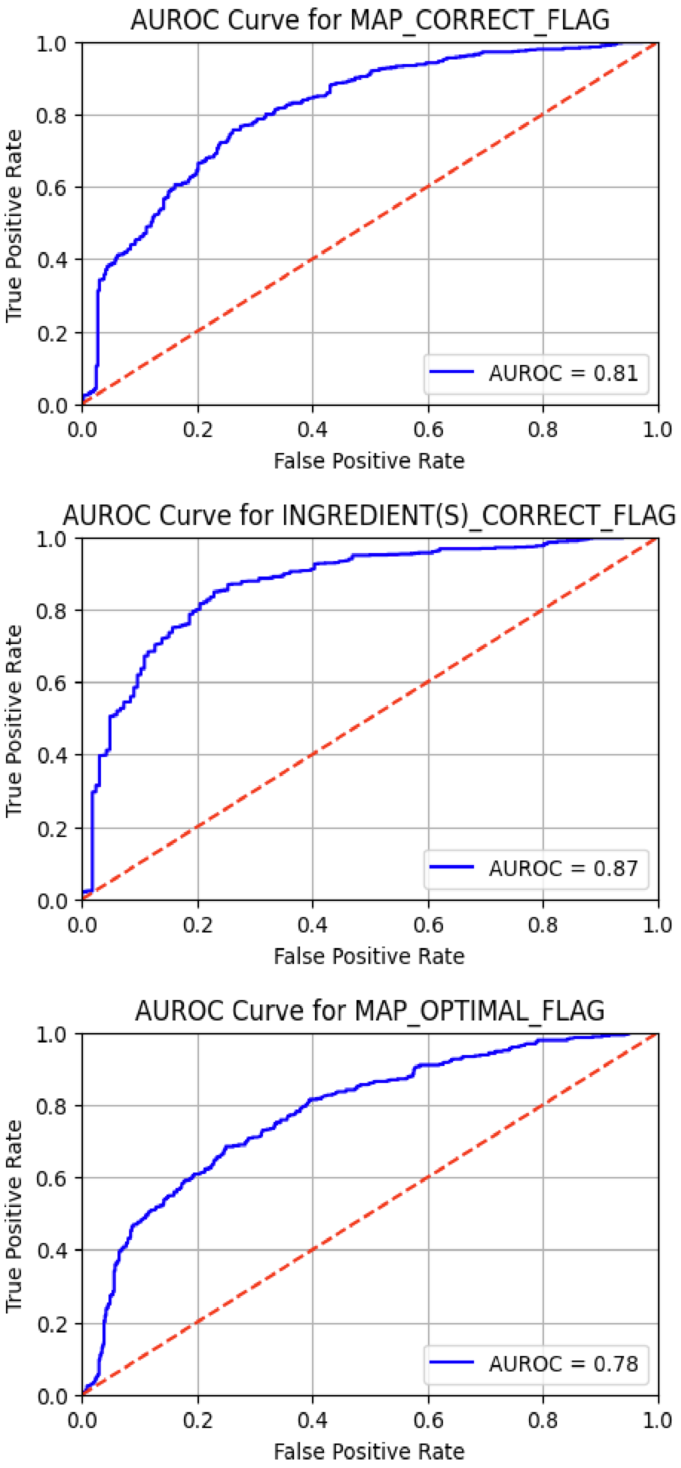
Since confidence probabilities were returned by the *Janus* model, we constructed Area Under the Receiver Operating Characteristic Curve (AUROC) measures in an attempt to determine a cutoff of confidence above which we could be assured of map quality.[5] We found that even though the AUROC overall metrics were fairly high, any confidence threshold that would have improved Ingredient Incorrectness to less than 10% (still a very high tolerance for highly incorrect data) would have eliminated up to 40% of all of the maps.This would leave ~38,000 source terms for our human expert to map - a substantial amount of human curation to deal with the volume of data remaining.

---

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8831439

©Copyright Atropos Health, 2023. All rights reserved.

Janus *AUROC Curves for Different Mapping Outcomes*



©Copyright Atropos Health, 2023. All rights reserved.

