Predicting drug outcome of population via clinical knowledge graph

Maria Brbić[1,*], Michihiro Yasunaga[2,*], Prabhat Agarwal[2,*], and Jure Leskovec[2,†]
[1] School of Computer and Communication Sciences, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland
[2] Department of Computer Science, Stanford University, Stanford, CA 94305, USA
*These authors contributed equally.
†Corresponding author. Email: jure@cs.stanford.edu

Optimal treatments depend on numerous factors such as drug chemical properties, disease biology, and patient characteristics to which the treatment is applied. To realize the promise of AI in healthcare, there is a need for designing systems that can capture patient heterogeneity and relevant biomedical knowledge. Here we present PlaNet, a geometric deep learning framework that reasons over population variability, disease biology, and drug chemistry by representing knowledge in the form of a massive clinical knowledge graph that can be enhanced by language models. Our framework is applicable to any sub-population, any drug as well drug combinations, any disease, and to a wide range of pharmacological tasks. We apply the PlaNet framework to reason about outcomes of clinical trials: PlaNet predicts drug efficacy and adverse events, even for experimental drugs and their combinations that have never been seen by the model. Furthermore, PlaNet can estimate the effect of changing population on the trial outcome with direct implications on patient stratification in clinical trials. PlaNet takes fundamental steps towards AI-guided clinical trials design, offering valuable guidance for realizing the vision of precision medicine using AI.

Introduction

A variety of different factors — environmental and biological at the molecular and cellular level — shape the treatment response. The same treatment may result in very different effectiveness and the likelihood of causing side effects when applied to different populations [1–6]. For example, the bias towards testing drugs on younger male Caucasian participants has led to missed patient-safety markers, raising awareness about the importance of population properties in investigating treatment efficacy and safety [7]. An overarching question is whether we can design more safe and effective treatments by changing the population properties to which the intervention is applied to [8].

Current approaches for predicting population response to treatment typically focus on a single disease and are designed for a specific task of interest [9–11]. On the other hand, general approaches for predicting outcome of a treatment that capture large space of underlying biological interactions, typically as networks [12–15], do not account for variability between patients. As such, these approaches fail to model population or individual response to a particular treatment and are unable to discover interventions effective only in certain groups. Finally, existing approaches are unable to reason about factors that cause certain side effects or effectiveness of interventions [16]. These approaches are typically black-box models that do not provide insights about causal relationships between interventions, population properties and the outcome.

Here, we present PlaNet, a geometric deep learning framework that predicts outcome of a treatment by reasoning over population variability, disease chemistry and drug biology. PlaNet is built over a massive clinical knowledge graph that captures treatment information in form of the *(drug, condition, population)* triplets grounded in biomedical knowledge that captures underlying chemical and biological interactions. PlaNet first learns general-purpose representations of all treatment, biological and clinical entities in the knowledge graph in an unsupervised fashion. This is achieved by pretraining the model to capture the structure of the network and semantics of the terms. PlaNet can then be fine-tuned on many downstream pharmacological tasks.

We demonstrate the utility of the PlaNet framework on clinical trials data. We structure the entire clinical trials database and incorporate it in PlaNet's framework, resulting in a knowledge graph of $330,915$ nodes and $13,928,443$ heterogenous edges where population variability is described by clinical trials' eligibility criteria. We use PlaNet to predict outcome of clinical trials including trial efficacy as survival endpoint, likelihood of causing side effects, and exact side effect category. By representing knowledge as a graph, PlaNet is equally applicable to drug combinations

as well as single treatments even for experimental drugs or their combinations that have never been seen in any clinical trial in the labeled data. Moreover, PlaNet captures causal relationships between population variability and treatment outcome, suggesting populations at risk of developing adverse events whose exclusion can impact the outcome of the trials and reduce the likelihood of side effects.

Results

Overview of PlaNet knowledge graph. PlaNet integrates the treatment information with an underlying biological and chemical knowledge. PlaNet consists of two knowledge graphs (KG): *(i)* a foreground clinical KG, and *(ii)* a background biological KG that captures relevant biology and chemistry. Clinical KG consists of a *(drug, condition, population)* triplets describing drug that is applied, condition or disease that the given population or patient has, and population/patient properties such as gender, age and medical history. Thus, *(drug, disease, population)* triplet defines a core triplet of the clinical KG that describes an application of a drug to a particular population or an individual. We then connect the foreground clinical KG with the background KG that captures underlying biology and chemistry. To create background biological KG, we integrate 9 biological and chemical databases to capture knowledge of disease biology and drug chemistry such as genomic variants associated with human diseases [17, 18], drug targets [19], physical interactions between human proteins [12], protein functions [20], chemical similarities between drugs [21], molecular, cellular and physiological phenotypes of chemicals [22] (Fig 1b; Supplementary Note 2). In total, PlaNet captures $5,751$ diseases, $14,300$ drugs augmented with $4,825$ drug structural classes, and $17,660$ proteins with $28,734$ protein functions.

