Database, 2024, baae102
DOI: https://doi.org/10.1093/database/baae102
Original article

CPMKG: a condition-based knowledge graph for precision medicine

Jiaxin Yang[1,‡], Xinhao Zhuang[1,‡], Zhenqi Li[2], Gang Xiong[3], Ping Xu[2,*], Yunchao Ling[1,*], Guoqing Zhang[1,4,*]

[1]National Genomics Data Center & Bio-Med Big Data Center, Chinese Academy of Sciences Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[2]Shanghai Information Center for Life Sciences, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China
[3]Shanghai Southgene Technology Co., Ltd., Shanghai 201203, China
[4]Shanghai Sixth People's Hospital, Shanghai 200233, China

[*]Corresponding authors. Guoqing Zhang. National Genomics Data Center & Bio-Med Big Data Center, Chinese Academy of Sciences Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Xuhui District, Shanghai 200031, China. E-mail: gqzhang@sinh.ac.cn; Yunchao Ling. National Genomics Data Center & Bio-Med Big Data Center, Chinese Academy of Sciences Key Laboratory of Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Xuhui District, Shanghai 200031, China. E-mail: lingyunchao@sinh.ac.cn; Ping Xu. Shanghai Information Center for Life Sciences, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Xuhui District, Shanghai 200031, China. E-mail: xuping@sinh.ac.cn.
[‡]Equal contribution.

Citation details: Yang, J., Zhuang, X., Li, Z. *et al.* CPMKG: a condition-based knowledge graph for precision medicine. *Database* (2024) Vol. 2024: article ID baae102; DOI: https://doi.org/10.1093/database/baae102

Abstract

Personalized medicine tailors treatments and dosages based on a patient's unique characteristics, particularly its genetic profile. Over the decades, stratified research and clinical trials have uncovered crucial drug-related information—such as dosage, effectiveness, and side effects—affecting specific individuals with particular genetic backgrounds. This genetic-specific knowledge, characterized by complex multirelationships and conditions, cannot be adequately represented or stored in conventional knowledge systems. To address these challenges, we developed CPMKG, a condition-based platform that enables comprehensive knowledge representation. Through information extraction and meticulous curation, we compiled 307 614 knowledge entries, encompassing thousands of drugs, diseases, phenotypes (complications/side effects), genes, and genomic variations across four key categories: drug side effects, drug sensitivity, drug mechanisms, and drug indications. CPMKG facilitates drug-centric exploration and enables condition-based multiknowledge inference, accelerating knowledge discovery through three pivotal applications. To enhance user experience, we seamlessly integrated a sophisticated large language model that provides textual interpretations for each subgraph, bridging the gap between structured graphs and language expressions. With its comprehensive knowledge graph and user-centric applications, CPMKG serves as a valuable resource for clinical research, offering drug information tailored to personalized genetic profiles, syndromes, and phenotypes.

Database URL: https://www.biosino.org/cpmkg/

Introduction

The primary challenge in drug therapy is the wide variation in individual responses to medications. This is due to differences in drug metabolism and physiological conditions [1–3]. Precision medicine, unlike the traditional one-size-fits-all approach, tailors treatments to each patient's unique genetic makeup and health profile [4, 5]. It recognizes the individuality of each person, customizing therapies accordingly. Currently, the most accurate drug information is found on labels and in the Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines, which explain how genomic data can inform decisions on dosages, metabolism, and potential adverse reactions to certain drugs [6, 7]. However, the FDA has detailed genomic information for only ∼380 drugs on their labels [8]. A significant portion of precision medication data remains buried in basic drug research databases and academic literature [9].

Existing data resources like PharmGKB [10], focusing on pharmacology and pharmacogenomics, CTD [11], with its specialization in toxicological data, and DrugBank [12], offering comprehensive drug information, contain a wealth of drug-related knowledge. However, their usefulness is hindered by a lack of a unified knowledge representation framework and their scattered presence across various platforms. This fragmentation makes it difficult for researchers and clinicians to fully utilize these resources [13]. Other

Received 18 March 2024; Revised 22 August 2024; Accepted 27 August 2024

knowledge resources such as the Precision Medicine Knowledgebase (PreMedKB) [14] consolidate ∼500 000 structured precision medicine relationships. But a detailed analysis shows that many of these relationships are oversimplified, labeled merely as "associate" or "effect," and fail to reflect the nuanced information in the original text. This issue arises from the traditional reliance on "triples" in knowledge graphs as the basic unit of knowledge representation. Traditional triples—comprising a subject, predicate, and object—are limited in conveying complex biomedical information, especially in representing the conditions for the establishment of knowledge [15]. They fail to capture the detailed genomic context required for precision medicine. Therefore, there is an urgent need for methods that can more accurately represent precision medicine knowledge by integrating these genetic conditions into knowledge representation. Furthermore, strategies in integration, mining, and governance should be developed to generate knowledge resources that better meet the requirements of precision medicine. Such resources would resolve ambiguities in current knowledge bases and focus on precise and accurate knowledge representation and dissemination.

To tackle existing challenges in precision medicine knowledge representation, we have developed Condition-based Precision Medicine Knowledge Graph (CPMKG), a comprehensive and advanced knowledge graph based on conditional precision medicine data. Our achievements include the following:

(i) We introduced the "Hyper-Triple," a novel framework that redefines the core data units in knowledge graphs. This framework overcomes the limitations of traditional triples by incorporating specific conditions (genetic backgrounds), ensuring that certain relationships are valid only under specific circumstances. This approach enhances the accuracy and depth of our knowledge representation, distinguishing CPMKG from traditional knowledge graphs.
(ii) We developed a "knowledge pattern" approach for organizing data. This method summarizes events involving multiple entities into a model, serving as an abstraction of conditional domain knowledge derived from the literature or expert input. These patterns are represented using the "Hyper-Triple" framework, guiding the collection of domain-specific knowledge.
(iii) CPMKG defines four key knowledge patterns for precision medicine research and application. It integrates 307 614 pieces of knowledge, addressing the needs of personalized medicine and drug discovery. The graph presents explicit relationships and constraints, enhancing the precision of therapies.
(iv) The knowledge graph offers a drug-centric exploration landscape, merging insights from molecular and clinical research into a comprehensive reasoning map. This facilitates the discovery of valuable evidence, supports medication synergy, and incorporates pharmacogenomics for holistic drug recommendations.
(v) CPMKG employs a large language model (LLM) to improve understanding of the knowledge graph. It provides clear explanations for subgraphs generated through system reasoning, balancing structured information with language expression for better user comprehension.

