Exploring Novel Drug Research Area using Large Language Models Based on Research Trends in Biomedical Literature

Afnan[1], Michael Van Supranes[1,2], Tomohiro Nishiyama[1], Shoko Wakamiya[1], Eiji Aramaki[1]

[1]Nara Institute of Science and Technology, Japan, [2]University of the Philippines, Philippines

Abstract

The rapid expansion of biomedical literature makes manual identification of novel drug-disease relationships increasingly difficult. Existing approaches have leveraged LLMs to mine abstracts or construct knowledge graphs for drug repurposing. There are two key limitations: finite context windows for capturing macro-level research trends, and single-pass black-box pipelines make it difficult to verify outputs. This paper proposes a pipeline for discovering new drug targets by combining disease and drug research trends using Large Language Models (LLMs). Our method extracts PICO components from PubMed abstracts, normalizing the Population and Intervention Component to ICD and ATC codes, respectively. A *temporal frequency delta* matrix is constructed to capture publication count shifts across 2013 to 2022, then used to discover novel drug areas. Compared with the abstract-based baseline, our approach showed qualitative signs of generating combinations that were more closely aligned with observed research trends and, in some cases, more clinically plausible. These findings suggest the potential usefulness of structured trend information for LLM-based exploration, although the differences between the two methods were limited and the results remain preliminary. Future work will focus on validating the consistency and reliability of these candidates.

Keywords: LLMs, drug re-purposing, biomedical NLP, research trend, PubMed

1 Introduction

The biomedical literature has grown exponentially with time, creating a rich and complex body of knowledge for scientific discovery. Although experienced researchers can generate novel ideas from prior work, there are inherent limits to the amount of information that can be effectively processed and compared within a given time. As a result, identifying relevant studies and distilling key insights that can lead to breakthroughs has become increasingly challenging amid the vast and rapidly expanding volume of publications.

This challenge is pronounced in domains such as oncology, where thousands of abstracts are published each year, advancing research on drug-disease interactions, clinical trial outcomes, and molecular pathways (Wang et al., 2025; Yan et al., 2024). In this context, large language models (LLMs) have emerged as powerful tools for assisting researchers in navigating large-scale literature. LLMs can support automatic synthesis of clinical evidence and uncover hidden knowledge connections that would otherwise take a long time and require extensive manual review (Ghafarollahi and Buehler, 2025). In addition to literature review, recent innovations have demonstrated that LLMs can also assist in generating novel, plausible clinical hypotheses, such as identifying drug re-purposing candidates for Alzheimer's disease and Acute Myeloid Leukemia (Yan et al., 2024; Gottweis et al., 2025).

Despite the potential of LLM-based scientific discovery, direct application of LLMs to large-scale literature analysis has the following limitations:

Limited ability to capture overall research trends: LLM-based methods are constrained by a finite context window, limiting the number of abstracts or related work that can be included in a single prompt. LLMs may also fail to fully integrate all important findings when the input is too long. This phenomenon is called *lost in the middle* or retrieval failure during long-context processing (Du et al., 2025; Li et al., 2025). Thus, using LLMs requires selective inclusion of related work, which may hinder the model's ability to capture broader research trends.

Limited transparency in LLM-based pipelines: LLMs are black boxes and often deployed in single-pass pipelines for complex tasks. As a result, it is

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 997–1013

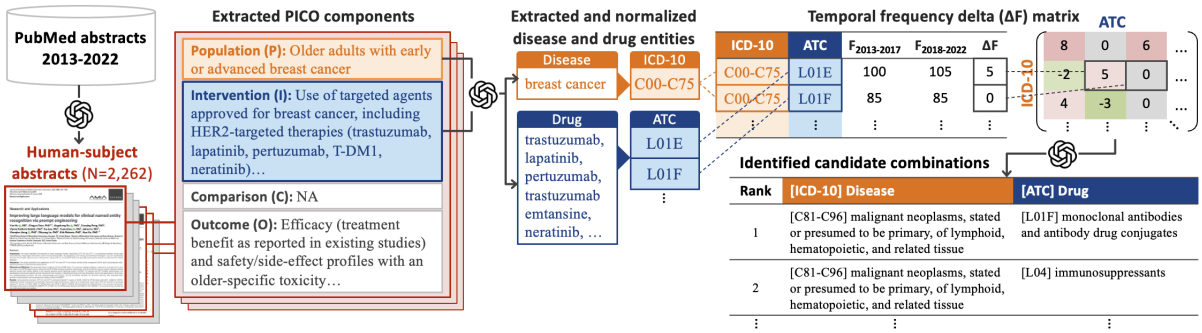


Figure 1: Analysis pipeline of the study. Abstracts were first filtered to include only human populations, then PICO components (Population, Intervention, Comparison, and Outcome) were extracted. The Population (P) and Intervention (I) components were then used to construct a matrix of publications for each P–I combination across two periods: 2013–2017 and 2018–2022. This matrix served as input for an LLM to generate novel potential P–I combinations. Multiple reasoning and an impact score in each step of generating are used to mitigate hallucination.

difficult to tell whether the model has correctly identified and incorporated relevant information or has generated hallucinated content (Wang et al., 2025).

Addressing these challenges remains an open problem. In this paper, we present a novel approach to support drug discovery by augmenting LLMs with temporally grounded bibliometric signals (Figure 1). Our contributions are two-fold. First, to address the difficulty LLMs face in detecting macro-level research trends across fields and time periods, we introduce a frequency delta matrix to represent research trends in a compact, context-efficient manner, enabling the incorporation of research momentum signals into prompts without requiring extensive collections of raw abstracts. Second, we develop an end-to-end pipeline that extracts structured information from abstracts based on the PICO framework (Population (**P**), Intervention (**I**), Comparison or Control (**C**), and Outcome (**O**)), constructs co-occurrence matrices, and leverages them to guide LLMs in finding candidate disease–drug research combinations. We compare this structured, trend-explicit formulation with a baseline that uses raw abstracts and publication years as unstructured, trend-implicit input, and examine whether LLMs can improve in generating more clinically plausible disease–drug research combinations.

2 Related Work

2.1 LLMs in Biomedical Discovery

The integration of LLMs into the medical field has shown immense promise, particularly in automating knowledge synthesis. Early benchmarks, such as BioRED (Du et al., 2025), established a gold standard for extracting explicit semantic triples (e.g., Drug A treats Disease B) from unstructured PubMed abstracts. Beyond simple extraction, LLMs are now being utilized as autonomous agents capable of generating novel and scientifically valid research ideas (Baek et al., 2025; Gottweis et al., 2025). However, a persistent challenge remains in ensuring the interpretability and reliability of these discoveries, especially when models move from sentence-level connections to complex, multi-faceted scientific problem-solving (Li et al., 2025; Williams, 2026).

2.2 Automated Hypothesis Generation in Drug Discovery

The primary objective of AI-driven drug discovery is to identify novel combinations of diseases and pharmacological agents that have not yet been clinically explored. Traditional methods relied on building discrete concept corpora and identifying relationships via word vector similarities (Islamaj et al., 2024) or link prediction within knowledge graphs (Yang et al., 2023). While these methods are effective for validating known biochemical pathways, they often fail to capture emerging research frontiers or trending therapeutic dyads that exhibit high growth momentum. Recent advances use the scientific knowledge in LLMs to move beyond predicting localized connections and toward identifying "hot topics" in science, a task traditionally handled by bibliometric analysis (Qi et al., 2023; Tshitoyan et al., 2019). Our work builds on this by investigating whether LLMs can automate the interpretive step of translating statistical clusters into actionable research hypotheses.