To demonstrate the usage of PlaNet, we instantiate clinical KG on the clinical trials database[1] (Fig 1c). We structure the database and represent it in the form of treatment *(drug, condition, population)* triplets by extracting drug-disease-population information from free-text trial protocol description using various named entity recognition approaches (Supplementary Note 1). Drug corresponds to intervention whose effectiveness or safety is investigated in the trial, disease is a condition that is being studied in a trial, and population is defined by eligibility criteria. By structuring the clinical trials database, we avoid natural language bias and allow grounding the structured entities in a background biomedical KG of PlaNet (Fig 1d). Overall, the KG is built

[1] https://clinicaltrials.gov

over $69,595$ interventional clinical trials and $205,809$ trial arms. It comprises $13,928,443$ edges between $330,915$ nodes (Supplementary Tables 1-2).

Learning general-purpose embeddings using PlaNet. PlaNet learns general-purpose representations (embeddings) of all entities in the KG including clinical entities in the clinical foreground KG, as well as biological and chemical entities defined in the biomedical background KG. The encoder takes a KG as input and for each entity in the graph generates low-dimensional embeddings that preserve information about the graph topology, while capturing heterogeneity of the graph by learning relation-specific transformations that depend on the type of an edge considered. To learn general-purpose embeddings, we perform self-supervised learning by defining an auxiliary task as predicting the existence of an edge between two entities in the KG (Methods). This auxiliary task does not require any labels and enables PlaNet to learn meaningful embeddings from the prior knowledge data.

Pretraining step generates embeddings of every entity in the KG, in total $330,915$ entities. We visualize resulting trial arm entities in the two-dimensional UMAP space [23] (Fig. 2a). We find that trial arm nodes cluster based on disease groups and trial arms that investigate more similar diseases are embedded next to each other confirming that learnt embeddings are meaningful. For example, mental and nervous system diseases, and cardiovacular and nutritional/metabolic diseases are embedded close to each other. We demonstrate that these embeddings can be used for knowledge graph query answering over the structured clinical trials and biomedical knowledge databases (Supplementary Note 3). For example, one can ask PlaNet to generate all diseases associated with a protein that a particular drug targets, suggesting potential candidates for drug repurposing (Supplementary Fig. 1). By fine-tuning the PlaNet using task-specific annotations, PlaNet is applicable to a variety of downstream tasks. In particular, we next demonstrate PlaNet's ability to reason about efficacy and safety of clinical trials.

Predicting efficacy of clinical trials using PlaNet. We applied PlaNet to predict efficacy of drugs in the clinical trials database. We focused on predicting a survival endpoint as the most frequently used primary and secondary outcome. We parsed the survival information from the results section of the clinical trials and ensured that a higher value indicates more positive outcome, obtaining $1,307$ labeled trial arms across 625 trials. Given two arms of the same trial testing different drugs, we aimed at predicting which drug will result in more favorable prognosis (Fig. 2b). We represent trial arm as a set of study protocol embeddings including arm, drug, disease, primary outcome and

eligibility criteria embeddings and fine-tune PlaNet using survival information.

We compared PlaNet to drug-disease-outcome (DDO) model and transformer–based language model BERT pretrained on the PubMed abstracts and full PubMed Central articles [24, 25] and fine-tuned on clinical trials protocol text information (Supplementary Note 4). PlaNet achieves 0.70 area under receiver operating characteristic curve (AUROC), outperforming the PuBMedBERT model by 15% (Fig. 2c). For instance, PlaNet is the only model that correctly predicted higher overall survival of the atezolizumab group compared to docetaxel group in Phase II non-small-cell lung cancer trial [26] (Supplementary Fig. 2a), as well as the outcome of the recently initiated trial which showed that immunomodulatory agent lenalidomide can increase the activity of rituximab and leads to significantly higher progression-free-survival [27] (Supplementary Fig. 2b). To further boost PlaNet with a textual knowledge, we developed a joint knowledge- language model (PlaNetLM) that allows joint reasoning over text and KG, allowing the two modalities to interact with each other [28, 29] (Methods). We observed an additional 5% improvement in the performance in the fused language-KG PlaNetLM model (Fig. 2c). The substantial improvements of PlaNet models are not dependent on the evaluation metric (Supplementary Fig. 3-4).

Given that the number of training examples is limited to clinical trials that reported results [30, 31], we further tested whether a larger dataset could provide further boosts in the PlaNet's performance. We sampled without replacement our training set to artificially reduce its size and we found that with larger training set size PlaNet substantially improved performance (Fig. 2d). This shows that substantial performance improvements can be expected by increasing the training set size even by only a few hundred examples. While PlaNet is able to reason about drug effectiveness, we also investigated whether we can use PlaNet to search for candidate drugs that have a potential to be more effective than an FDA approved drug for a particular disease by creating artifical AI-generated clinical trials (Supplementary Note 6). We focused our question on capecitabine, an FDA approved treatment for metastatic breast cancer [32]. Among 7 top ranked drugs, all drugs have been investigated for breast cancers in isolation or combination with other drugs with a number of ongoing clinical trials, supporting immediate practical applicability of PlaNet in providing insights in potentially effective treatments.