Materials and methods

Data collection

To delve into precision medicine knowledge, CPMKG has aggregated and restructured 15 727 studies from nine key drug-related databases: PharmGKB [10], SIDER [16], CIViC [17], DrugBank [12], TTD [18], CTD [11], DCDB [19], DoCM [20], and PharmacotherapyDB (https://github.com/dhimmel/indications). We meticulously extracted and mined information on relationships between various entities, such as drug side effects, sensitivities, molecular mechanisms, and treatments.

Knowledge pattern and conditional knowledge representation

CPMKG knowledge pattern

Biomedical literature serves as a vital repository of knowledge, where authors craft sentences to define and delineate key concepts. However, not every sentence contains critical information. Effective distillation of pertinent knowledge enhances the precision of text mining tools, bolstering their ability to expand knowledge bases and clarify the application scope of knowledge graphs. Our research primarily examines pharmacogenomics and drug research papers. In clinical drug studies, the focus often lies on treatment and side effects, whereas drug discovery research prioritizes understanding drug sensitivity and mechanisms. Despite varying linguistic expressions, these knowledge sources share semantic and entity-level commonalities. We harness these similarities through knowledge patterns to capture essential information accurately.

In CPMKG, knowledge is organized into four distinct patterns: Pattern 1 links drug side effects to disease treatment, highlighting genetic variations that signal increased risk (e.g. the C allele of olanzapine and metabolic syndrome in schizophrenia, Fig. 1a) [21–23]. Pattern 2 focuses on drug sensitivity, showing how genetic variations impact treatment outcomes (e.g. reduced tacrolimus need in liver transplant patients with the CC genotype, Fig. 1b) [24–26]. Pattern 3 elucidates drug mechanisms by connecting drug usage to changes in gene expression influenced by genetic variations (e.g. reduced AKR1C3 enzyme activity with the A allele during daunorubicin treatment, Fig. 1c) [27]. Pattern 4 outlines drug indications, providing therapeutic insights for specific drugs or combination therapies (e.g. enzalutamide for prostate cancer, Fig. 1d) [28]. These knowledge patterns often use genetic variations as conditions, highlighting the need for a more nuanced representation.

Conditional knowledge representation framework and storage

The knowledge patterns we designed, including node-to-edge (e.g. conditional) and edge-to-edge (e.g. causal) connections, cannot be intrinsically represented by conventional knowledge graphs, which only support node-to-node triples. To address this limitation, we extend to a "hypergraph" that accommodates complex information and allows for these non-node-to-node connections. In our knowledge graph, tuples can denote relationships between both entities and relations. For a tuple (S, R, and O), where S, R, and O stand for subject, relation, and object, respectively, R acts as a predicate,

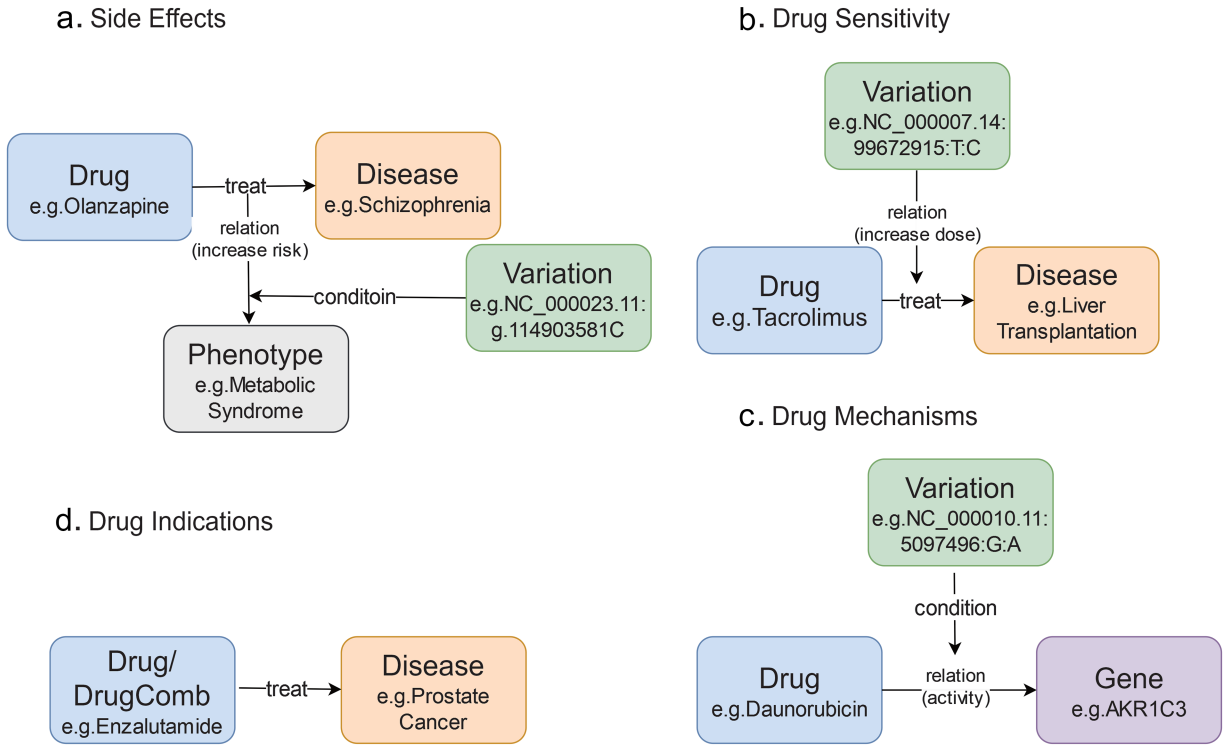


Figure 1. Knowledge patterns in CPMKG. (a) Side effects refer to side effects or complications that occur during the use of medication. Example: association between olanzapine and metabolic syndrome risk in schizophrenia patients. (b) Drug sensitivity refers to an individual's propensity to exhibit a heightened or exaggerated response to medication compared to the average population. Example: influence of the CC genotype on tacrolimus requirement in liver transplant patients. (c) Drug mechanism refers to the relationship between an individual's genome and their response to medications. Example: reduction of AKR1C3 enzyme activity during daunorubicin treatment in patients with the A allele variant. (d) Drug indication refers to the formal recommendation for medication use in treating specific diseases or pathological conditions. Example: utilization of enzalutamide in prostate cancer treatment.