Approach 2: UMLS API Alone

UMLS API Alone Correctness Metrics

Metric	Count	Percent of Total (n=998)
All Aspects Correct	441	44.2%
Ingredients Correct	923	92.5%
Map Optimal	323	32.4%

Two of the top 1,000 terms were lost in the pipeline run due to an API error and were discarded. We compared the RxNorm codes mapped by the UMLS API to the gold standard established by our mapper.

In the UMLS run alone, the rate of Ingredient Correctness increased by 9% over the *Janus* NLP mapping, from 83% to 92%. Since the model does not output confidence, we are unable to generate AUROC for this output. However, this threshold of 92% would still allow a large number of rows with incorrect ingredients (69.4 million rows, or 5.3% of the total rows accounted for in the top 1,000 medication terms) into downstream data. Nevertheless, the reduction in ingredient error rate by 54.8% relative to using *Janus* is notable, particularly because the UMLS RxNorm tool is a free-to-use API available for use under generous Terms of Use for any entity with a free UMLS license. The UMLS API documentation can be found online.[6]

Although this finding was initially somewhat surprising, it made more sense considering that the NLP model was originally intended for use on medication references in free text notes, while this data task was operating on structured data from EHR systems. The UMLS API model was specifically designed for exactly this use case - inputting short medication name strings and resolving to RxNorm codes.

Despite the higher performance in terms of Ingredient Correctness, the NLP model was able to generate more maps that were optimal than the UMLS model (42.2% vs 32.5%). Furthermore, compared to the NLP model, many maps outputted by the UMLS API were to RxNorm codes that instantiate a strength value, which wasn't specified in the input terms (with the exception of "traditional" normal saline mapping to 0.9% sodium chloride solutions).

While many of these maps showed correct ingredients, the errors would be significant for any analyses involving medication dose, form, or other features. None of the approaches we tested could approximate the performance of an expert human in all of these dimensions. As a result, the

---

[6] https://documentation.uts.nlm.nih.gov/rest/home.html

©Copyright Atropos Health, 2023. All rights reserved.

specific method outlined here is only appropriate for use cases where ingredient correctness alone is sufficient.

Nevertheless, given that ingredient correctness is the primary goal, the UMLS API proved to be significantly more compelling to use as a tool given its ~50% lower rate of generating such errors and the fact that it remains free and open to use. While 92% may be sufficient for some use cases, this level of ingredient incorrectness would not be acceptable for our use cases that demand precision and correctness. For example, the UMLS API mapped the source term "MAGNESIUM HYDROXIDE INTERNAL POWDER ORAL" to the RxNorm 370822 - "aluminum hydroxide / magnesium hydroxide Oral Suspension." This represents an example of the RxNorm API "hallucinating" an additional ingredient not mentioned in the source.

©Copyright Atropos Health, 2023. All rights reserved.

Approach 3: Generative Terminology Mapping

Taking the 998 generated source-term maps above, we created a pipeline to submit these sequentially to OpenAI's GPT-4 through its Python API using an enterprise account, and using the prompt previously described. Each term took approximately 5 seconds to send, process, and retrieve from the UMLS API and 8 seconds from the GPT-4 API. In addition to this, we implemented several layers of retry, backoff, and caching logic to overcome intermittent interruptions, slowdowns, and timeouts as GPT-4 was under considerable load limitations. We used a temperature of 0 to provide deterministic results, following this workflow:

1. Distill unique source terms
2. Map using the UMLS API to generate a suggested source term to RxNorm map
3. Filter out the low-quality ingredient-incorrect maps using our GPT-4 layer
    a. In practice, this means removing any maps that GPT-4 indicated contained incorrect or missing ingredients
4. In practice, this would mean trusting the UMLS API's mapping unless our Generative AI review concluded that the ingredients did not match between source and mapped term.

Therefore, we can fully automate the pipeline and then remove any maps that GPT-4 flagged as being potentially incorrect in terms of their ingredients. These residual maps can be funneled into a human review process, or at least they will not enter the dataset as an RxNorm mis-map.

©Copyright Atropos Health, 2023. All rights reserved.