PlaNet predicts outcome of novel drugs. We next questioned whether PlaNet can be applied to new drugs. This ability is crucial to be able to make predictions for experimental drugs that have never been investigated before. To test that, we train the model on 1040 drugs and then apply it to

a new set of 224 drugs that have never been applied in any clinical trial seen in the labeled data. Remarkably, we find that PlaNet achieves comparable performance on novel drugs compared to drugs abundantly present in the training set (Fig. 2e), demonstrating that PlaNet effectively generalizes to novel drugs, never-before-tested in the clinical trials. Such strong generalization ability is achieved by exploiting similarities between novel drugs and well investigated drugs through their connections in the KG.

When analyzing individual examples, we find that PlaNet predicted with high confidence lower survival of the novel investigational anticancer agent tasisulam-sodium compared to chemotherapy drug paclitaxel even though the model has never seen any labeled example that investigated tasisulam (Fig. 2f). In this phase III study conducted on metastatic melanoma patients, tasisulam resulted in 2.6 months lower overall survival and the trial was early terminated due to the possibly tasisulam-related deaths that were identified by the external data monitoring committee [33]. PlaNet is also applicable to drug combinations which is a highly non-trivial capability. For example, PlaNet correctly predicted improved progression-free survival (PFS) of combination of dabrafenib and trametinib compared to trametinib alone for melanoma patients without ever seeing any labeled example of trametinib or dabrafenib in the training set (Fig. 2g). Combination of these drugs was shown to be superior compared to monotherapy with 3-year PFS 22% with dabrafenib plus trametinib and 12% with trametinib alone [34] and it was later approved by FDA for melanoma patients with BRAF V600E or V600K mutations.

Predicting safety of clinical trials using PlaNet. We next applied PlaNet to reason about safety of clinical trials by extracting information about side effects of clinical trials from the results section. While previous works used machine learning models to predict adverse events of a drugs and drug combinations [35–38], these prior works neglect the effect of population to which the drug is applied on the occurrence of adverse events. Same drug applied to different populations may have caused different adverse events. To investigate dependence of adverse events on the change of population, we compared the adverse events frequency distributions between trials that apply the same drug to populations suffering from the same disease and trials in which disease is changed. We find that a high percentage of drug-disease combinations have significantly different adverse events frequency distributions when drug is applied to a different population (Supplementary Fig. 5).

We defined safety of a clinical trial with respect to a prior probability that a population suf-

fering from a particular condition will experience an adverse event without any intervention. We use placebo arm to estimate this prior probability and predict if the occurrence of a particular event is enriched in the intervention arm compared to the placebo arm when no intervention is given to the participants (Methods). We apply PlaNet to two safety prediction tasks: *(i)* predicting occurrence of a serious adverse event, and *(ii)* predicting exact adverse event category defined based on the preferred term in MedDRA hierarchy [39] (Fig. 3a). On the serious adverse event prediction task, PlaNet achieves a high AUROC score of 0.79 (Fig. 3b). Similar performance is observed on non-cancer clinical trials, confirming that the model is not biased to cancer trials that have higher probability of serious adverse events. We next evaluate whether PlaNet can predict the exact adverse event category. PlaNet achieves average 0.85 AUROC score across 554 adverse event categories, retaining high performance across different adverse event categories (Fig. 3c). Since many adverse events have a small number of positives, we additionally measure performance using AUPRC score as a function of the number of positives in the training set. For all bins, PlaNet consistently outperforms all baselines (Supplementary Fig. 6). We next assess the generalization ability of the model to predict safety of drugs and diseases that have never been seen during training. Similar to efficacy results, we again find that PlaNet effectively generalizes to novel drugs and diseases, achieving similar performance on novel drugs and diseases compared to drugs/diseases previously seen (Supplementary Fig. 7).

In the real-world setting, one would like to apply PlaNet to predict outcomes of new clinical trials by using historical data for training. To check how applicable is PlaNet in this setting, we use clinical trials data up to June 2017 for training, and then apply PlaNet to predict safety of newer trials that posted results after that date. We find that PlaNet achieves similar performance as when splitting the data by ensuring unique drug-disease pairs (Fig. 3d), demonstrating its applicability in the real-world setting in which the model needs to generalize to future trials. Interestingly, we find that PlaNet assigned very high confidence to pneumonia as an adverse event of everolimus given to patients with tuberous sclerosis complex with refractory partial-onset seizures in a phase III trial which is a very rare adverse event of everolimus [40] (Fig. 3e). However, we find that in this trial pneumonia was reported as a very common adverse event with one patient dying from pneumonia, which was even suspected to be treatment-related [41]. In a phase II trial that investigated lenvatinib safety for thyroid cancer patients, PlaNet correctly assigned highest confidence to uncontrolled hypertension as an adverse event (Fig. 3f). Hypertension was indeed later reported as the most frequent adverse event occurring in 80.5% patients [42]. PlaNet also correctly pre-

dicted with high confidence two other adverse events with the highest frequencies: fatigue (58.3%) and diarrhea (36.1%) (Supplementary Fig. 8a). Moreover, in three recent COVID-19 trials that investigated efficacy of remdesivir, PlaNet increased the probability of hemorrhage and breathing difficulty in all trials, which have been consistently reported in COVID-19 patients [43, 44] (Supplementary Fig. 8b). The model has never seen examples with COVID-19 or remdesivir drug during model training. In another COVID-19 trial completed in 2021 that investigated the protective role of proxalutamide in COVID-19 infection, PlaNet correctly increased the probability of gastrointestinal spasm as a side effect (Supplementary Fig. 8c), which was reported as the most common treatment emergent adverse event in this trial [45].