while S and O can be entities or other tuples. This structure is defined as follows:

$$G = \langle V, E \rangle$$

$$E = E_{vv}, E_{ev}, E_{ee}$$

Here, E_{vv} represents edges that connect a vertex or an entity to another. E_{ev} denotes edges that link a vertex to an edge or vice versa. Finally, E_{ee} signifies edges that connect two edges. In a semantic context, E_{ev} is often utilized to denote a constraint condition for a tuple, while E_{ee} usually describes how one tuple (or event) leads to another. This framework allows for a more nuanced representation of complex relationships and conditions within the knowledge graph (Supplementary Fig. S2).

Our graph structure incorporates causes, conditions, and other crucial information, enabling detailed exploration of relationships. This intricate hypergraph comprises three fundamental structures: node-to-node, node-to-relation, and relation-to-relation connections, as shown in Supplementary Fig. S1a–c. Our knowledge representation framework also includes concept composition, representing combined medication as a collective "ALL" union of various drugs. We introduce "gate" nodes, inspired by logic gates, to amalgamate concepts and relations into new entities. We use two primary gates: the "AND gate," integrating all members, and the "OR gate," combining some members (see Supplementary Fig. S1d). Multiple nodes or edges directly connected without a gate node are considered independent. Additionally, our framework allows for the expression of negation and likelihood in all relations, making it highly expressive and adaptable for various scenarios.

Due to our updated knowledge representation, the classic knowledge graph storage method cannot accommodate our framework. To address this, we can adapt the data structure to better fit the storage capabilities of contemporary graph databases. Our approach involves integrating a helper node within a relationship, serving as a meaningful predicate. Specifically, we augment relationship predicates to function similarly to entities. We introduce a special relationship node that represents the original relationship predicate. This node uses "from" and "to" edges to indicate the direction of the relationship between the subject and object nodes. As a result, a triple's relationship is routed through this node, facilitating connections to entities or other triples. Furthermore, we can insert an auxiliary node within relationships to convey complex relationships more effectively, as demonstrated in Supplementary Fig. S2D. Additionally, this method enhances the flexibility and scalability of our knowledge graph, allowing for more intricate data representations and improving query performance.

Conditional knowledge mining

To ensure better alignment with our knowledge patterns, we refined our approach using several methods for knowledge mining and data integration.

We employed automated entity and relationship extraction techniques, utilizing regular expressions to parse unstructured text from databases such as PharmGKB [10], DrugBank [12], and CTD [11]. This approach allowed us to identify and extract entities and their relationships, converting raw text into structured data for our knowledge patterns. For a comprehensive description, refer to the Supplementary Methods for Processing Each Database.

From PharmGKB [10], we extracted 13 055 entries, and from CTD [11], we extracted 143 280 entries, which were then standardized and integrated.

To create a unified dataset for databases like SIDER [16], where data were scattered across multiple files, we merged and standardized various data tables. This process ensured consistency and completeness, resulting in a consolidated dataset of 89 491 entries.

For databases like TTD [18], which provided data in HTML format, we extracted relevant information by parsing the HTML content and reorganized it into a standardized dataset format, processing a total of 1002 entries.

In databases like PharmacotherapyDB (https://github.com/dhimmel/indications), where files contained duplicates, we performed deduplication by comparing data points and removing redundant entries. This process ensured each entry was unique and accurate, resulting in 11 751 unique entries post-deduplication.

For databases like DCDB [19], which listed multiple drugs in a single table, we isolated each drug entry and combined relevant data points to form a comprehensive dataset, processing 496 entries.

For databases like CIViC [17] and DoCM [20], which lacked structured relationships or complete data, we employed manual curation. Experts identified and added missing entities and relationships to ensure completeness and accuracy. This resulted in the curation of 1998 entries from CIViC [17] and 89 entries from DoCM [20], filling gaps and ensuring accurate representation.

Graph interpretation by LLM

To improve the user experience of CPMKG, we have incorporated the ChatGPT API, specifically the gpt-4o, into our web application. The integration of the AI-generated content model equips our web application with advanced natural language processing capabilities, significantly enhancing its interactivity and intelligence. We have designed prompts for four distinct application scenarios (Supplementary Table S3). The graphic descriptions generated in response to these prompts are tailored to user needs and supported by evidence from our knowledge graph. In this setup, the LLM refines and consolidates our knowledge, producing content that is both user-friendly and accurate. Importantly, we ensure that the model strictly adheres to the information in the knowledge graph, preventing the introduction of unsupported information and avoiding the risk of model hallucination.

Entity disambiguation

Entity disambiguation is a critical process for resolving ambiguities among entities sharing the same name. In CPMKG, this technique is applied to standardize five distinct types of entities: drugs, diseases, phenotypes, genetic variations, and genes. Each of these entities is associated with its own ontology, encompassing controlled vocabularies of standard names, synonyms, and IDs (the ontologies used for these entities are detailed in Supplementary Table S2).

For entities where a direct correlation exists between the source database entity ID and the target ontology ID, we employ ID mapping for straightforward standardization. For entities lacking this direct link, name mapping is utilized. Using controlled vocabulary $M = (m1, m2, \ldots, mn)$, we map the original structured names N of these entities against M to identify the most accurate terms. When this mapping results in a unique ID, it is adopted as the entity's external ID. In cases where the mapping leads to multiple "best matches," manual correction is undertaken.

$$\Gamma_{\text{best}} = \text{argmax}_{\Gamma} \sum_{i=0}^{n} \varphi(m_i, N)$$

Results

Statistics on entities and knowledge in CPMKG

CPMKG aims to advance precision medicine and drug discovery in clinical research. Utilizing a unified knowledge representation framework, CPMKG consolidates comprehensive pharmaceutical knowledge through processes like knowledge acquisition, element mining, and restructuring (Supplementary Fig. S3). This process has yielded 307 614 pieces of detailed drug knowledge, including 139 824 entries on side effects, 9819 on drug sensitivity, 144 269 on drug mechanisms, and 13 702 on drug indications. This comprehensive data encompasses 2150 drugs, 1689 diseases, 1719 phenotypes, 5029 genetic variations, and 20 111 genes (Table 1). For a more detailed statistical breakdown, including filtering and merging across various databases, refer to Supplementary Table S1.