2.3 Topic Trends as Scientific Signals

The use of publication frequency and temporal shifts to identify emerging research directions is a well-established principle in scientometrics. Traditional bibliometric methods, such as co-word analysis and citation burst detection, have long been used to map the "intellectual turning points" of a scientific field (Qi et al., 2024). For instance, a sudden spike in publications on a specific drug-disease dyad often signals a transition from basic research to clinical interest. Our work builds on the premise that scientific discovery is a temporal process, in which "information remnants" in literature, captured via frequency shifts, can predict future high-impact research areas (Tshitoyan et al., 2019). By transforming bibliometric data into a *frequency delta matrix* (Δ), we construct a compact representation of research trends that can be efficiently incorporated into LLM prompts, providing signals of research momentum that are difficult to capture when using raw abstracts alone.

2.4 Reasoning over Tabular Data

While LLMs are natively trained on unstructured natural language, their ability to interpret structured formats like CSV, JSON, and SQL has become a focal point of research. Recent studies indicate that while LLMs can perform reasoning over tabular data, their performance is highly sensitive to the data's serialization format (Sybrandt et al., 2017). In the context of drug discovery, the ability of a model to ingest structured co-occurrence matrices is essential for identifying statistical trends. Researchers have found that providing LLMs with structured summaries of biomedical metadata enables more precise "*chain-of-thought*" reasoning than processing raw, disconnected abstracts (Kim et al., 2022). This study extends this line of inquiry by comparing structured tabular inputs.

3 Dataset Construction

The primary dataset for this study comprises scholarly abstracts retrieved from PubMed, spanning a 10-year period from 2013 to 2022. This period was selected to enable a balanced comparison between two consecutive five-year intervals (2013–2017 and 2018–2022). Using equal time intervals helps ensure that observed differences reflect meaningful shifts in research activity. In addition, this period captures the accelerated growth in oncology research.

Year	# of filtered articles	Avg. word count
2013	128	251
2014	169	242
2015	152	245
2016	176	245
2017	171	251
2018	168	249
2019	172	252
2020	293	250
2021	433	251
2022	400	177
Total	**2262**	–
Min	**128**	–
Max	**400**	–

Table 1: Number of abstracts after filtering (not animal or biomedical abstracts, only clinical abstracts), along with average word counts per year.

To construct a high-quality dataset for analysis, it is essential to accurately identify clinically relevant studies. However, traditional keyword-based Boolean searches often suffer from low precision due to polysemy and the inclusion of non-human studies.

The dataset was constructed in a two-stage process. First, an initial pool of 5,000 abstracts was retrieved from PubMed[1] via the API[2], with 500 abstracts randomly sampled per year to balance analytical coverage and computational feasibility. Second, each abstract was screened using an LLM (GPT-5, OpenAI[3]) to retain only those describing human studies (see Section 4.1 for details). This process yielded 2,262 abstracts, which constitute the final dataset used in this study.

Further details on the dataset distribution are provided in Table 1, which shows that annual publications remained below 180 between 2013 and 2018, exceeded 200 by 2020, and nearly doubled by 2021–2022. These trends have been attributed to advances in genomics and immunotherapy, increased clinical trial activity, and the growing use of artificial intelligence in data analysis (Yao et al., 2023; Wang et al., 2024; Azmi and Howe Chan, 2025).

4 Method

The overall methodology of this study consists of four main stages: LLM-based abstract screening and PICO extraction (Section 4.1), entity extraction

[1] https://pubmed.ncbi.nlm.nih.gov/

[2] https://www.ncbi.nlm.nih.gov/books/NBK25497/

[3] https://developers.openai.com/api/docs/models/gpt-5

and normalization (Section 4.2), construction of the temporal frequency delta matrix (Section 4.3), and identification of potential disease-drug combinations from trending topics (Section 4.4). The complete pipeline is illustrated in Figure 1.

4.1 PICO Extraction

We use an LLM-based approach to identify the Population, Intervention, Comparison, and Outcome (PICO) components in abstracts. For each abstract, the LLM first determines whether the study concerns a human population; only if this criterion is satisfied does it proceed to extract the PICO components. This design integrates dataset screening with information extraction in a single process.

To ensure high accuracy and mitigate potential hallucinations, we employ a one-shot prompting strategy with multi-step reasoning. Before producing the final output, the LLM is also instructed to generate intermediate reasoning steps. This approach helps ensure that the extracted PICO components are grounded in the terminology used in each abstract. The detailed prompt provided in Appendix A and the example of the output is provided in Appendix G.

4.2 Extraction and Normalization of Disease and Drug Entities

Based on the extracted PICO components, we identify structured biomedical entities for downstream analysis. The `P` component identifies the underlying pathology or patient condition, while the `I` component captures the therapeutic agents evaluated in the study.

To ensure standardization and interoperability, we use an LLM-based approach for entity extraction and normalization. In this approach, the same LLM is used for both entity extraction and ontology mapping. First, we extract all the disease and drug entities (therapeutic agents) mentioned in the `P` and `I` components. These entities are then mapped to established biomedical ontologies: disease entities to the International Classification of Diseases (ICD-10, version 2019) and drug entities to the Anatomical Therapeutic Chemical (ATC) classification.

Both ontologies provide hierarchical structures, enabling each entity to be represented at multiple levels of granularity, with deeper levels corresponding to more specific clinical concepts. In the mapping process, we begin at the most specific category level. If no match is found at that level, the LLM escalates to the next higher category, continuing upward until a suitable match is identified. Regardless of the level at which the match is found, all hierarchical levels above it are retained, preserving the full ontological path of each entity. To balance clinical specificity with statistical robustness, we standardize the granularity of our variables, as detailed in Figures 3 (ICD-10) and 4 (ATC) in Appendix H. This approach enables a high-resolution view of therapeutic trends while avoiding over-segmentation or non-representative categories.

This LLM-driven matching process enables flexible semantic normalization of noisy biomedical text, including synonym resolution and implicit concept alignment, without relying on explicit string matching or rule-based heuristics. For further information, refer to Appendix B.

4.3 Temporal Frequency Delta Matrix

To investigate the evolution of research priorities, we divide the dataset into two distinct five-year cohorts: an early period (2013–2017) and a recent period (2018–2022).

By segmenting the data into these time windows, we quantify shifts in scientific interest over a decade and identify therapeutic interventions that have gained or declined in prominence.

We then construct a matrix to capture the frequency of disease-drug combinations within each period. Based on these matrices, we define the primary metric for our heatmap, the *Temporal Frequency Delta* (ΔF), which represents the net change in publication volume for a given pairing between the two cohorts.

$$\Delta F(P, I) = \sum_{t=C+T}^{C+2T-1} F_t(P, I) - \sum_{t=C}^{C+T-1} F_t(P, I) \quad (1)$$

where C denotes the starting year (set to 2013 in this study), and T denotes the length of each time window (set to 5 years). $F_t(P, I)$ denotes the number of unique publications discussing a specific disease–drug pair in year t, where P and I denote the population (disease) and intervention (drug), respectively. This formulation captures the temporal trend as the difference in publication counts between the two periods (2013-2017 and 2018-2022).

The ΔF value can be positive, negative, or zero:

- **Positive** ΔF indicates trending therapeutic combinations, where research activity has increased in the recent period. These regions are the primary focus of our LLM-based analysis.
- **Negative** ΔF indicates combinations with declining research activity over time.

- **Zero** ΔF indicates no change in the number of publications for a given disease–drug pairing between the two periods.