Causal reasoning with PlaNet. The fundamental question of trial design and precision medicine is whether we can change population or patient properties to lead to more favorable outcomes of treatments. To analyze the sensitivity of PlaNet to subtle changes of population terms, we identified all clinical trials that investigate the same drug, study the same disease and have the same primary outcome, but define different inclusion/exclusion criteria and result in a different adverse event (Fig. 4a). Given these matched trials, we aim at analyzing whether PlaNet correctly adjusts probability of an adverse event when the population is changed. We count pairs of matched trials as correct or wrong only if the difference between probability of adverse event occurrence is larger than the predefined threshold that we initially set to 0.2. We find that PlaNet correctly adjusted probability in 91% of matched pairs (6575 out of 7261), while wrong adjustments were observed in only 9% of pairs (Fig. 4b). With higher probability thresholds PlaNet achieves even higher differences between the correct and wrong predictions: with probability threshold of 0.3 PlaNet has 22 times more correct than wrong probability adjustments, while with 0.4 threshold PlaNet has 90 times more correct adjustments (Fig. 4c).

We next develop a methodology for assigning node importance scores to each term in the eligibility criteria (Methods). Given a population term, *i.e.*, inclusion/exclusion term in case of clinical trials, PlaNet computes the change in adverse event probability when the term is removed from the inclusion or exclusion criteria. High score indicates that removing a term from the criteria has a high influence on the occurrence of an adverse event. We then rank terms based on their influence on adverse event probability change (Fig. 4d). Using this methodology, we find that in a trial that investigated efficacy and safety of exemestane for breast neoplasms, PlaNet indicates that excluding terms 'metastasis', 'exemestane', 'tamoxifen' and 'aromatase inhibitors' leads to

the lower probability of breathing difficulty (Fig. 4e). We validate this finding by identifying another related trial that also studied exemestane for breast neoplasms but it does not have these terms in the exclusion criteria, being focused on metastatic breast neoplasms. Indeed, breathing difficulty is significantly enriched in a metastatic breast cancer trial compared to placebo and comparing PlaNet's predictions between these two trials PlaNet correctly adjusted probabilities and assigned 21.8% higher probability of breathing difficulty for the metastatic breast neoplasm trial. Additionally, external validation in literature and drug reports confirms that breathing difficulty is a known symptom of metastatic breast cancer [46] and a potential adverse event of tamoxifen and aromatase inhibitors including exemestane [47].

Discussion

PlaNet is a geometric deep learning framework for predicting treatment response of a population by reasoning over a massive clinical knowledge graph. The clinical knowledge graph in PlaNet captures population heterogeneity and prior knowledge of biological and chemical interactions. PlaNet learns low-dimensional embeddings of heterogeneous node types in an unsupervised manner and can use them on downstream pharmacological tasks of interest, such as predicting drug efficacy and likelihood of serious adverse events. If text data is additionally available, PlaNet can be further complemented with the language models [24, 48] and trained as a joint knowledge-language foundation model [29].

The unique ability of PlaNet is its ability to generalize to drugs, diseases and population terms that have never been part of the annotated datasets. By modelling clinical terms as nodes in the massive knowledge graph, PlaNet finds similarity of the novel terms to existing terms. This enables PlaNet to make predictions for experimental drugs, new emerging disease states, or population properties that have not been tested before. In three COVID-19 trials that investigated efficacy of remdesivir – disease and drug for which PlaNet has never seen any annotated example – PlaNet increased the probability of hemorrhage and breathing difficulty, side effects that have been consistently reported in COVID-19 patients [43, 44]. While previous works showed advantage in using network-based methods to identify clinically efficacious drug combinations [15], PlaNet extends this capability not only by considering population heterogeneity, but also by making predictions for combinations that include novel, experimental drugs.

PlaNet is scalable, flexible and easily extendable. Without retraining the model, PlaNet can be applied to new entities in the treatment knowledge graph such as new drugs, new diseases and new population terms. This important feature allows obtaining predictions for new drugs and population properties without retraining the model on these new terms.

PlaNet is uniquely able to reason about treatment effects over a complex population space and suggest how to change population to reduce the negative effects of the treatment. This opens opportunities to design more safe and effective treatments by intervening in the population design, but also to discover interventions effective only in certain groups. So far, such discovery has been happening rarely and often by chance [2].

Finally, PlaNet is a general framework: although we demonstrate its usage on clinical trials data, it could also be used to represent individual patients and integrated with existing clinical knowledge graphs [16]. In that case, the population properties would correspond to individ-

ual patient characteristics such as personal omics assays [49] paving the way towards precision medicine [50].

Methods

Knowledge graph construction. We develop a computational framework for systematically extracting structured information from the clinical trials database.[2]. We focus on interventional clinical trials that study at least one drug, resulting in $69,595$ trials. Given free-form text description of a clinical trial, our framework automatically extracts and structures clinical trials protocol information, including disease, drug/intervention, primary outcome and eligibility criteria. After extracting key terms, we standardize them by mapping the extracted terms to external databases. We provide details of the knowledge graph construction pipeline in Supplementary Note 1.