In terms of entity disambiguation, the process resulted in the standardization of 30 698 entities, with 918 entities not aligned with external database mappings. Notably, there are 618 entities in CPMKG that can be classified as both diseases and phenotypes. These entities are represented in different knowledge patterns: diseases are associated with treatments, while phenotypes are linked to side effects. Users can choose the appropriate classification based on their specific use case.

Unlike traditional methods that integrate databases from diverse sources, CPMKG focuses on aligning data sources to predefined knowledge patterns. This methodology involves literature mining and manual curation based on original evidence within these patterns. This strategy not only gathers essential knowledge elements, such as drug interactions and genomic variations, but also enhances existing data by introducing new knowledge patterns.

Conditional knowledge-based schema of CPMKG

The conditional knowledge-based schema of CPMKG is constructed from four primary knowledge patterns: drug side effects, drug sensitivity, drug mechanisms, and drug indication. These patterns form the foundational elements of the schema, including entities such as drugs, diseases, phenotypes, genes, and variations. The relationships among these entities establish the schema's structure (Fig. 2). Differing from traditional knowledge graphs, CPMKG incorporates critical yet long-missing causal and conditional associations, embodied in relationships between entities and triples, as well as

Table 1. Statistics on entities and knowledge in CPMKG

Knowledge pattern	Knowledge source	Knowledge	Drug	Disease	Phenotype	Variant	Gene
Side effects	SIDER, PharmGKB, and DrugBank	139 824	1712	544	1719	1447	597
Drug sensitivity	CIViC, TTD, DrugBank, DoCM, and PharmGKB	9819	496	336	–	3478	880
Drug mechanisms	CIViC, CTD, and PharmGKB	144 269	1039	–	–	682	19 984
Drug indications	PharmacotherapyDB, DCDB, and SIDER	13 702	1394	1544	–	–	–
Total	–	307 614	2150	1689	1719	5029	20 111

En-dashes (–) : data is not available for this knowledge pattern.

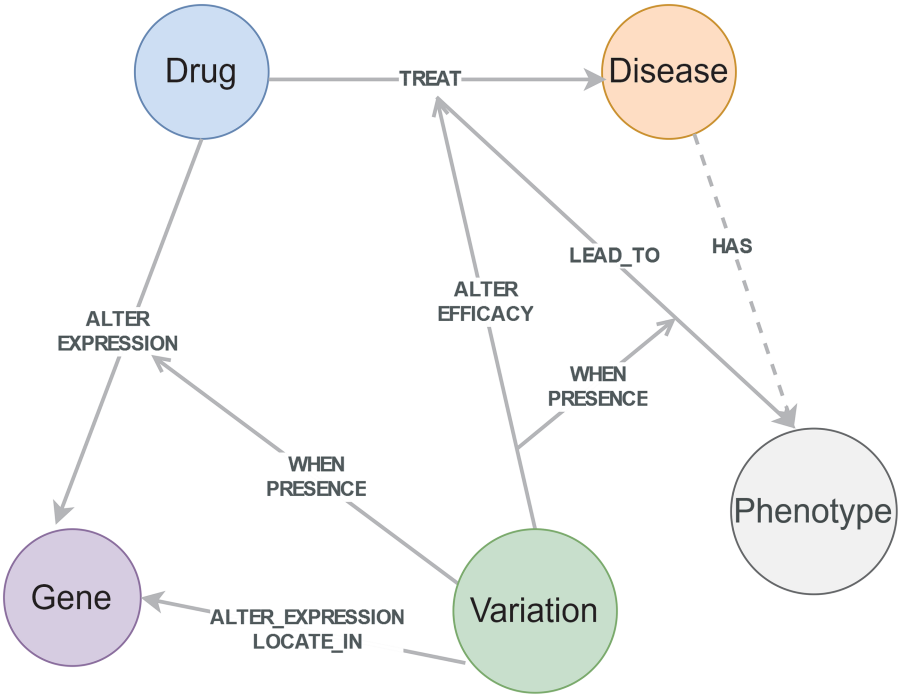


Figure 2. Conditional knowledge-based schema of CPMKG. This schema includes foundational elements such as drugs, diseases, phenotypes, genes, and variations. "Drugs" cover pharmacological substances, "diseases" encompass pathological conditions, "variations" refer to differences in the human genome, "phenotypes" include side effects or complications, and "genes" pertain to human genes. This schema illustrates the integration of these entities and their detailed relationships, highlighting the four conditional knowledge patterns in precision medicine.

among triples themselves. This allows for a nuanced representation of precision medicine knowledge, highlighting differences in individual genetic backgrounds and population characteristics.

Drug-centered conditional knowledge exploration

CPMKG empowers researchers to delve into drug-centered research in precision medicine. It offers users access to knowledge across four categories, including medication recommendations and pharmacogenomics, anchored in core elements like drugs and genetic variations. For example, Fig. 3a displays a knowledge list centered on "warfarin." It provides switchable lists covering four types of precision medicine knowledge, enabling users to deeply understand and compare warfarin's effects across different genetic backgrounds and explore personalized drug recommendations via entity-condition-relationship pairs.

Furthermore, CPMKG presents each knowledge unit graphically, allowing users to visualize the type of knowledge, the conditions under which it was established, and its evidence sources within the knowledge graph. This is accompanied by detailed knowledge descriptions and annotations for each entity. For instance, Fig. 3b demonstrates that under the genetic background *NC_000010.11:g.94981296A>C*, warfarin treatment for venous thromboembolism heightens bleeding risk [29]. This graphical representation provides an intuitive understanding of the knowledge, while the descriptions and entity annotations offer an in-depth comprehension of the graph.