Generative Terminology Mapping Results: 90% Reduction in Errors, 91% Coverage, 98% Reduced Cost

Generative Terminology Mapping (UMLS API Plus GPT-4) Correctness Metrics

Metric	Count	Percent of Total (n=923)[7]
All Aspects Correct	441	47.8%
Ingredients Correct	916	99.24%
Map Optimal	323	35.0%

Generative Terminology Mapping vs. Ground Truth Results and Confusion Matrix

Ground Truth	Generative AI Verdict	Count	Percent of Total (n=998)	Error (Positive = Ingredient Correct)
Ingredient Correct	Ingredient Incorrect	7	0.7%	False Negative
Ingredient Correct	Ingredient Correct	916	91.78%	True Positive
Ingredient Incorrect	Ingredient Correct	7	0.7%	False Positive
Ingredient Incorrect	Ingredient Incorrect	68	6.81%	True Negative

When we completed this process, there were only 7 maps out of 998 total that GPT-4 agreed was correct in terms of its ingredients between the source term and the UMLS API-derived mapped RxNorm code that our human reviewer disagreed with. In other words, if we used the combined UMLS and GPT-4 approach, we would achieve an overall map rate of 91.78% of the input maps, and, of these, our human reviewer would disagree with the ingredients in only 7 (.7% of all maps, .76% of completed maps). 7.5% (75) of the maps would be removed by the GPT-4 ingredient check.

Of these maps removed as being ingredient incorrect, only 7 would be maps where our human expert concluded the map was ingredient correct while the Generative AI review disagreed. Adding GPT-4 as a second reviewer to the UMLS mapped data reduced the error rate of ingredient correctness as measured against our human expert's ground truth by ~90% (7.5% without GPT-4 to .7% with our generative terminology mapping approach).

---

[7] n=923 after 75 maps flagged by GPT-4 as being incorrect at the ingredient level were removed

©Copyright Atropos Health, 2023. All rights reserved.

Overall, GPT-4 and our human expert agreed in 984 cases, and disagreed in 14 (998 cases were passed through both models). This rate of agreement– 98.6% –was extraordinarily high and surprised us. Again, zero fine-tuning was performed and the model used (GPT-4) had absolutely no healthcare or informatics domain-specific focus. These numbers result in a Cohen's κ (kappa) coefficient of interrater reliability of 0.899 (95% CI: [0.842, 0.947]), which can be characterized as to be right on the margin between "strong" and "almost perfect" agreement even in a very demanding healthcare research context.[8] Future work could be established that leveraged multiple expert raters to establish a baseline expectation for human vs. human inter-rater reliability would be for this use case. Our experience is that, even with a detailed curation editorial policy and extremely knowledgeable expert mappers, differences can occur between human experts when performing this kind of mapping.

Data Engineering Challenges

The principle engineering challenge for this approach is that, as of October 2023, neither GPT-4 nor the UMLS API offered a bulk API interface that a flat file could be submitted to, and an operation simultaneously called on each row of a spreadsheet with thousands of entries. Instead, we had to engineer a framework to take each source-term, map it through the UMLS API, avoid hitting rate limits and handling network errors, and then submitting these to GPT-4 through its Python API one source-mapped term pair at a time. Doing this in a non-fragile way that could handle delays and failures proved to be challenging, but possible. The UMLS API processed each map in approximately 5 seconds on average. GPT-4 took on average 8 seconds to process and return each of our "chats", for a total time of about 2.5 hours runtime for the ChatGPT portion for 1,000 terms.

Scaling our pipeline for 1,000 terms to ~95,000 using this framework will be challenging, but it should be a quite surmountable problem, particularly in light of the cost of manual mapping.

---

[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/

©Copyright Atropos Health, 2023. All rights reserved.

Estimated Cost and Comparisons

We estimate that the marginal cost per map generated from this approach will be approximately 2.5 cents (~15 seconds of compute per map, plus roughly $.02 for GPT-4, so ~2.5 cents per map overall as a high-end estimate).[9] This estimate suggests we could complete an automatic mapping of the 95,000 terms for approximately $2,500. In contrast, the human approach would cost $300 per 150 terms per hour, or $2 per term, so roughly 80x as much per term. Runtime will be more of a concern for the original run of 95,000 terms, since this would suggest a continuous runtime of roughly two weeks if not parallelized.