Model Overview. PlaNet knowledge graph is represented as a directed and labeled multi-graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{T})$ where $v_i \in \mathcal{V}$ are nodes/entities, $(v_i, r, v_j) \in \mathcal{E}$ are relations/labeled edges, $t_i \in \mathcal{T}$ are node types and $r \in \mathcal{R}$ denote relation types. Additionally, entities have associated entity attributes depending on the entity type (Supplementary Note 1). PlaNet learns a low-dimensional representation z_i for all the entities in the graph $\mathcal{G}$. The low-dimensional entity representations are learnt to capture both structural properties of an entity's neighborhood as well as entity's attribute representations.

Encoder. The encoder model takes node/entity in the PlaNet and maps it to a low-dimensional embedding vector that captures entity attributes and its local neighborhood. Formally, the encoder is a function $ENC : \mathcal{V} \rightarrow \mathbb{R}^d$ that takes entity $v_i \in \mathcal{V}$ and generates its low-dimensional embedding $z_i \in \mathbb{R}^d$ that captures entity structural properties as well as entity attributes. We build our encoder model as the relational graph neural networks (R-GCN) [51] encoder. Given a latent low-dimensional representation $h_i^{(l)}$ of entity v_i in the l-th layer of the neural network, single layer of the encoder has the following form:

$$h_i^{(l+1)} = \sigma\left(\sum_{r \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \frac{1}{c_{i,r}} W_r^{(l)} h_j^{(l)} + W_0^{(l)} h_i^{(l)}\right), \tag{1}$$

where $W_r^{(l)}$ is the transformation matrix for relation $r \in \mathcal{R}$, $\mathcal{N}_i^r$ denotes the set of neighbor indices of node i under relation $r \in \mathcal{R}$, $c_{i,r}$ denotes normalization constant defined as $c_{i,r} = |\mathcal{N}_i^r|$ and the operator σ defines the non-linear function in the neural network model. We use PReLU as an activation function. The key idea of the relational encoder is to learn propagation and transformation operators across different parts of the graph defined by the entity and relation types. Since

[2] https://clinicaltrials.gov

the transformation matrix depends on the relation type, the encoder propagates latent node feature information across edges of the graph while taking into account the type of an edge. In this way local neighborhoods are accumulated differently depending on the entity type. Thus, for each entity in the graph the encoder has a different neural network architecture defined by the network neighborhood of the given entity.

In the first layer, $h_i^{(0)}$ is initialized with entity attributes. The entity feature vectors are associated with different entity types, so we first learn a linear projection W_{t_i} for each entity type $t_i \in \mathcal{T}$ and use the projected attributes as the input to the first layer of the network:

$$h_i^{(0)} = W_{t_i} x_i \tag{2}$$

where x_i is an entity of type t_i. In other layers, the output of the previous layer becomes the input to the next layer representing latent low-dimensional entity representations that capture neighborhood structure. Stacking multiple layers allows successive application of propagation/transformation operators, giving the ability to the model to capture higher-order network neighborhoods. Final representation of entity v_i in the last (L-th) layer of the encoder gives us entity embeddings $z_i \in \mathbb{R}^d$, that is $ENC(v_i) = z_i = h_i^{(L)}$.

To efficiently handle rapid growth in the number of parameters with the number of relations in the graph, we use the basis decomposition regularization technique [51] and represent transformation matrix as a linear combination of basis transformations:

$$W_r^{(l)} = \sum_{b=1}^{B} a_{rb}^{(l)} V_b^{(l)}, \tag{3}$$

where $V_b^{(l)} \in \mathbb{R}^{d^{(l+1)} \times d^{(l)}}$ define basis and $a_{rb}^{(l)}$ are coefficients that depend on relation *r*.

Self-supervised learning. To leverage a large amount of unlabeled data, we first perform self-supervised learning by defining an auxiliary task. We define the auxiliary task as the edge mask/link prediction task. In particular, for each triplet (h, r, t) consisting of head, relation and tail entities, we first construct a k-hop subgraph of the head and tail entities. Then, we randomly drop α edges in the subgraph and the model is asked to reconstruct the dropped edges by assigning scores $f(h, r, t)$ to possible edges (h, r, t) in order to determine how likely those edges belong to $\mathcal{E}$. Our model for the task is a graph auto-encoder model, consisting of an entity encoder and an edge scoring function as the decoder. The encoder maps each entity $v_i \in \mathcal{V}$ to a real-valued vector $z_i \in \mathbb{R}^d$. The decoder assigns scores to (h, r, t)-triplets through a scoring function $f : \mathbb{R}^d \times \mathcal{R} \times \mathbb{R}^d \to \mathbb{R}$

denoting the probability of the triplet belonging to the graph. To define the scoring function for the triplets, we use DistMult factorization decoder [52]:

$$f(h, r, t) = z_h^T R_r z_t \,. \tag{4}$$

where every relation r is associated with a diagonal matrix $R_r \in \mathbb{R}^{d \times d}$, while z_h and z_t denote head and tail embeddings, respectively. We train the model with negative sampling [51, 52] meaning that for each observed example we sample n negative edges by randomly corrupting either the head or the tail of each positive triplet but not both. We use negative sampling loss with self-adversarial negative sampling [53] as defined below:

$$L = -\log \sigma(f(h, r, t)) - \sum_{i=1}^{n} p(h'_i, r, t'_i) \log \sigma(-f(h'_i, r, t'_i)))\,, \tag{5}$$

with

$$p(h'_j, r, t'_j | \{(h_i, r, t_i)\}) = \frac{e^{\alpha f(h'_j, r, t'_j)}}{\sum_i e^{\alpha f(h'_i, r, t'_i)}}\,, \tag{6}$$

where α is the sampling temperature, σ is the sigmoid function, and (h'_i, r, t'_i) is the i-th corrupted triplet for the positive triplet (h_i, r, t_i).