Knowledge inference with multiple evidence

Each knowledge unit, representing a specific knowledge pattern, can effectively communicate the author's intended information. However, the scope of knowledge conveyed by a single piece of literature remains confined. CPMKG, as a knowledge graph, empowers researchers to integrate multiple knowledge instances into subgraphs, each drawing on numerous evidence sources. This allows for systematic reasoning about pertinent knowledge connections within these subgraphs.

Take, for example, the frequent occurrence of *KRAS mutations* (*KRAS* is a proto-oncogene that encodes a GTPase) in cancer, a factor in >20% of human cancers. These mutations are also present in patients with malignant pleural

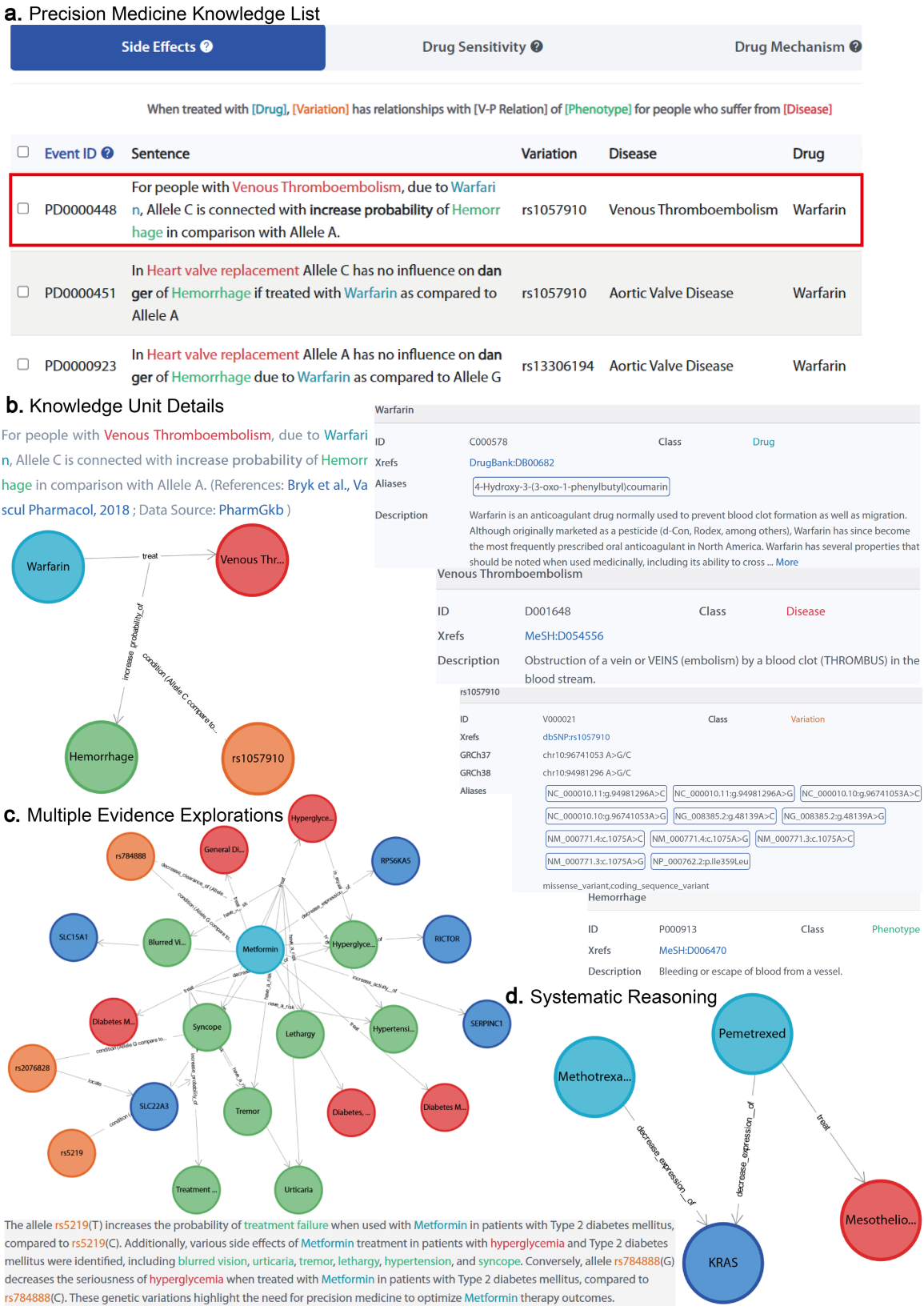
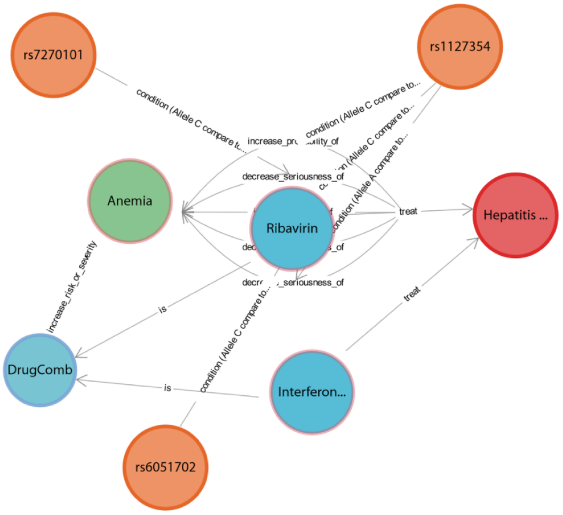


Figure 3. Knowledge exploration in CPMKG. (a) Precision medicine knowledge list. A list centered on "warfarin," comprising four distinct patterns. (b) Knowledge unit details. Illustrated by "warfarin treatment side effects," it includes graphical representation, established conditions, evidence sources, and entity details. (c) Multiple evidence explorations. Subgraph exploration centered on metformin, along with knowledge description. (d) Systematic reasoning. Illustration of pemetrexed's efficacy in MPM treatment and its correlation with reduced *KRAS* expression, suggesting a shared mechanism with methotrexate, supporting methotrexate's potential effectiveness in MPM treatment.