We estimate the total non-personnel cost of running this experiment on 1,000 samples from the UMLS API plus GPT-4 approach was under $100, with most of it coming from AWS compute and Databricks costs on a i3.xlarge Amazon AWS EC2 ($0.312 per hour for EC2 and $1 per hour for Databricks) instance that ran for about 50 hours total across all project phases. The cost of using 1,000 calls to the GPT-4 API was around $25.

Between the quarterly refresh of *Eos* upon which this experiment was conducted and the next refresh, *Eos* changed some of its internal mappings on the source data to consolidate terms. This improvement actually led to a reduction in coverage of the 1,000 previously most common medication source terms from 50% of the unmapped records to less than 10% in the new data. This indicates another major pitfall of manual mapping – changes in the source data can quickly cause your map rate to collapse, resulting in a near-total loss of the time and money invested previously. If we had invested nearly $200,000 in a one-time medication mapping project based on *Eos*, much of its utility would have been lost in less than 3 months due to this change. This challenge is addressed by using a much less resource-intensive and automatic approach that can be added to or re-run as needed, such as the generative terminology mapping approach outlined in this report.

---

[9] As of Nov 14, 2023, GPT-4 charged $.03 per 1,000 tokens of input and $.06 per 1,000 tokens of output. One of our maps constitutes ~250 tokens of input and ~100 tokens of output. Therefore the per-map cost should be roughly ((.03/1000)*250) + ((.06/1000)*100) = .0135 or 1.35 cents. Calculating the compute plus GPT-4: (1.312/60/60)*15 +.02 (rounded up to 2 cents from 1.35 to account for possible increases in charges or unknown additional costs to be very conservative) = ~$.025 or 2.5 cents per map. This represents a high end estimate of the per-map cost.

©Copyright Atropos Health, 2023. All rights reserved.

Conclusion

Using the generative terminology mapping approach outlined here will enable Atropos Health to correctly map 91% of our *Eos* medication data with 99.2% accuracy for the approved maps.[10] Approximately 7% of the maps generated would be sent back for human review, and only 0.7% would be sent back unnecessarily. This means that, instead of taking more than 600 hours and over $190,000 to manually map 100% of these terms, we believe that we will be able to map the vast majority of *Eos* medication data in a totally automated pipeline that can be re-run with every data refresh.

We estimate that a run of our automated pipeline would cost ~98% less per map performed vs a 100% human-driven process. This approach being scaled to the full 95,000 terms would leave a residue of several thousand maps that our generative AI reviewer flagged as being potentially incorrect that a human expert would need to review. However, the Cohen's κ measure of interrater reliability between GPT-4 and our human informatics expert was .899, even though GPT-4 had no access to our detailed curation/mapping policy beyond the instruction: 'The primary criterion is that the ingredients of the mapped term must match those in the source term. If the mapped term introduces an ingredient not found in the source term, it's deemed incorrect.'

Generative Terminology Mapping Produces Research-Grade RxNorm-Mapped Medication Data at Scale

The UMLS Plus GPT-4 mappings generated in our experiment were of such high quality that even the 0.7% flagged by our expert reviewer as "incorrect" are near-misses, or a consequence of a very detailed curation policy that we did not provide to GPT-4. We assess that even a reduction of performance down from 99.2% correctness, or .8% error (95% CI - 0.28% to 1.44% error) would be acceptable. Unlike errors produced by the *Janus* model–which were often extremely puzzling and nonsensical– the 'errors' from our generative terminology mapping approach tend to be near-misses in addition to only comprising 0.7% of the maps. Many mappings of messy, aggregated EHR data are performed by non-experts that introduce errors of their own, and we believe strongly that this approach will lead to a much better than average mapping accuracy rate.

When examining the cases where our human expert and the GPT-4 model disagreed, it's possible to find cases that a fairly plausible inference was made by GPT-4, but the very specific curation policy set by our human informatics expert led him to choose differently. Speaking broadly, there

---

[10] While it is possible that the top 1,000 source terms by incidence are not representative of the entirety of the "long tail," we believe that this is not generally the case, as the performance does not seem to be highly correlated with whether the terms are high incidence or low incidence within the top 1,000.

©Copyright Atropos Health, 2023. All rights reserved.

were no cases where the model tried to defend a completely incorrect approach. While its decision differed from our informatics expert in 1.4% of cases (evenly split with positive and negative), it generally provided a plausible and clinically appropriate argument.