Outcome prediction. To fine-tune PlaNet on downstream prediction tasks, we represent a trial arm as the set of entities defining the trial protocol information including trial arm, diseases, drugs, primary outcomes, included and excluded population. To obtain trial arm embedding, we first obtain a representation vector for each entity type of trial protocol entity by computing the average embedding of all entities of a given type. Resulting embeddings represent protocol embeddings, *i.e.*, drug embedding, disease embedding, included/excluded population embeddings and primary outcome embedding. Finally, we concatenate all entity embeddings including arm embedding to obtain final trial representation h_T. Formally, the final trial arm embedding h_T is computed by aggregating information from all protocol entities using a parameter free convolution layer:

$$h_T = \left(\|_{r \in \mathcal{R}_\mathcal{T}} \frac{1}{|\mathcal{N}_i^r|} \sum_{j \in \mathcal{N}_i^r} h_j^{(L)} \right) \| \, h_T^{(L)} \tag{7}$$

where R_T denotes relations of a trial arm and $h_T^{(L)}$ is trial arm representation in the last layer L.

Trial outcome classifier takes as input final trial arm embedding and predicts the outcomes of the clinical trials, namely efficacy, safety and exact adverse events category. For efficacy prediction, outcome classifier takes as input pair of trial arm embeddings, while for safety and efficacy

tasks classifier takes as input single trial arm embedding. Task-specific classifier consists of two fully connected layers and outputs the probability that a particular event occurs. Specifically, the trial encoder is followed by a fully connected layer with non-linear ReLU activation function. Given trial arm embedding h_T, the forward-pass update of the first fully connected classifier layer is the following:

$$h'_T = ReLU(W_{T'} h_T + b_{T'}) \,, \tag{8}$$

where $W_{T'}$ is a parameter matrix and $b_{T'}$ is a bias vector. Finally, the model outputs probabilities in the second layer:

$$p = \sigma(W_t h'_T + b_t), \tag{9}$$

where W_t is the task specific weight matrix, b_t is the task specific scalar bias, and σ is the logistic sigmoid function.

Efficacy prediction. In the efficacy task, we predict which arm will have more favorable outcomes. We consider only survival-related primary and secondary outcomes including overall survival, progression-free survival, recurrence-free survival and disease-free survival. Depending on the unit, higher value may indicate better or worse outcome and we correct all examples with the opposite direction. The output of the model represents the probability that the first arm will have higher survival than the second arm. Specifically, given a pair of arms, we concatenate their trial arm embeddings computed from Equation (7), and then apply Equations (8) and (9) for prediction. We use the binary cross-entropy loss for training.

Safety and adverse event prediction. In the safety prediction task, the output corresponds to the probability of occurrence of serious adverse events, while in the adverse event prediction task the output corresponds to the probability of the occurrence of a particular adverse event category. We define both tasks with respect to the placebo arm. The placebo arm represents the prior probability that the adverse event will occur given the disease and population that the clinical trial is investigating. For each disease, we aggregate information from all tested placebo arms and use it as the estimation of the expected safety issues/adverse events. Given an intervention, we then construct a contingency table of frequency distributions between treatment and the estimated placebo arm and check whether the enrichment of adverse events is higher in the treatment arm than in the placebo arm at the particular odds ratio threshold. We use the odds ratio 2 as the default threshold. Importantly, the frequency between true placebo arms and estimated placebo arms is not significantly different between true and estimated placebo arms in 99.4% trials (t-test, , FDR $< 10\%$),

confirming that our estimates are trustworthy.

For predicting adverse events we consider MedDRA Primary Term (PT) level terms with at least 50 positive examples and at least 15 positive examples in the test set. In the adverse events prediction task, many categories are scarcely labeled. To transfer useful information from abundantly labeled categories to scarcely labeled categories, we train our model in the multi-task setting, In particular, our loss function is a multi-task binary cross entropy loss:

$$\mathcal{L}_{AE} = -\sum_{c \in \mathcal{C}} \frac{1}{N_c} \sum_{j=1}^{N_c} y_{jc} \log p_{jc} + (1 - y_{jc}) \log(1 - p_{jc}) \,, \tag{10}$$

where $\mathcal{C}$ is the set of adverse event categories, N_c is the number of learning examples for category task c, y denotes outcome binary labels and p denotes probability at the output of the model defined in Equation (9). Encoder is shared across all tasks, while each task has its own task-specific classifier. In particular, classifier parameters in Equation (8) are shared across all tasks, while parameters in Equation (9) are task-specific. For the safety prediction task, we use binary cross-entropy loss. We split the data into train, validation and test sets by ensuring that same trial and same drug-disease pairs can not appear in different splits, meaning that the model needs to generalize to unseen drug-disease combinations.