Finally, using the combinations of P and I, we define the *temporal frequency delta* matrix, $\Delta F \in \mathbb{R}^{n \times m}$, as follows:

$$\begin{pmatrix} \Delta F(P_1, I_1) & \Delta F(P_1, I_2) & \cdots & \Delta F(P_1, I_m) \\ \Delta F(P_2, I_1) & \Delta F(P_2, I_2) & \cdots & \Delta F(P_2, I_m) \\ \vdots & \vdots & \ddots & \vdots \\ \Delta F(P_n, I_1) & \Delta F(P_n, I_2) & \cdots & \Delta F(P_n, I_m) \end{pmatrix} \quad (2)$$

The lists of all $P_{1..n}$ and all $I_{1..m}$ are provided in Appendix E and F. In this study, $n = 51$ represents the number of diseases analyzed (focused on Neoplasm). While $m = 256$ denotes the total number of drugs analyzed.

4.4 Identification of Candidate Disease–Drug Combinations from Emerging Trends

To identify candidate combinations for future research, the proposed method uses a structured, matrix-based representation derived from temporal publication signals. This design enables a controlled evaluation of how input formulation influences an LLM's ability to synthesize meaningful and novel research areas. The method is provided with tabular data consisting of P–I pairs and *temporal frequency deltas* (ΔF), which capture shifts in research priorities.

The LLM is instructed to select the top-k ($k = 10$ in this study) most promising P–I combinations based on four scoring criteria: growth over the past five years, translational plausibility, clinical relevance, and strategic novelty. Each criterion is weighted and aggregated into a composite impact score ranging from 0 to 100. All outputs are generated by the LLM, not only to identify high-potential combinations but also to reduce the risk of hallucination. Further details are provided in Appendix C.

5 Experiment Setup

5.1 Model Implementation

We employed GPT-5-2025-08-07 (OpenAI), a general-purpose LLM with strong reasoning capabilities, as the primary model in our pipeline to support structured information extraction and multi-step reasoning. The model was applied across three key stages: PICO extraction (Section 4.1), extraction and normalization of disease and drug entities (Section 4.2), and identification of candidate disease–drug combinations from emerging trends (Section 4.4).

By utilizing a consistent LLM architecture across these stages, we ensured semantic continuity and reduced the risk of formatting discrepancies that often occur when switching between different models.

5.2 Domain Selection

To provide a focused evaluation setting, this analysis is restricted to neoplastic diseases. After the entity normalization stage, only records with P disease entities mapped to the ICD-10 chapter "Neoplasms" were retained for analysis. This constraint was applied solely for experimental purposes and does not limit the general applicability of the proposed method. We are only focusing on neoplastic diseases as the representative domain due to their extensive research, well-defined disease taxonomy, and clinical importance in drug discovery. This provides a suitable testbed for evaluating the proposed approach.

5.3 Baseline

For the baseline, we use the raw biomedical abstracts with their publication years. This means the baseline approach operates on unstructured, abstract texts, requiring the LLM to implicitly extract and reason about relevant entities, relationships, and trends in the research area. As in the proposed method, the instruction to the LLM is to select the top 10 most promising P and I combinations. The generated combination should follow the normalized version of P–I. This means the output should be in ICD-10 codes for the P component and ATC codes for the I component. The generated combinations are evaluated using 4 scoring criteria: growth over the past 5 years, translational plausibility, clinical relevance, and strategic novelty. Each criterion is weighted to produce a composite impact score ranging from 0 to 100. For more details, refer to the Appendix D.

5.4 Evaluation Procedure

The evaluation is conducted using two approaches: visualization and expert interpretation. First, evaluation is performed through the interpretation of visualization, specifically heatmaps. In the heatmap, it is possible to identify where new combinations are located. This means analyzing whether the combinations lie around trending drugs or not. If

the results are located around the red region, it suggests that the model successfully captures trending drugs and may propose new diseases that can be treated using those drugs (Section 6.1). Second, a qualitative evaluation was performed by a domain expert and co-author of this paper with expertise in pharmaceutical science to determine whether the results are medically meaningful (Section 6.2).

6 Results and Discussion

6.1 Heatmap Results

We present a heatmap showing all disease–drug combinations along with the candidate combinations identified by the LLM. Out of all possible combinations of 51 diseases and 256 drugs, the proposed method identifies 32 unique disease–drug pairs. These results form the basis for the heatmap, enabling comparison between the proposed method (Figures 2a and 2c) and the baseline method (Figures 2b and 2d) in terms of identified candidate combinations. The candidate combinations are highlighted in yellow, as shown in Figure 2. We focus on candidate combinations that correspond to emerging trends (i.e., the red regions in the heatmap).

A comparison with the baseline reveals that the baseline approach identifies some candidate combinations that do not overlap with previously observed trends. Specifically, among the top 10 combinations identified by the baseline, only 2 overlap with the red regions, indicating limited alignment with emerging trends.

In contrast, the proposed method produces candidate combinations that are largely consistent with emerging trends: nearly all of the top 10 combinations overlap with the red regions. This suggests that the proposed method more effectively captures captures temporally grounded research dynamics more effectively a tabular comparison of the top 10 most plausible combinations in Table 2.

6.2 Qualitative Clinical Interpretation

Both the proposed and baseline methods show a tendency for the extracted drug–disease combinations to be concentrated in oncology, particularly in hematologic malignancies. This suggests that both methods may, to some extent, reflect the broad structure of current anticancer treatment. At the same time, however, the differences between the two methods were limited, making it difficult to draw a clear conclusion about the superiority of the proposed method based on these results alone.

When examined in more detail, the baseline included drug classes such as systemic antifungals among its top-ranked outputs, which are more commonly used for supportive care, such as infection prevention, rather than as direct anticancer therapies. This suggests that when inference is based on aggregated categories, drug classes that are not central to disease treatment may be included in the results. In contrast, the proposed method included drug classes such as corticosteroids and alkylating agents for hematologic malignancies, which are more closely aligned with the context of anticancer treatment. This may indicate that, at least in some cases, the proposed method is better at prioritizing candidates with higher clinical plausibility. However, combinations that remain difficult to interpret, such as ocular vascular disorder agents for carcinoma in situ of the digestive organs, were also observed. Therefore, it cannot be concluded that the proposed method consistently extracts meaningful candidates.

7 Conclusion & Future Work

This study aims to identify emerging areas in drug discovery by leveraging LLMs. Our basic idea is to decompose the research into `P` (patient) and `I` (intervention) combinations. Under this paradigm, given the limitation of context length, handling the entire research trend is an issue. Our solution is to represent the trend in a matrix (we all *temporal frequency delta* matrix) to incorporate it with in-context learning.

As a result, it demonstrates that while the baseline approach is more likely to suggest novel, non-overlapping 'P' and 'I' combinations, the proposed approach excels at pattern recognition within established "trending" zones. Expert interpretation confirms that, although the LLM successfully prioritizes clinically plausible treatments such as alkylating agents for hematologic malignancies, it can still produce improbable pairings, suggesting that the model currently functions best as a high-level exploratory tool for visualizing therapeutic landscapes. This expert evaluation also shows that the idea of using multiple reasoning and CIS to mitigate hallucination does not fully eliminate the risk of hallucinated outputs.