Finally, the fact that this approach can generate 1) maps at an incredible volume 2) maps with high accuracy in terms of ingredient correctness and 3) provide a “reasoning” for GPT-4’s evaluation of each source term and mapped term pair represents a step forward for programmatic mapping and curation that was previously only available to those utilizing very specialized vendor services for terminology mapping. The “reasoning” provided by the model helps to bolster trust in the data, and provides lineage information about precisely why maps were accepted, increasing the alignment of this data to emerging standards of Real World Data / Real World Evidence that emphasize data quality and data lineage. Lineage information for mapping structured data to known ontologies is a particularly emerging area, and we believe this is an exciting step forward.

An example of reasoning that demonstrates that, even without fine-tuning, advanced generative AI models appear to be able to generate plausible and clinically correct reasoning about drug ingredients is the following example from our evaluation the top 1,000 maps:

Source Term	**Mapped UMLS RxNorm Code**	**Mapped UMLS RxNorm Name**	**GPT-4’s Reasoning for why this map is ‘ingredient_correct’**
ALENDRONATE SODIUM TABLET ORAL	904419	alendronic acid 10 MG Oral Tablet	Alendronate sodium and alendronic acid are the same drug, they both refer to bisphosphonate class used for osteoporosis

Our human expert judged this particular mapped term to be ‘ingredient incorrect,’ because, according to our very stringent editorial policy, aledronate sodium and alendronic acid should not be considered to be the same ingredient. Nevertheless, GPT-4’s reasoning includes the correct fact that both of these are biphosphate drugs used for osteoporosis. The difference here is very subtle, and a slightly less restrictive editorial policy might have judged GPT-4 to be correct.

In at least one case, GPT-4 also caught an error that the human expert made in marking a map as “ingredient correct.” To err is human. An error rate of 1/1000 is commendable for any human-centric process of applying mental energy. This indicates that, even for human experts, a second check from a generative AI agent may be useful to hone in on the .01% of times that even a human expert may make an error.

In the Appendix, you can see that GPT-4 does sometimes produce clinical reasoning errors, which is not surprising given that this model is not healthcare-specific and was not fine tuned or otherwise trained at all for this type of task. More research and/or review could be done to

©Copyright Atropos Health, 2023. All rights reserved.

establish the rate of clinical reasoning errors or 'hallucinations' in the data, but since GPT-4 agreed with our expert reviewer nearly 100% of the time, it seems clear that such 'hallucination' or errors did not have a large effect on the outcome.

The question of whether the "reasoning" provided here represents true reasoning or not seems largely philosophical – even if the model is simply parroting "reasoning," it seems to be able to reproduce the same conclusions drawn by a human expert the vast majority of the time and the reasoning appears to generally relate to the model's conclusions.

Caveats and Future Directions

This evaluation was based on medication terms from one source, generated from multiple data fields in a very particular way, as described above. Caution should be exercised in generalizing the findings to other types of medication terms, particularly those that differ in their internal lexical consistency and/or lexical similarity to RxNorm terms.

The next step of our evaluation would be to take a totally random sample of maps and repeat this validation to test our assumption and hypothesis that the performance of the mapping on the top 1,000 terms is comparable to the performance of the mapping on all 95,000 terms. Additionally, our calculation of agreement would benefit from multiple human experts' establishing multiple interrater reliability measures to compare human vs. human and human vs. model's level of agreement.

Although we believe that this approach allows us to produce maps that will be fit for specific purposes at scale due to their low rate of errors in terms of ingredient correctness, it's important to note that this generative terminology approach cannot match the rate of "optimal" outputs by a human expert. This approach still leads to maps which 'hallucinate' or impute aspects of the medication such as strength, and sometimes form when this information was not indicated in the source term.

Interestingly, this 'hallucination' is entirely due to the UMLS API's RxNorm mapping model, and not the generative AI model applied afterwards. Future applications of the generative terminology approach may actually allow us to catch and remedy these "hallucinations" or imputations. Therefore, we can identify how generative AI models can actually act as a factor to remove hallucinations from other modeling approaches, and provide transparency by generating reasoning that appears to relate to the conclusions drawn by the model.

©Copyright Atropos Health, 2023. All rights reserved.