Knowledge graph-language model framework (PlaNetLM). The PlaNet model discussed above uses our constructed PlaNet knowledge graph as the primary information for efficacy/safety prediction. In addition, the raw text of clinical trial protocols could provide additional context (*e.g.*, description of the exact way dosage is given to participants), and improve robustness and safety of the model. With this motivation, we introduce a version of PlaNet model that incorporates the textual information (PlaNetLM), where we augment the R-GCN encoder with a text encoder, inspired by the DRAGON method [29, 54]. Specifically, letting $\texttt{text}_T$ denote the protocol text of the input trial arm T, we use a Transformer encoder [55] to obtain a text embedding of the arm, $g_T = \text{Transformer}(\texttt{text}_T)$. We then fuse the R-GCN embedding of the arm h_T and the text embedding of the arm g_T by concatenating them and passing them to an MLP. We use this architecture for both the pre-training and fine-tuning phases.

Neural network architecture. Our encoder consists of 2 message passing layers with 512 embedding size in each layer and basis decomposition with 15 bases. We use layer normalization, and PReLU [56] activation after the first layer of message passing. Additionally we use a Dropout [57] of 0.2 for the encoder after each layer. Other parameters are reported in Supplementary Note 5.

Causal reasoning. To provide explanations behind the predictions for the input trial arm, we develop a methodology for assigning node influence scores to each term in the eligibility criteria, inspired by [58]. Given a term, we compute the change in adverse event probability when the term is removed from the inclusion or exclusion criteria. Concretely, denoting the input trial arm node as T, the eligibility criterion term node as e, and the TrialNet KG as G, we prepare a KG *without* the edges between T and e: $G' = G \backslash \{(e, T)\}$. Then the influence score of the eligibility criterion e for the trial arm T in the adverse event category c is computed as:

$$S_c^{e \to T} := \Delta p_c = p(y_c; G') - p(y_c; G) \tag{11}$$

If the score is positive, it indicates that removing this eligibility criterion makes the adverse event probability higher, meaning that having this eligibility criterion reduces the adverse event probability.