This highlights that transparency and reliability remain open challenges in LLM-based scientific discovery frameworks. Future work can also ex-

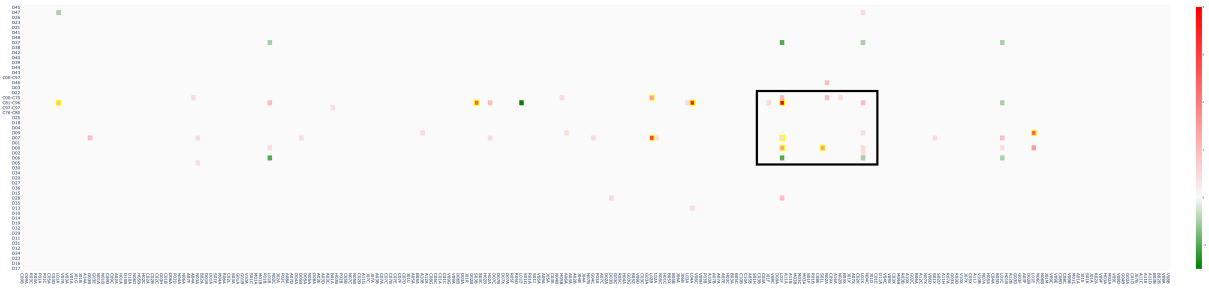
Rank	Proposed		Baseline	
	Disease (ICD-10)	Drugs (ATC)	Disease (ICD-10)	Drugs (ATC)
1	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L01F]** monoclonal antibodies and antibody drug conjugates	**[D07]** carcinoma in situ of other and unspecified genital organs	**[G03E]** androgens and female sex hormones in combination
2	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L04]** immunosuppressants	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[L01E]** protein kinase inhibitors
3	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[L02B]** hormone antagonists and related agents	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[L01F]** monoclonal antibodies and antibody drug conjugates
4	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[H02]** corticosteroids	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L01E]** protein kinase inhibitors
5	**[D00]** carcinoma in situ of the oral cavity, esophagus, and stomach	**[L01F]** monoclonal antibodies and antibody drug conjugates	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[D01]** antifungals for systemic use
6	**[D01]** carcinoma in situ of other and unspecified genital organs	**[L02B]** hormone antagonists and related agents	**[D01]**carcinoma in situ of other and unspecified digestive organs	**[L01E]** protein kinase inhibitors
7	**[D01]** carcinoma in situ of other and unspecified genital organs	**[L01F]** monoclonal antibodies and antibody drug conjugates	**[C81–C96** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[L01F]** monoclonal antibodies and antibody drug conjugates
8	**[D07]** carcinoma in situ of other and unspecified sites	**[L01E]** protein kinase inhibitors	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic, and related tissue	**[C03D]** aldosterone antagonists and other potassium-sparing agents
9	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic, and related tissue	**[LO1A]** alkylating agents	**[D07]** carcinoma in situ of other and unspecified genital organs	**[G03B]** androgens
10	**[D00]** carcinoma in situ of the oral cavity, esophagus, and stomach	**[S01L]** ocular vascular disorder agents	**[C81–C96]** malignant neoplasms, stated or presumed to be primary, of specified sites, except for lymphoid, hematopoiesis, and related tissue	**[L04A]** immunosuppressants

Table 2: Comparison between the proposed approach and baseline approach with promising rank. Each entry consists of a code and its corresponding category name (ICD-10 for diseases and ATC for drugs). The rows with a gray background represent the combinations of the baseline and proposed approach that have the same output.

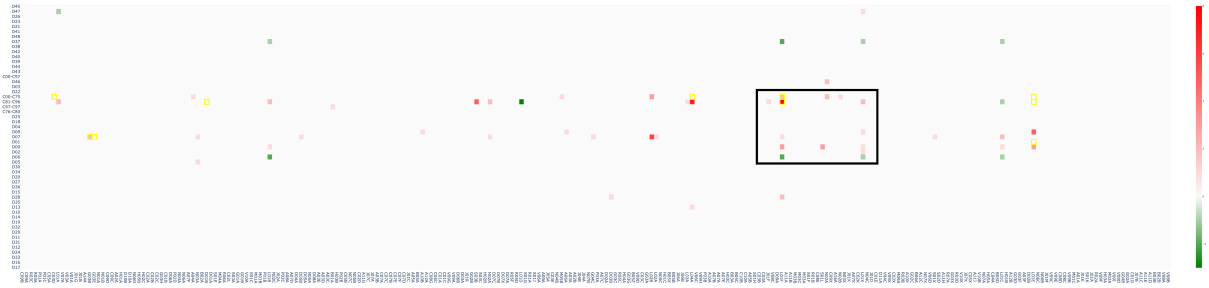
tend the analysis to other form of output such as the exact PICO output so we can evaluate the quality and the correctness of the idea generated by LLM.

8 Limitations

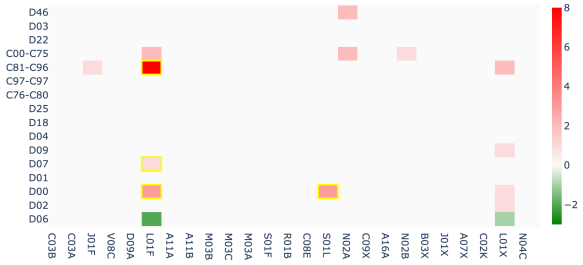
There are several limitations: First, we focus only on PubMed papers, specifically on cancer abstracts, which limits the generalizability. Second, the mapping process still needs further investigation to de-



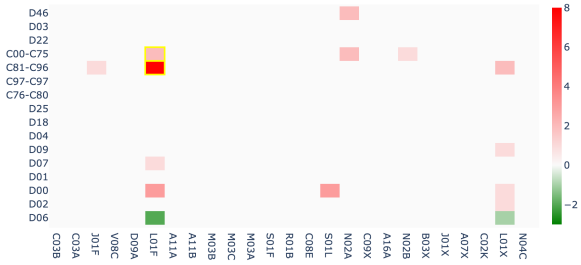
(a) Proposed method



(b) Baseline method



(c) Zoomed-in view of a selected region in (a)



(d) Zoomed-in view of a selected region in (b)

Figure 2: Heatmaps of disease (ICD-10)–drug (ATC) combinations based on temporal frequency delta (ΔF). The vertical axis represents diseases (ICD-10) and the horizontal axis represents drugs (ATC). Each cell corresponds to the ΔF value for a disease–drug pair. Red indicates combinations with increasing research activity (positive ΔF), while green indicates combinations with decreasing activity (negative ΔF); light gray indicates zero change ($\Delta F = 0$). Yellow boxes highlight top-ranked candidate combinations based on composite impact scores.

termine the optimal level of categories used for the normalization process. Third, the *temporal frequency delta* matrix ranged from -3 to 8, which may be sensitive to cumulative errors throughout the pipeline. Therefore, hallucination in each step should be considered; there may be a chance of hallucination along the way. Finally, in the evaluation framework, we only used qualitative methods that leads to missing information regarding how is the quality, the correctness, and the novelty of the new research combination. This can also be future research focusing on the evaluation using quantitative methods. Beyond these limitations, future work can be applied across more papers and sources and expanded beyond a focus on cancer. Also, the LLM models used in this research can be further explored selecting models from the medical domain is another option.

Acknowledgements

This research was supported by JST CREST JPMJCR22N1.

References

Nur Sabrina Azmi and Weng Howe Chan. 2025. Recent trends on multi-omics studies in cancer research: A bibliometric study. *International Journal of Innovative Computing*, 15(2):149–158.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas*

Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6709–6738.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Babu Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A Huerta, and Hao Peng. 2025. Context length alone hurts LLM performance despite perfect retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23281–23298.

Alireza Ghafarollahi and Markus J Buehler. 2025. SciAgents: automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, and 15 others. 2025. Towards an AI co-scientist. *Preprint*, arXiv:2502.18864.

Rezarta Islamaj, Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Tiago Almeida, Richard A A Jonker, Sofia I R Conceição, Diana F Sousa, Cong-Phuoc Phan, Jung-Hsien Chiang, Jiru Li, Dinghao Pan, Wilailack Meesawad, Richard Tzong-Han Tsai, M Janina Sarol, Gibong Hong, Airat Valiev, Elena Tutubalina, Shao-Man Lee, and 4 others. 2024. The overview of the BioRED (Biomedical Relation Extraction Dataset) track at BioCreative VIII. *Database*, 2024:baae069.