Appendix

Terms That Both Human Expert and GPT-4 Indicated Ingredient Incorrect (Excerpts)

Medication Source Term	UMLS RxNorm Name	UMLS RxNorm Code	GPT-4 Reasoning
LIDOCAINE HCL/DEXTROSE 5 %/PF IV SOLUTIONS	lidocaine hydrochloride 5 MG/ML Injectable Solution	1010900	The source term includes 'dextrose' as an ingredient which is missing in the mapped term. The mapped term only includes 'lidocaine hydrochloride', hence the ingredient mismatch.
MAG HYDROX/AL HYDROX/SIMETH ORAL SUSPENSION	magaldrate / simethicone Oral Suspension [Ri-Mag Plus]	1086969	The MAPPED_TERM introduces 'magaldrate', which is not found in the SOURCE_TERM. The SOURCE_TERM includes 'AL HYDROX' (aluminum hydroxide), which is missing in the MAPPED_TERM.
ACETAMINOPHEN DROPS	Viva-Drops Lubricating Eye Drops	1090069	The source term indicates 'ACETAMINOPHEN' as the ingredient, while the mapped term suggests 'Viva-Drops Lubricating Eye Drops', which introduces a completely different ingredient.
GUAIFENESIN SUSTAINED RELEASE TABLET ORAL	sulfamethazine Extended Release Oral Tablet [Sustain]	1101615	The source term ingredient is Guaifenesin while the mapped term ingredient is Sulfamethazine. These are different ingredients.
ACETAMINOPHEN SUSTAINED RELEASE TABLET ORAL	sulfamethazine Extended Release Oral Tablet [Sustain]	1101615	The mapped term includes 'sulfamethazine' which is not included in the source term 'ACETAMINOPHEN'. Therefore, ingredients do not align.
OMEPRAZOLE MAGNESIUM ORAL	magnesium hydroxide / omeprazole / sodium bicarbonate Oral Product	1157482	The MAPPED_TERM introduces 'sodium bicarbonate' and 'magnesium hydroxide' which are not found in the SOURCE_TERM 'Omeprazole Magnesium Oral'.
IRON ORAL	iron carbonyl Oral Product	1159146	The source term only includes iron, whereas the mapped term introduces the ingredient 'carbonyl'. Therefore, the ingredients do not align perfectly.
FLUTICASONE/UMECLIDIN/VILANTER ORAL DISKUS	fluticasone Inhalant	1165655	The MAPPED_TERM only includes fluticasone, while the SOURCE_TERM also

©Copyright Atropos Health, 2023. All rights reserved.

INHALER INHALATION	Product		includes Umeclidin and Vilanter.
VANCOMYCIN HCL IN 5 % DEXTROSE SOLUTION INTRAVENOUS	Isolyte S in 5 % Dextrose	1189612	The mapped term introduces a new ingredient, Isolyte S, which is not found in the source term. The source term's main ingredient is Vancomycin HCL, not present in the mapped term.
FENTANYL CITRATE MCG SOLUTION	fentanyl 0.004 MG/ML / ropivacaine hydrochloride 2 MG/ML Injectable Solution	1233687	The MAPPED_TERM introduces ropivacaine hydrochloride, an ingredient not found in the SOURCE_TERM, making the mapping incorrect.
FENTANYL CITRATE INJECTABLE	fentanyl 0.004 MG/ML / ropivacaine hydrochloride 2 MG/ML Injectable Solution	1233687	The MAPPED_TERM introduces ropivacaine hydrochloride, an ingredient not found in the SOURCE_TERM, leading to an incorrect mapping.
FENTANYL CITRATE SOLUTION	fentanyl 0.004 MG/ML / ropivacaine hydrochloride 2 MG/ML Injectable Solution	1233687	The source term only mentions fentanyl citrate while the mapped term introduced another ingredient, ropivacaine hydrochloride.

©Copyright Atropos Health, 2023. All rights reserved.