a. Drug Suggestion

In patients with Breast Neoplasms carrying rs1056836, different alleles (C or G) influence response to various drugs differently. Allele C generally associates with increased nausea and decreased response to several chemotherapy medications but has a decreased probability of peripheral neuropathy. The choice of medication should consider these pharmacogenomic interactions to minimize adverse effects and optimize therapeutic outcomes. Specific recommendations are made based on the presence of C or G alleles.

c. Medication Synergy Assistant



Based on the given statements, it's evident that the patient's genetic variants and current treatment regimen are crucial in determining the best approach to managing the ribavirin-induced anemia, especially in the context of chronic Hepatitis C treatment. Variants rs1127354 (Allele C), rs7270101 (Allele C), and rs6051702 (Allele C) have distinct influences on the seriousness and probability of anemia in patients undergoing Ribavirin therapy. The combination of Interferon alfa-2b and Ribavirin increases the risk or severity of anemia, which complicates the choice of medication. Each genetic variant's relation to anemia is considered in the decision-making process to suggest the most suitable drug regimen for reducing anemia risk in a patient with chronic Hepatitis C.

b. Pharmacogenetics

The medication Ivacaftor has varied effects on CFTR activity depending on the patient's genetic variants. Specifically, Ivacaftor increases CFTR activity in patients carrying rs113993960(CTTdel), rs397508288(G), rs150212784(G), rs121908752(G), and rs75541969(C). However, Ivacaftor does not influence CFTR activity in patients carrying rs397508276(G) or rs80055610(C). Additionally, rs113993960(CTTdel) is associated with increased CFTR transport when treated with Ivacaftor but does not affect protein thermostability.

Figure 4. Advanced application of CPMKG. (a) Personalized drug suggestion offers tailored medical advice based on diagnostic outcomes and genetic backgrounds. Example: crucial factors for prescribing medication to breast cancer patients with the *NC_000002.12:g.38071060G>A,C* variant. (b) Pharmacogenomics focuses on understanding drug mechanisms for personalized medicine and novel drug discovery. Example: ivacaftor's effects on various *CFTR* alleles and genotypes in CF. (c) Medication synergy assistant optimizes treatment outcomes and patient safety, particularly with multiple drugs. Example: effects of interferon α-2b and ribavirin in treating chronic hepatitis C.

mesothelioma (MPM). As shown in Fig. 3d, pemetrexed effectively treats MPM and reduces KRAS protein expression, a mechanism shared with methotrexate [30]. This shared mechanism led us to hypothesize methotrexate's effectiveness in treating MPM, a hypothesis supported by the literature [31].

To make these subgraphs more accessible, we have innovatively combined LLMs with precise, specialized medical knowledge from our database, including reference articles, to provide clear descriptions for each subgraph. We offer four distinct scenarios for varied graph descriptions, skillfully connecting structured graphs with narrative language. For instance, Fig. 3c displays a subgraph centered on "metformin," illustrating gene variations and their impact on diabetic patients' responses and side effects. The *rs784888(G)* allele correlates with a better response to metformin, reducing hyperglycemia severity compared to *rs784888(C)* [32], while the *rs5219(T)* allele is linked to an increased likelihood of treatment failure compared to *rs5219(C)* [33]. Common side effects for hyperglycemia patients taking metformin include blurred vision, urticaria, pruritus, skin rash, tremor, lethargy, hypertension, and syncope.

Advanced application of CPMKG (case study)

CPMKG enhances the drug usage experience with three user-centric applications: personalized drug suggestion, which offers tailored medical advice; pharmacogenomics application, accelerating drug mechanism research and discovering new applications for existing drugs; and medication synergy assistant, aiding in the selection of effective drugs or drug combinations.

Case study 1: personalized drug suggestion

Personalized drug suggestion aids clinical research by enabling individualized medical advice based on patients' diagnostic outcomes and genetic backgrounds. Consider Fig. 4a, which delineates crucial factors for prescribing medication to breast cancer patients carrying the *NC_000002.12:g.38071060G>A,C* genetic variant. For the GG genotype, docetaxel may be ineffective [34]. In contrast, the CC genotype can increase nausea risk with doxorubicin [35] and potentially lead to diminished efficacy with epirubicin [36]. However, this variant does not impact the effectiveness of gemcitabine and paclitaxel [37]. Importantly, cyclophosphamide has the potential to reduce peripheral neuropathy risk in patients with the C allele [38]. CPMKG provides valuable genotype-specific references to support prescription choices in clinical research.

Case study 2: pharmacogenomics application

Pharmacogenomics plays a crucial role in understanding drug mechanisms for personalized medicine and novel drug discovery, particularly in diseases like cystic fibrosis (CF). CF, caused by mutations in the CF transmembrane conductance regulator (*CFTR*) gene [39], can be treated with ivacaftor, a CFTR potentiator that enhances CFTR protein function [40]. Figure 4b demonstrates ivacaftor's effects on various *CFTR* alleles and genotypes, highlighting 11 genomic variations that significantly influence the drug's pharmacological response in the human body. For instance, ivacaftor treatment alters *CFTR* activity in the *NC_000007.14:g.117603654T>A,C* and *NC_000007.14:g.117611620A>C* variants. Additionally, the *NC_000007.14:g.117559592_117559594del* variant is linked with increased CFTR transport [41] but does not affect the protein's thermal stability [42]. Such insights are invaluable for developing targeted treatments for patients with CFTR-related conditions.

Case study 3: medication synergy assistant

In drug indication, both efficacy and side effects are of paramount importance. Medication synergy assistants can optimize treatment outcomes and bolster patient safety, particularly when multiple drugs are used, either in combination or individually. For example, Fig. 4c demonstrates the effects of interferon α-2b and ribavirin in treating chronic hepatitis C [43]. Both drugs, whether used separately or together, significantly increase the risk and severity of anemia. However, patients with the *NC_000020.11:g.3271278A>C* and *NC_000020.11:g.3213247A>C* genetic variants experience less severe anemia after ribavirin treatment [44, 45]. Consequently, ribavirin therapy is recommended for patients with these specific genetic profiles.