Yoonbee Kim, Yi-Sue Jung, Jong-Hoon Park, Seon-Jun Kim, and Young-Rae Cho. 2022. Drug-Disease Association Prediction Using Heterogeneous Networks for Computational Drug Repositioning. *Biomolecules*, 12(10):1497.

Zhijing Li, Yunwen Yu, Wenhao Gu, Tiantian Zhu, Haohua Song, Wenbin Guo, Xiao Yang, and Zexuan Zhu. 2025. Dual-LLM Adversarial Framework for Information Extraction from Research Literature. *Preprint*, bioRxiv:2025.09.11.675507.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers. In *Proceedings of the NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Wenhao Qi, Xiaohong Zhu, Danni He, Bin Wang, Shihua Cao, Chaoqun Dong, Yunhua Li, Yanfei Chen, Bingsheng Wang, Yankai Shi, Guowei Jiang, Fang Liu, Lizzy M M Boots, Jiaqi Li, Xiajing Lou, Jiani Yao, Xiaodong Lu, and Junling Kang. 2024. Mapping knowledge landscapes and emerging trends in AI for dementia biomarkers: bibliometric and visualization analysis. *Journal of Medical Internet Research*, 26:e57830.

Justin Sybrandt, Michael Shtutman, and Ilya Safro. 2017. MOLIERE: Automatic Biomedical Hypothesis Generation System. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, page 1633–1642.

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A Persson, Gerbrand Ceder, and Anubhav Jain. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763):95–98.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(509):1–14.

Ziheng Wang, Yang Zhao, and Lin Zhang. 2024. Emerging trends and hot topics in the application of multi-omics in drug discovery: A bibliometric and visualized study. *Current Pharmaceutical Analysis*, 21(1):20–32.

Norman R Williams. 2026. Old drugs, new opportunities: Advancing cancer care through repurposing. *United European Gastroenterology Journal*, 14(1):e70184.

Chao Yan, Monika E Grabowska, Alyson L Dickson, Bingshan Li, Zhexing Wen, Dan M Roden, C Michael Stein, Peter J Embí, Josh F Peterson, QiPing Feng, Bradley A. Malin, and Wei-Qi Wei. 2024. Leveraging generative AI to prioritize drug repurposing candidates for Alzheimer's disease with real-world clinical validation. *npj Digital Medicine*, 7(46):1–6.

Rui Yang, Ting Fang Tan, Wei Lu, Arun James Thirunavukarasu, Daniel Shu Wei Ting, and Nan Liu. 2023. Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.

Zhichao Yao, Zhiyong Lin, and Weijia Wu. 2023. Global research trends on immunotherapy in cancer: A bibliometric analysis. *Human Vaccines & Immunotherapeutics*, 19(2):2219191. PMID: 37314453.

A Details of Filtering and PICO Extraction via LLM

LLM Prompt for PICO Extraction

```
You are an expert biomedical research analyst.

The objective is to extract structured PICO/PECO
elements from biomedical abstracts while filtering out
non-human studies.

TASK:
Extract concise PICO/PECO elements and return a single
JSON object.

REQUIREMENTS:
- Output MUST contain exactly these keys: "Population",
```

```
"Intervention/Exposure", "Comparison", "Outcome", and
"Explain".
- Each field MUST contain 1–2 concise sentences using
canonical clinical phrasing.
- Do NOT include citations or bracketed identifiers.

HUMAN SUBJECT FILTER:
- If the study does NOT involve human subjects (e.g.,
animal studies, in vitro, biomedical experiments), set
ALL fields to "NA".
- Provide a clear justification in the "Explain" field.

STRICT RULES:
- Do NOT hallucinate or infer missing information.
- If any PICO element is not explicitly stated, write
"NA".
- Ensure all outputs are directly grounded in the
abstract.

WORKFLOW:
1. Perform initial screening for human subjects.
2. Identify study design, treatment arms, and
comparators.
3. Map findings into PICO/PECO components.
4. Validate that all required JSON keys are present
and correctly formatted.

OUTPUT FORMAT (JSON ONLY):

{
"Population": "...",
"Intervention/Exposure": "...",
"Comparison": "...",
"Outcome": "...",
"Explain": "..."
}

INPUT DATA:
{abstracts}
```

B Prompts for Entity Extraction & Normalization

PICO NER Prompt

```
You are a professional biomedical Named Entity
Recognition (NER) system specialized in PICO-based
extraction.

TASK:
Extract structured biomedical entities from the input
text:
- P (Population/Problem): underlying diseases or
patient conditions
- I (Intervention): therapeutic agents (drugs,
biologics)

REQUIREMENTS:
- Extract ONLY entities explicitly mentioned in the
text.
- DO NOT infer or hallucinate entities.
- Resolve abbreviations ONLY if clearly defined in the
text.
- Return ONLY the expanded (full) form.

ENTITY DEFINITIONS:
- Disease (P): pathological conditions, diagnoses,
syndromes
- Drug (I): therapeutic agents, drugs, biologics

EXCLUDE:
- Procedures, measurements, symptoms, and general terms

REASONING REQUIREMENTS:
- Each entity MUST include a "reason".
- The reason MUST be directly grounded in the text.
- Prefer exact phrase or minimal paraphrase from the
text.
- DO NOT use external knowledge.
- Maximum 15 words.

STRICT CONSTRAINTS:
- DO NOT include entities that are not directly grounded
in the text.
- DO NOT guess abbreviation expansions unless clearly
supported.
- If no entities exist, return empty lists.

VALIDATION STEP (INTERNAL):
1. Identify candidate entities directly from the text.
2. Remove duplicates and abbreviations.
3. Ensure each entity appears explicitly.
4. Ensure each reason is grounded in the text.

INPUT:
{text}

OUTPUT FORMAT (JSON ONLY):
{
"Disease": [
{"name": "...", "reason": "..."}
],
"Drug": [
{"name": "...", "reason": "..."}
]
}
```

ATC Drug Normalization Prompt

```
You are an expert biomedical terminologist performing
drug normalization using ATC classification.

TASK:
Map the input drug to the MOST SPECIFIC ATC concept
using the full hierarchy.

INPUT:
"{term}"

ONTOLOGY STRUCTURE:
- ATC hierarchy: Level 1 → Level 2 → Level 3 → Level
4 → Level 5
- Level 5 = most specific (chemical substance)

REQUIREMENTS:
- Search across ALL ATC levels
- START from Level 5 (most specific)
- If no match, move upward level-by-level
- Select the MOST SPECIFIC semantically equivalent
concept

SEMANTIC MATCHING:
- Accept synonyms, brand names, and drug combinations
- For combination drugs → choose dominant therapeutic
class

STRICT RULES:
- DO NOT choose a broader class if a specific match
exists
- DO NOT assign incorrect therapeutic class
- DO NOT hallucinate mappings
- If no valid match → return "None"

VALIDATION (INTERNAL):
1. Identify active ingredient or drug class
2. Match against candidates (all levels)
3. Start from Level 5 → move upward
4. Ensure therapeutic equivalence
5. Confirm no more specific match exists

OUTPUT FORMAT (STRICT JSON):
{
"input": "{term}",
"match": {
"code": "...",
"name": "...",
"hierarchy": [
{"level": "level1", "code": "...", "name": "..."},
{"level": "level2", "code": "...", "name": "..."},
{"level": "level3", "code": "...", "name": "..."},
{"level": "level4", "code": "...", "name": "..."},
{"level": "level5", "code": "...", "name": "..."}
]
}
}
```

```
CANDIDATES (ALL LEVELS):
{candidates}
```