Terms That Both Human Expert and GPT-4 Indicated Ingredient Correct (Excerpts)

Medication Source Term	UMLS RxNorm Name	UMLS RxNorm Code	GPT-4 Reasoning
BUPROPION HCL TABLET ORAL	bupropion hydrochloride 100 MG Oral Tablet	993687	Both the source term and mapped term refer to the same active ingredient, bupropion hydrochloride (HCL is a common abbreviation for hydrochloride). In addition, both are orally consumed tablets.
CHLORHEXIDINE GLUCONATE SOLUTION	chlorhexidine gluconate 5 MG/ML Topical Solution	834176	Both SOURCE_TERM and MAPPED_TERM share the same active ingredient, which is chlorhexidine gluconate. No additional ingredients are introduced in the MAPPED_TERM.
HYDROXYZINE HCL TABLET ORAL	hydroxyzine hydrochloride 25 MG Oral Tablet	995258	Both SOURCE_TERM and MAPPED_TERM contain the same active ingredient which is hydroxyzine hydrochloride.
BECLOMETHASONE DIPROPIONATE INHALANT INHALATION	120 ACTUAT beclomethasone dipropionate 0.084 MG/ACTUAT Metered Dose Inhaler	966675	Both the source term and mapped term contain the same active ingredient beclomethasone dipropionate. The mode of application for both is inhalation.
PROCHLORPERAZINE EDISYLATE SOLUTION INJECTION	prochlorperazine 5 MG/ML Injectable Solution	2102949	The active ingredient in both SOURCE_TERM and MAPPED_TERM is prochlorperazine. The form is also consistent (injectable solution), indicating an accurate mapping.
PHENYTOIN SODIUM EXTENDED CAPSULE ORAL	phenytoin sodium 100 MG Extended Release Oral Capsule	855671	Both SOURCE_TERM and MAPPED_TERM contain the same active ingredient 'phenytoin sodium'. The dosage form mentioned, 'oral capsule', is also consistent.

©Copyright Atropos Health, 2023. All rights reserved.

HALOPERIDOL LACTATE SOLUTION INJECTION	haloperidol 5 MG/ML Injectable Solution	204416	The mapped term, 'haloperidol 5 MG/ML Injectable Solution', has the same active ingredient with the source term, 'HALOPERIDOL LACTATE SOLUTION INJECTION'. No extra ingredient is introduced.
PROMETHAZINE HCL SOLUTION	promethazine hydrochloride 1.25 MG/ML Oral Solution	992432	Both the source term and mapped term refer to Promethazine hydrochloride, which indicates accurate mapping.
TRAZODONE HCL TABLET	trazodone hydrochloride 150 MG Oral Tablet	856364	Both SOURCE_TERM and MAPPED_TERM refer to 'trazodone hydrochloride', with 'HCL' being short for hydrochloride. The ingredients align properly.

©Copyright Atropos Health, 2023. All rights reserved.

Terms Human Expert Indicated Ingredient Incorrect, GPT-4 Indicated Ingredient Correct

These maps would be propagated into the data and so represent the most important type of error, if truly incorrect. Other clinical reviewers that we engaged had mixed opinions on whether all of these should be considered true errors from a clinical point of view. There are additional errors in GPT-4's clinical reasoning that are interesting, but fall outside of the scope of this white paper.

Medication Source Term	UMLS RxNorm Name	UMLS RxNorm Code	GPT-4 Reasoning
ALENDRONATE SODIUM ORAL	alendronic acid 10 MG Oral Tablet	904419	Although the SOURCE_TERM and MAPPED_TERM use different names (Alendronate Sodium vs Alendronic Acid), they both refer to the same medication. Alendronic acid is another name for Alendronate Sodium.
ALENDRONATE SODIUM TABLET ORAL	alendronic acid 10 MG Oral Tablet	904419	Alendronate sodium and alendronic acid are the same drug, they both refer to bisphosphonate class used for osteoporosis.
PROMETHAZINE HCL/CODEINE SYRUP	codeine phosphate 2 MG/ML / promethazine hydrochloride 1.25 MG/ML Oral Solution	991486	Both the SOURCE_TERM and MAPPED_TERM involve the same ingredients: Promethazine and Codeine, despite variations in the pharmaceutical names and concentrations.
INSULIN,REGULAR INJECTABLE INJECTION	insulin, regular, human Injection	2179743	Both source and mapped terms contain the same active ingredient, Insulin. Although the mapped term specifies it as human Insulin, it aligns with the source term.
INSULIN ASPART PROT/INSULN ASP SOLUTION	insulin aspart, human Injectable Solution	378914	Both terms contain 'insulin aspart', meaning they represent the same active ingredient. 'Prot/Insuln Asp' in the source term indicates the protraction and aspartic acid modification, which aligns with 'human' in the mapped term.
HEPARIN SODIUM SOLUTION	heparin sodium, porcine 1000 UNT/ML Injectable Solution	1361226	Both SOURCE_TERM and MAPPED_TERM have the same main ingredient, which is 'Heparin Sodium'. The mapped term does specify 'porcine', indicating the origin, but this doesn't introduce a new medicinal ingredient.
ZOSTER VACCINE LIVE/PF VIALS INJECTABLE SUBCUTANEOUS	varicella-zoster virus vaccine live (Oka-Merck) strain Injection	1658006	The active ingredient in the source term 'Zoster Vaccine Live' and the mapped term 'varicella-zoster virus vaccine live' are the same, representing live versions of the varicella-zoster virus.