Discussion

Despite the advancements achieved with CPMKG, some detailed aspects still require deeper exploration. Precision medicine demands a thorough understanding of both entities and their attributes, such as gene variant genotypes and clinical indices. Transitioning from traditional triples to hyper-triples presents challenges in making accurate inferences due to the detailed and specific conditions involved. However, genetic information in drug databases is limited, and even the original literature often lacks necessary genomic details. As this area is under-researched, we aim to refine this in future studies and encourage broader contributions. This shift also underscores the need for further research into advanced knowledge reasoning methods. Our forward-looking approach leverages the sophisticated understanding capabilities of LLMs to decode complex semantics and hyper-triples. Additionally, we aim to utilize natural language interpretation based on intricate knowledge reasoning, driving the advancement of application-focused knowledge graphs.

Conclusion

CPMKG revolutionizes traditional drug knowledge by incorporating refined elements like specific conditions, making it ideal for precision medicine. Our knowledge graph offers personalized medication recommendations based on patients' genetic profiles, serving as a reference for clinical practice. It also supports researchers by facilitating drug metabolism studies and targeted drug discovery. Unique in its approach, CPMKG employs the "hyper-triple" concept in knowledge representation, capturing the complex nuances of precision medicine with remarkable accuracy. It merges and rationalizes various precision medicine knowledge pieces through innovative knowledge graph construction methods. This process not only uncovers information overlooked in current research but also enhances the understanding and application of these knowledge graphs in clinical research. Furthermore, the hypergraph structure can be seamlessly integrated into any graph database, accommodating existing database

technologies while ensuring minimal information loss compared to the original research publications. This effectively preserves the depth and complexity of the relationships, providing a robust and comprehensive foundation for future clinical-related research.

To make our knowledge graph more user-friendly, we have integrated LLM for graph interpretation. This integration not only advances our construction methods but also enriches the fusion of structured graphs with textual data. It broadens the spectrum of user engagement with knowledge graphs, paving the way for new perspectives in their representation, storage, and interpretation.

Acknowledgements

We would like to thank Dr Qingwei Xu from Ezhou Industrial Technology Research Institute, Huazhong University of Science and Technology for his support of website development.

Supplementary data

Supplementary data is available at *Database* online.

Conflict of interest

None declared.

Funding

This research was supported by the National Key Research and Development Program of China (Grant Nos 2021YFC2301502, 2021YFF0703702, 2016YFC0901904, and 2023YFA0915501); Key disciplines in the three-year Plan of Shanghai municipal public health system (2023–2025) (GWVI-11.1-42); Shanghai Science and Technology Innovation Action Plan (Grant No. 23JS1401500); Shanghai Municipal Science and Technology Major Project; and R&D Program of Guangzhou National Laboratory (Grant No. GZNL2024A01002).

Data availability

CPMKG is publicly accessible through https://www.biosino.org/cpmkg/. All data and resources hosted on the platform are freely accessible.

References

1. Jian J, He D, Gao S *et al*. Pharmacokinetics in pharmacometabolomics: towards personalized medication. *Pharmaceuticals (Basel)* 2023;**16**:1568. https://doi.org/10.3390/ph16111568
2. Hamburg MA, Collins FS. The path to personalized medicine. *N Engl J Med* 2010;**363**:301–04. https://doi.org/10.1056/NEJMp1006304
3. Dingemanse J, Appel-Dingemanse S. Integrated pharmacokinetics and pharmacodynamics in drug development. *Clin Pharmacokinet* 2007;**46**:713–37. https://doi.org/10.2165/00003088-200746090-00001
4. Schee Genannt Halfmann S, Evangelatos N, Schröder-Bäck P *et al*. European healthcare systems readiness to shift from 'one-size fits all' to personalized medicine. *Per Med* 2017;**14**:63–74.
5. Naithani N, Sinha S, Misra P *et al*. Precision medicine: concept and tools. *Med J Armed Forces India* 2021;**77**:249–57.
6. Caudle KE, Klein TE, Hoffman JM *et al*. Incorporation of pharmacogenomics into routine clinical practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process. *Curr Drug Metab* 2014;**15**:209–17.
7. Relling MV, Klein TE, Gammal RS *et al*. The Clinical Pharmacogenetics Implementation Consortium: 10 years later. *Clin Pharmacol Ther* 2020;**107**:171–75.
8. Kim JA, Ceccarelli R, Lu CY. Pharmacogenomic biomarkers in US FDA-approved drug labels (2000-2020). *J Pers Med* 2021;**11**:179.
9. Scott SA. Personalizing medicine with clinical pharmacogenetics. *Genet Med* 2011;**13**:987–95.
10. Barbarino JM, Whirl-Carrillo M, Altman RB *et al*. PharmGKB: a worldwide resource for pharmacogenomic information. *Wiley Interdiscip Rev Syst Biol Med* 2018;**10**:e1417.
11. Davis AP, Wiegers TC, Wiegers J *et al*. CTD tetramers: a new online tool that computationally links curated chemicals, genes, phenotypes, and diseases to inform molecular mechanisms for environmental health. *Toxicol Sci* 2023;**195**:155–68.
12. Wishart DS, Feunang YD, Guo AC *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–D1082.
13. Wilkinson MD, Dumontier M, Aalbersberg IJ *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 2016;**3**:160018.
14. Yu Y, Wang Y, Xia Z *et al*. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. *Nucleic Acids Res* 2019;**47**:D1090–101.
15. Anderson C, Müller H, Hanbury A *et al*. Formal ontologies in biomedical knowledge representation. *Yearb Med Inform* 2013;**22**:132–46.
16. Kuhn M, Letunic I, Jensen LJ *et al*. The SIDER database of drugs and side effects. *Nucleic Acids Res* 2016;**44**:D1075–79.
17. Griffith M, Spies NC, Krysiak K *et al*. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 2017;**49**:170–74.
18. Zhou Y, Zhang Y, Zhao D *et al*. TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res* 2024;**52**:D1465–77.
19. Liu Y, Wei Q, Yu G *et al*. DCDB 2.0: a major update of the drug combination database. *Database (Oxford)* 2014;**2014**:bau124.
20. Ainscough BJ, Griffith M, Coffman AC *et al*. DoCM: a database of curated mutations in cancer. *Nat Methods* 2016;**13**:806–07.
21. Mulder H, Franke B, van Der-beek van der AA *et al*. The association between HTR2C gene polymorphisms and the metabolic syndrome in patients with schizophrenia. *J Clin Psychopharmacol* 2007;**27**:338–43.
22. Risselada AJ, Vehof J, Bruggeman R *et al*. Association between HTR2C gene polymorphisms and the metabolic syndrome in patients using antipsychotics: a replication study. *Pharmacogenomics J* 2012;**12**:62–67.
23. Ma X, Maimaitirexiati T, Zhang R *et al*. HTR2C polymorphisms, olanzapine-induced weight gain and antipsychotic-induced metabolic syndrome in schizophrenia patients: a meta-analysis. *Int J Psychiatry Clin Pract* 2014;**18**:229–42.
24. Chen YK, Han LZ, Xue F *et al*. Personalized tacrolimus dose requirement by CYP3A5 but not ABCB1 or ACE genotyping in both recipient and donor after pediatric liver transplantation. *PLoS One* 2014;**9**:e109464.
25. Wei-lin W, Jing J, Shu-sen Z *et al*. Tacrolimus dose requirement in relation to donor and recipient ABCB1 and CYP3A5 gene polymorphisms in Chinese liver transplant patients. *Liver Transpl* 2006;**12**:775–80.
26. Monostory K, Tóth K, Kiss Á *et al*. Personalizing initial calcineurin inhibitor dosing by adjusting to donor CYP3A-status in liver transplant patients. *Br J Clin Pharmacol* 2015;**80**:1429–37.