ICD-10 Disease Normalization Prompt

```
You are an expert biomedical terminologist performing
disease normalization using ICD-10 (2019).

TASK:
Map the input disease to the MOST SPECIFIC matching
ICD-10 concept using the full hierarchy.

INPUT:
"{term}"

ONTOLOGY STRUCTURE:
- ICD-10 is hierarchical Level 1 → Level 2 → Level 3
→ Level 4 → Level 5 → Level 6 → Level 7
- Lower levels = more specific concepts

REQUIREMENTS:
- Search across ALL levels of ICD-10 candidates
- START from the LOWEST (most specific) level
- If no exact match at lower level, move upward
step-by-step
- Select the MOST SPECIFIC semantically equivalent
concept

SEMANTIC MATCHING:
- Accept synonyms and clinical rephrasings
- Use meaning, NOT string similarity

STRICT RULES:
- DO NOT select a broader category if a more specific
match exists
- DO NOT select a narrower category if it changes
meaning
- DO NOT hallucinate mappings
- If no valid match exists → return "None"

VALIDATION (INTERNAL):
1. Identify the core disease concept
2. Compare against candidates (all levels)
3. Start from the most specific level
4. Ensure semantic equivalence (not related, not
approximate)
5. Confirm no better (more specific) match exists

OUTPUT FORMAT (STRICT JSON):
{
"input": "{term}",
"match": {
"code": "...",
"name": "...",
"hierarchy": [
{"level": "chapter", "code": "...", "name": "..."},
{"level": "block", "code": "...", "name": "..."},
{"level": "category", "code": "...", "name": "..."},
{"level": "subcategory", "code": "...", "name": "..."}
]
}
}

CANDIDATES (ALL LEVELS):
{candidates}
```

C Prompts for Identifying Candidate Combination

Proposed Method Prompt

```
You are analyzing structured trend data derived from
publication dynamics.

Each row contains:
- Disease_ICD_L4
- Drug_ATC_L3
- Delta_Publications

DELTA = (2018–2022 publications) − (2013–2017
publications)

Meaning:
Positive → increasing research activity
Zero → no change
Negative → declining research

INPUT DATA:
{delta_json_text}

TASK:
Select the TOP 10 most promising Problem–Intervention
(P–I) combinations for future research.

SELECTION CRITERIA:
- Strong positive growth over the past 5 years
- Translational plausibility
- Clinical relevance
- Strategic novelty

SCORING SCHEME (Impact Score: 0–100):
- Growth intensity (0–30)
- Translational plausibility (0–25)
- Clinical relevance (0–25)
- Strategic novelty (0–20)

STRICT RULES:
- You MUST use EXACT labels as provided in the dataset.
- Do NOT invent, modify, or normalize labels.
- All reasoning MUST be grounded in the provided data.

OUTPUT FORMAT (JSON ONLY):

[
{
"P-Group": "...",
"I-Group": "...",
"Impact Score": 0–100,
"Insight": "Max 2 sentences reasoning"
}
]
```

D Prompts for Baseline

Baseline Prompt

```
You are analyzing biomedical abstracts along with their
publication years.

Each abstract includes a publication year. You MUST
use this temporal information to identify trends in
research activity.

The objective is to identify potential
Problem–Intervention (P–I) combinations for future
research.

TASK:
Select the TOP 10 most promising P and I combinations.

REQUIREMENTS:
- P MUST be normalized to ICD-10 Level 3 and MUST be
selected ONLY from the following valid disease labels:
{disease_labels}.
- Do NOT generate new or modified labels.
- I MUST be normalized to ATC Level 4 and MUST be
selected ONLY from the following valid drug labels:
{drug_labels}.
- Do NOT generate new or modified labels.

TEMPORAL CONSTRAINT:
- You MUST explicitly use publication years to infer
research growth.
- Compare the most recent 5 years with the earlier 5
years.
- If no temporal signal is observable, do NOT prioritize
that combination.

SCORING SCHEME (Impact Score: 0–100):
- Growth intensity (0–30)
- Translational plausibility (0–25)
- Clinical relevance (0–25)
- Strategic novelty (0–20)

STRICT RULES:
```

```
- Do NOT invent unsupported entities.
- All reasoning MUST be grounded in the abstracts AND
their publication years.

OUTPUT FORMAT (JSON ONLY):
{
"P-Group": "ICD-10 Level 3",
"I-Group": "ATC Level 4",
"Impact Score": 0-100,
"Insight":    "Max  2  sentences  including  temporal
reasoning"
}


INPUT DATA:
{abstracts}
```

E List of ICD-10 Code

ICD-10 2019 Codes

Code	Category	Code	Category
C00-C75	Malignant neoplasms, stated or presumed to be primary, of specified sites, except of lymphoid, hematopoietic and related tissue	C81-C96	Malignant neoplasms, stated or presumed to be primary, of lymphoid, hematopoietic and related tissue
C76-C80	Malignant neoplasms of ill-defined, secondary, and unspecified sites	D48	Neoplasm of uncertain or unknown behavior of other and unspecified sites
C97-C97	Malignant neoplasms of independent (primary) multiple sites	D00	Carcinoma in situ of oral cavity, esophagus, and stomach
D01	Carcinoma in situ of other and unspecified digestive organs	D02	Carcinoma in situ of the middle ear and respiratory system
D03	Melanoma in situ	D04	Carcinoma in situ of skin
D05	Carcinoma in situ of the breast	D06	Carcinoma in situ of cervix uteri
D07	Carcinoma in situ of other and unspecified genital organs	D09	Carcinoma in situ of other and unspecified sites
D10	Benign neoplasm of the mouth and pharynx	D11	Benign neoplasm of major salivary glands
D12	Benign neoplasm of colon, rectum, anus, and anal canal	D13	Benign neoplasm of other and ill-defined parts of the digestive system
D14	Benign neoplasm of the middle ear and respiratory system	D15	Benign neoplasm of other and unspecified intrathoracic organs
D16	Benign neoplasm of bone and articular cartilage	D17	Benign lipomatous neoplasm
D18	Haemangioma and lymphangioma, any site	D19	Benign neoplasm of mesothelia tissue
D20	Benign neoplasm of soft tissue of the retroperitoneum and peritoneum	D21	Other benign neoplasms of connective and other soft tissue
D22	Melanocytic naevi	D23	Other benign neoplasms of skin
D24	Benign neoplasm of breast	D25	Leiomyoma of uterus
D26	Other benign neoplasms of uterus	D27	Benign neoplasm of ovary
D28	Benign neoplasm of other and unspecified female genital organs	D29	Benign neoplasm of male genital organs
D30	Benign neoplasm of urinary organs	D31	Benign neoplasm of eye and adnexa
D32	Benign neoplasm of meninges	D33	Benign neoplasm of brain and other parts of the central nervous system
D34	Benign neoplasm of thyroid gland	D35	Benign neoplasm of other and unspecified endocrine glands
D36	Benign neoplasm of other and unspecified sites	D37	Neoplasm of uncertain or unknown behavior of the oral cavity and digestive organs
D38	Neoplasm of uncertain or unknown behavior of the middle ear and respiratory and intrathoracic organs	D39	Neoplasm of uncertain or unknown behavior of female genital organs
D40	Neoplasm of uncertain or unknown behavior of male genital organs	D41	Neoplasm of uncertain or unknown behavior of urinary organs
D42	Neoplasm of uncertain or unknown behavior of meninges	D43	Neoplasm of uncertain or unknown behavior of brain and central nervous system
D44	Neoplasm of uncertain or unknown behavior of endocrine glands	D45	Polycythaemia vera
D46	Myelodysplastic syndromes	D47	Other neoplasms of uncertain or unknown behavior of lymphoid, hematopoietic and related tissue