©Copyright Atropos Health, 2023. All rights reserved.

Terms Human Expert Indicated Ingredient Correct, GPT-4 Indicated Ingredient Incorrect

These maps would be returned to a human mapper for review and wouldn't be propagated into the data. Note that in some cases, GPT-4 is making errors in its reasoning and misstatements of fact or interpretation. In one case, there is an outright factual error, where GPT-4 asserts that "ophthalmic solution" is a drug ingredient. In many of these cases, the missing ingredient that GPT-4 has identified is more of a "carrier" or inactive ingredient in a mixture, which is likely why our human expert did not flag these as incorrect. There are additional errors in GPT-4's clinical reasoning that are interesting, but fall outside of the scope of this white paper.

Medication Source Term	UMLS RxNorm Name	UMLS RxNorm Code	GPT-4 Reasoning
CEFAZOLIN SODIUM/DEXTROSE,ISO SOLUTION INTRAVENOUS	cefazolin sodium	203171	SOURCE_TERM includes both 'cefezolin sodium' and 'dextrose', but MAPPED_TERM only accounts for 'cefezolin sodium'.
MORPHINE SULFATE/NORMAL SALINE PIGGYBACKS NON-FROZEN INTRAVENOUS	morphine sulfate	30236	The MAPPED_TERM 'morphine sulfate' is part of the SOURCE_TERM but it excludes 'normal saline' which is also an ingredient in the SOURCE_TERM.
CEFTRIAXONE SODIUM/LIDOCAINE KITS INJECTABLE INTRAMUSCULAR[11]	ceftriaxone sodium	203172	The mapped term only includes 'ceftriaxone sodium', missing the second ingredient 'lidocaine' found in the source term.
0.9 % SODIUM CHLORIDE (PHENOL) VIALS INJECTABLE	sodium chloride 9 MG/ML Injectable Solution	313002	The mapped term 'sodium chloride 9 MG/ML Injectable Solution' does not contain the 'phenol' ingredient present in the source term '0.9% SODIUM CHLORIDE (PHENOL) VIALS INJECTABLE'.
LIDOCAINE HCL/PALM OIL NON-AEROSOL SPRAY TOPICAL APPLICATION	lidocaine hydrochloride 20 MG/ML Topical Spray	1010769	The source term contains the ingredient 'Palm Oil' which is not found in the mapped term 'Lidocaine Hydrochloride 20 MG/ML Topical Spray'
PHENYLEPHRINE HCL IN 0.9% NACL SOLUTION INTRAVENOUS	phenylephrine hydrochloride 100 MG/ML Ophthalmic Solution	1234571	While both terms contain phenylephrine hydrochloride, the mapped term introduces an extra ingredient, an ophthalmic solution, not present in the source term.
VANCOMYCIN HCL IN 5 % DEXTROSE INJECTABLE/INTRAVENOUS FROZEN INTRAVENOUS	vancomycin 500 MG Injection	1807516	The mapped term lacks 'Dextrose', an ingredient mentioned in the source term. Hence, the ingredient mapping is considered incorrect.

---

[11] This appears to be the case where our human expert made a mis-map, either through an actual error or possibly a mistake by marking the map as correct when it omits an ingredient – lidocaine.

©Copyright Atropos Health, 2023. All rights reserved.