27. Bains OS, Grigliatti TA, Reid RE *et al*. Naturally occurring variants of human aldo-keto reductases with reduced in vitro metabolism of daunorubicin and doxorubicin. *J Pharmacol Exp Ther* 2010;**335**:533–45.
28. Li C, Lanman NA, Kong Y *et al*. Inhibition of the erythropoietin-producing receptor EPHB4 antagonizes androgen receptor overexpression and reduces enzalutamide resistance. *J Biol Chem* 2020;**295**:5470–83.
29. Bryk AH, Wypasek E, Plens K *et al*. Bleeding predictors in patients following venous thromboembolism treated with vitamin K antagonists: association with increased number of single nucleotide polymorphisms. *Vascul Pharmacol* 2018;**106**:22–27.
30. Moran DM, Trusk PB, Pry K *et al*. KRAS mutation status is associated with enhanced dependency on folate metabolism pathways in non-small cell lung cancer cells. *Mol Cancer Ther* 2014;**13**:1611–24.
31. Kuribayashi K, Miyata S, Fukuoka K *et al*. Methotrexate and gemcitabine combination chemotherapy for the treatment of malignant pleural mesothelioma. *Mol Clin Oncol* 2013;**1**:639–42.
32. Goswami S, Yee SW, Stocker S *et al*. Genetic variants in transcription factors are associated with the pharmacokinetics and pharmacodynamics of metformin. *Clin Pharmacol Ther* 2014;**96**:370–79.
33. Sesti G, Laratta E, Cardellini M *et al*. The E23K variant of KCNJ11 encoding the pancreatic beta-cell adenosine 5′-triphosphate-sensitive potassium channel subunit Kir6.2 is associated with an increased risk of secondary failure to sulfonylurea in patients with type 2 diabetes. *J Clin Endocrinol Metab* 2006;**91**:2334–39.
34. Tulsyan S, Chaturvedi P, Singh AK *et al*. Assessment of clinical outcomes in breast cancer patients treated with taxanes: multi-analytical approach. *Gene* 2014;**543**:69–75.
35. Tecza K, Pamula-Pilat J, Lanuszewska J *et al*. Pharmacogenetics of toxicity of 5-fluorouracil, doxorubicin and cyclophosphamide chemotherapy in breast cancer patients. *Oncotarget* 2018;**9**:9114–36.
36. Le Morvan V, Litière S, Laroche-Clary A *et al*. Identification of SNPs associated with response of breast cancer patients to neoadjuvant chemotherapy in the EORTC-10994 randomized phase III trial. *Pharmacogenomics J* 2015;**15**:63–68.
37. Lee SY, Im SA, Park YH *et al*. Genetic polymorphisms of SLC28A3, SLC29A1 and RRM1 predict clinical outcome in patients with metastatic breast cancer receiving gemcitabine plus paclitaxel chemotherapy. *Eur J Cancer* 2014;**50**:698–705.
38. Abraham JE, Guo Q, Dorling L *et al*. Replication of genetic polymorphisms reported to be associated with taxane-related sensory neuropathy in patients with early breast cancer treated with Paclitaxel. *Clin Cancer Res* 2014;**20**:2466–75.
39. Riordan JR, Rommens JM, Kerem B *et al*. Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. *Science* 1989;**245**:1066–73.
40. Van Goor F, Yu H, Burton B *et al*. Effect of ivacaftor on CFTR forms with missense mutations associated with defects in protein processing or function. *J Cyst Fibros* 2014;**13**:29–36.
41. Keating D, Marigowda G, Burr L *et al*. VX-445-tezacaftor-ivacaftor in patients with cystic fibrosis and one or two Phe508del alleles. *N Engl J Med* 2018;**379**:1612–20.
42. Liu X, Dawson DC. Cystic fibrosis transmembrane conductance regulator (CFTR) potentiators protect G551D but not ΔF508 CFTR from thermal instability. *Biochemistry* 2014;**53**:5613–18.
43. Gane EJ, Stedman CA, Hyland RH *et al*. Efficacy of nucleotide polymerase inhibitor sofosbuvir plus the NS5A inhibitor ledipasvir or the NS5B non-nucleoside inhibitor GS-9669 against HCV genotype 1 infection. *Gastroenterology* 2014;**146**:736–743.e731.
44. Thompson AJ, Fellay J, Patel K *et al*. Variants in the ITPA gene protect against ribavirin-induced hemolytic anemia and decrease the need for ribavirin dose reduction. *Gastroenterology* 2010;**139**:1181–89.
45. Fellay J, Thompson AJ, Ge D *et al*. ITPA gene variants protect against anaemia in patients treated for chronic hepatitis C. *Nature* 2010;**464**:405–08.

Database, 2024, 00, baae102, DOI: https://doi.org/10.1093/database/baae102, Original article