F List of ATC Classification Codes

ATC Classification			
Code	**Description**	**Code**	**Description**
A01A	Stomatological Preparations	A02A	Antacids
A02B	Drugs for Peptic Ulcer and GORD	A03A	Drugs for Functional Gastrointestinal Disorders
A03B	Belladonna and Derivatives, Plain	A03C	Antispasmodics in Comb. with Psycholeptics
A03D	Antispasmodics in Comb. with Analgesics	A03E	Antispasmodics and Anticholinergics in Comb.
A03F	Propulsives	A04A	Antiemetics and Antinauseants
A05A	Bile Therapy	A05B	Liver Therapy, Lipotropics
A06A	Drugs for Constipation	A07A	Intestinal Antiinfectives
A07B	Intestinal Adsorbents	A07C	Electrolytes with Carbohydrates
A07D	Antipropulsives	A07E	Intestinal Antiinflammatory Agents
A07F	Antidiarrheal Microorganisms	A07X	Other Antidiarrheals
A08A	Antiobesity Preparations	A09A	Digestives, Incl. Enzymes
A10A	Insulins and Analogues	A10B	Blood Glucose Lowering Drugs, Excl. Insulins
A10X	Other Drugs Used in Diabetes	A11A	Multivitamins, Combinations
A11B	Multivitamins, Plain	A11C	Vitamin A and D, Incl. Combinations
A11D	Vitamin B1, Plain and Comb.	A11E	Vitamin B-Complex, Incl. Combinations
A11G	Ascorbic Acid (Vitamin C), Incl. Comb.	A11H	Other Plain Vitamin Preparations
A11J	Other Vitamin Products, Combinations	A12A	Calcium
A12B	Potassium	A12C	Other Mineral Supplements
A14A	Anabolic Steroids	A16A	Other Alimentary Tract and Metabolism Products
B01A	Antithrombotic Agents	B02A	Antifibrinolytics
B02B	Vitamin K and Other Hemostatics	B03A	Iron Preparations
B03B	Vitamin B12 and Folic Acid	B03X	Other Antianemic Preparations
B05A	Blood and Related Products	B05B	I.V. Solutions
B05C	Irrigating Solutions	B05D	Peritoneal Dialytics
B05X	I.V. Solution Additives	B05Z	Hemodialytics and Hemofiltrates
B06A	Other Hematological Agents	C01A	Cardiac Glycosides
C01B	Antiarrhythmics, Class I and III	C01C	Cardiac Stimulants Excl. Cardiac Glycosides
C01D	Vasodilators Used in Cardiac Diseases	C01E	Other Cardiac Preparations
C02A	Antiadrenergic Agents, Centrally Acting	C02B	Antiadrenergic Agents, Ganglion-Blocking
C02C	Antiadrenergic Agents, Peripherally Acting	C02D	Arteriolar Smooth Muscle, Agents Acting On
C02K	Other Antihypertensives	C02L	Antihypertensives and Diuretics in Comb.
C03A	Low-Ceiling Diuretics, Thiazides	C03B	Low-Ceiling Diuretics, Excl. Thiazides
C03C	High-Ceiling Diuretics	C03D	Aldosterone Antagonists and K-Sparing Agents
C03E	Diuretics and K-Sparing Agents in Comb.	C03X	Other Diuretics
C04A	Peripheral Vasodilators	C05A	Agents for Hemorrhoids (Topical)
C05B	Antivaricose Therapy	C05C	Capillary Stabilizing Agents
C05X	Other Vasoprotectives	C07A	Beta Blocking Agents
C07B	Beta Blocking Agents and Thiazides	C07C	Beta Blocking Agents and Other Diuretics
C07D	Beta Blocking Agents, Thiazides and Diuretics	C07E	Beta Blocking Agents and Vasodilators
C07F	Beta Blocking Agents, Other Comb.	C08C	Selective Calcium Channel Blockers (Vascular)
C08D	Selective Calcium Channel Blockers (Cardiac)	C08E	Non-Selective Calcium Channel Blockers
C08G	Calcium Channel Blockers and Diuretics	C09A	ACE Inhibitors, Plain
C09B	ACE Inhibitors, Combinations	C09C	Angiotensin II Receptor Blockers (ARBs)
C09D	ARBs, Combinations	C09X	Other Renin-Angiotensin System Agents
C10A	Lipid Modifying Agents, Plain	C10B	Lipid Modifying Agents, Combinations
D01A	Antifungals for Topical Use	D01B	Antifungals for Systemic Use
D02A	Emollients and Protectives	D02B	Protectives Against UV-Radiation
D03A	Cicatrizants	D03B	Enzymes
D04A	Antipruritics	D05A	Antipsoriatics for Topical Use
D05B	Antipsoriatics for Systemic Use	D06A	Antibiotics for Topical Use
D06B	Chemotherapeutics for Topical Use	D07A	Corticosteroids, Plain
D07B	Corticosteroids, Comb. with Antiseptics	D07C	Corticosteroids, Comb. with Antibiotics
D07X	Corticosteroids, Other Combinations	D08A	Antiseptics and Disinfectants
D09A	Medicated Dressings	D10A	Anti-Acne Preparations for Topical Use
D10B	Anti-Acne Preparations for Systemic Use	D11A	Other Dermatological Preparations
G01A	Antiinfectives and Antiseptics (Vaginal)	G01B	Antiinfectives with Corticosteroids
G02A	Uterotonics	G02B	Contraceptives for Topical Use
G02C	Other Gynecologicals	G03A	Hormonal Contraceptives for Systemic Use
G03B	Androgens	G03C	Estrogens
G03D	Progestogens	G03E	Androgens and Female Sex Hormones in Comb.
G03F	Progestogens and Estrogens in Combination	G03G	Gonadotropins and Ovulation Stimulants
G03H	Antiandrogens	G03X	Other Sex Hormones
G04B	Urologicals	G04C	Drugs Used in Benign Prostatic Hypertrophy
H01A	Anterior Pituitary Lobe Hormones	H01B	Posterior Pituitary Lobe Hormones
H01C	Hypothalamic Hormones	H02A	Corticosteroids for Systemic Use, Plain
H02B	Corticosteroids for Systemic Use, Comb.	H02C	Antiadrenal Preparations
H03A	Thyroid Preparations	H03B	Antithyroid Preparations
H03C	Iodine Therapy	H04A	Glycogenolytic Hormones
H05A	Parathyroid Hormones	H05B	Anti-Parathyroid Agents
J01A	Tetracyclines	J01B	Amphenicols
J01C	Beta-Lactam Antibacterials, Penicillins	J01D	Other Beta-Lactam Antibacterials

ATC Classification

Code	Description	Code	Description
J01E	Sulfonamides and Trimethoprim	J01F	Macrolides and Lincosamides
J01G	Aminoglycoside Antibacterials	J01M	Quinolone Antibacterials
J01R	Combinations of Antibacterials	J01X	Other Antibacterials
J02A	Antimycotics for Systemic Use	J04A	Drugs for Treatment of Tuberculosis
J04B	Drugs for Treatment of Lepra	J05A	Direct Acting Antivirals
J06A	Immune Sera	J06B	Immunoglobulins
J07A	Bacterial Vaccines	J07B	Viral Vaccines
J07C	Bacterial and Viral Vaccines, Combined	J07X	Other Vaccines
L01A	Alkylating Agents	L01B	Antimetabolites
L01C	Plant Alkaloids	L01D	Cytotoxic Antibiotics
L01E	Protein Kinase Inhibitors	L01F	Monoclonal Antibodies and ADCs
L01X	Other Antineoplastic Agents	L02A	Hormones and Related Agents
L02B	Hormone Antagonists	L03A	Immunostimulants
L04A	Immunosuppressants	M01A	Antiinflammatory (NSAIDs)
M01B	Antiinflammatory Agents in Comb.	M01C	Specific Antirheumatic Agents
M02A	Topical Products for Joint/Muscular Pain	M03A	Muscle Relaxants, Peripherally Acting
M03B	Muscle Relaxants, Centrally Acting	M03C	Muscle Relaxants, Directly Acting
M04A	Antigout Preparations	M05B	Drugs Affecting Bone Structure
M09A	Other Musculo-Skeletal Drugs	N01A	Anesthetics, General
N01B	Anesthetics, Local	N02A	Opioids
N02B	Other Analgesics and Antipyretics	N02C	Antimigraine Preparations
N03A	Antiepileptics	N04A	Anticholinergic Agents
N04B	Dopaminergic Agents	N04C	Other Antiparkinson Drugs
N05A	Antipsychotics	N05B	Anxiolytics
N05C	Hypnotics and Sedatives	N06A	Antidepressants
N06B	Psychostimulants and Nootropics	N06C	Psycholeptics and Psychoanaleptics Comb.
N06D	Anti-Dementia Drugs	N07A	Parasympathomimetics
N07B	Drugs Used in Addictive Disorders	N07C	Antivertigo Preparations
N07X	Other Nervous System Drugs		
P01A	Agents Against Amoebiasis	P01B	Antimalarials
P01C	Agents Against Leishmaniasis	P02B	Antitrematodals
P02C	Antinematodal Agents	P02D	Anticestodals
P03A	Ectoparasiticides	P03B	Insecticides and Repellents
R01A	Nasal Decongestants (Topical)	R01B	Nasal Decongestants (Systemic)
R02A	Throat Preparations	R03A	Adrenergics, Inhalants
R03B	Other Inhalants for Obstructive Airway	R03C	Adrenergics for Systemic Use
R03D	Other Systemic Drugs for Obstructive Airway	R05C	Expectorants
R05D	Cough Suppressants	R05F	Cough Suppressants and Expectorants
R06A	Antihistamines for Systemic Use	R07A	Other Respiratory System Products
S01A	Antiinfectives (Ophthalmic)	S01B	Antiinflammatory (Ophthalmic)
S01C	Antiinflam. and Antiinfectives Comb.	S01E	Antiglaucoma and Miotics
S01F	Mydriatics and Cycloplegics	S01G	Decongestants and Antiallergics
S01H	Local Anesthetics (Ophthalmic)	S01J	Diagnostic Agents (Ophthalmic)
S01K	Surgical Aids (Ophthalmic)	S01L	Ocular Vascular Disorder Agents
S01X	Other Ophthalmologicals	S02A	Antiinfectives (Otic)
S02B	Corticosteroids (Otic)	S02C	Corticosteroids and Antiinfectives Comb.
S02D	Other Otologicals	S03A	Antiinfectives (Ophth/Otic)
S03B	Corticosteroids (Ophth/Otic)	S03C	Corticosteroids/Antiinfectives Comb.
V01A	Allergens	V03A	All Other Therapeutic Products
V04C	Other Diagnostic Agents	V06A	Diet Formulations for Obesity
V06C	Infant Formulas	V06D	Other Nutrients
V07A	All Other Non-Therapeutic Products	V08A	X-Ray Contrast Media, Iodinated
V08B	X-Ray Contrast Media, Non-Iodinated	V08C	MRI Contrast Media
V08D	Ultrasound Contrast Media	V09A	CNS (Radiopharmaceuticals)
V09B	Skeleton (Radiopharmaceuticals)	V09C	Renal (Radiopharmaceuticals)
V09D	Hepatic (Radiopharmaceuticals)	V09E	Respiratory (Radiopharmaceuticals)
V09F	Thyroid (Radiopharmaceuticals)	V09G	Cardiovascular (Radiopharmaceuticals)
V09H	Inflammation Detection (Radiopharma)	V09I	Tumour Detection (Radiopharmaceuticals)
V09X	Other Diagnostic Radiopharma	V10A	Antiinflammatory (Therapeutic Radiopharma)
V10B	Pain Palliation – Bone Seeking	V10X	Other Therapeutic Radiopharmaceuticals

G Example of Extracted PICO Components

Table 3: An example of PICO components extracted from abstract, drugs and disease extracted from `P` & `I` components, and results mapped to ICD-10 and ATC

Field	Content
PMID	31171494
Year	2021
Abstract	Older patients are one of the most relevant sub-groups of patients with breast cancer and will only gain in importance as demographic transition unfolds. Their management, in both the early and advanced settings, should take into consideration specific clinical needs and is made more difficult by the limited availability of evidence on the efficacy and safety of standard treatment regimens in older patients. At the root of this situation is the low rate of participation of older patients in clinical trials, often due to age limits for inclusion, and limitations on the participation of persons with significant comorbidities or organ dysfunction. Although this has begun to change in recent years, most agents currently in use have not been tested in a substantial number of older patients. This includes targeted agents that have changed prognosis in breast cancer.
Population	Older adults with early or advanced breast cancer, a subgroup underrepresented in clinical trials.
Intervention	Use of targeted agents including HER2-targeted therapies (trastuzumab, lapatinib, pertuzumab, T-DM1, neratinib), CDK4/6 inhibitors (palbociclib, ribociclib, abemaciclib), bevacizumab, everolimus, and PARP inhibitors (olaparib, talazoparib).
Comparison	NA
Outcome	Efficacy and safety profiles with an older-specific toxicity and tolerability context to inform treatment decisions.
Explain	Narrative review summarizing efficacy and safety data; no explicit comparator specified.
P_Disease	Breast cancer
I_Drugs	trastuzumab, lapatinib, pertuzumab, trastuzumab emtansine, neratinib, CDK4/6 inhibitors, palbociclib, ribociclib, abemaciclib, bevacizumab, everolimus, PARP inhibitors, olaparib, talazoparib
P_Code	C00–C75
P_Name	Malignant neoplasms of specified sites (excluding lymphoid and related tissue)
I_Code	L01F, L01E, L01X, L04A
I_Name	Monoclonal antibodies, Protein kinase inhibitors, Other antineoplastic agents, Immunosuppressants

H Entity Normalization by Ontology Mapping

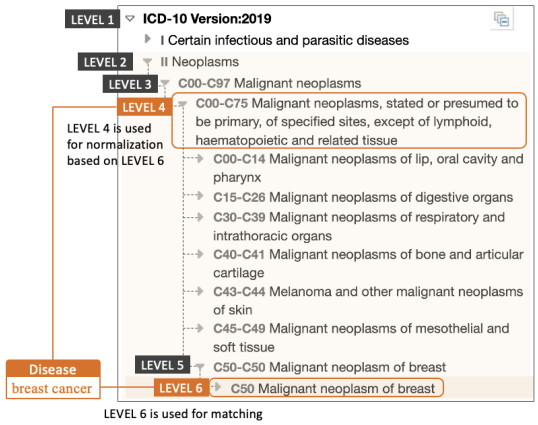


Figure 3: ICD-10 code mapping

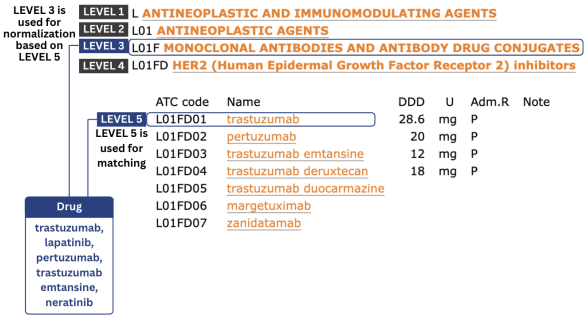


Figure 4: ATC code mapping