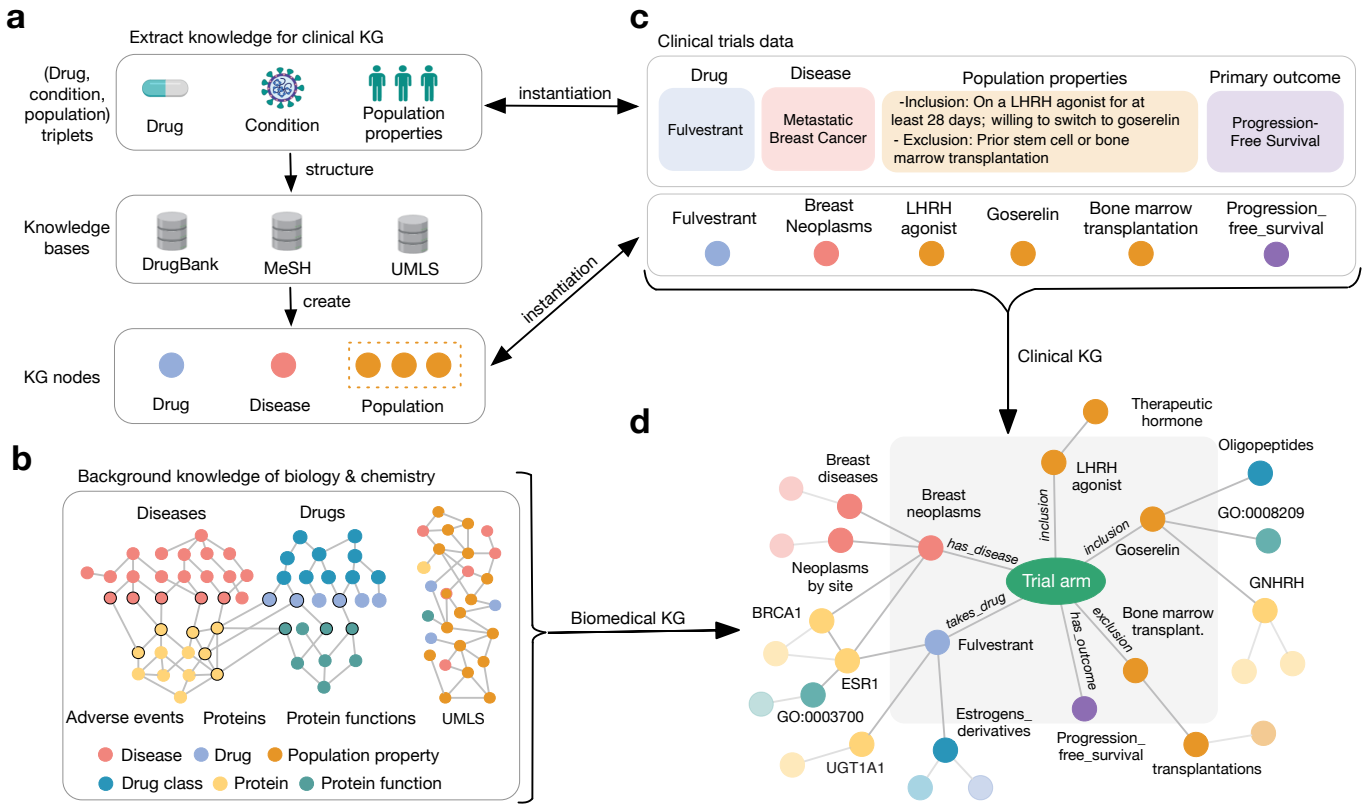

Figure 1: Overview of the PlaNet framework. PlaNet is built as a massive clinical knowledge graph (KG) that captures treatment information as well as underlying biology and chemistry. **(a)** The core of the PlaNet framework is a clinical KG that represents knowledge in the form of *(drug, disease, population)* triplets. These entities are then linked to external knowledge bases: diseases to Medical Subject Headings (MeSH) vocabulary [59], treatments to DrugBank database [19], and population properties to Unified Medical Language System (UMLS) terms [60]. **(b)** We integrate 11 biological and chemical databases to capture knowledge of disease biology and drug chemistry, such as databases of drug structural similarities, drug targets, disease-perturbed proteins, protein interactions and protein functional relations (Methods). These databases are integrated with the UMLS graph that captures population relations. **(c)** Instantiation of the PlaNet framework on the clinical trials data. We parse and standardize clinical trials database and extract information about diseases, drug treatments, eligibility criteria terms and primary outcomes. **(d)** Final KG is obtained by integrating the clinical KG (c) with biological and chemical networks (b).

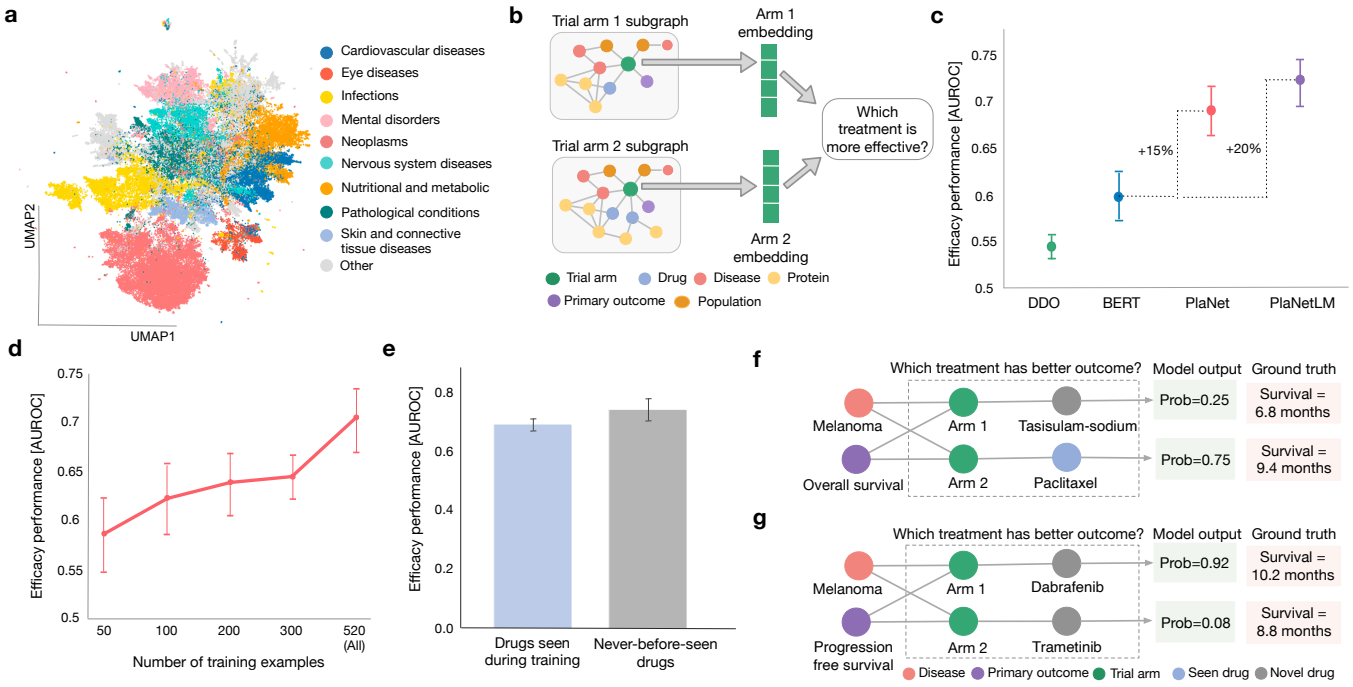

Figure 2: PlaNet reasons about efficacy of drugs in clinical trials even for experimental drugs that have never been tested before. (a) UMAP space of all trial arm embeddings in the clinical trials database obtained by pretraining PlaNet on the self-supervised task (Methods). Arms are colored according to disease information. Only major disease groups according to MeSH hierarchy [59] are shown. Grey color denotes minor disease groups. The arm embeddings learned by PlaNet exhibit clustering according to disease groups. **(b)** Given embeddings of two trial arms to which different drug treatments were applied, PlaNet predicts which of the treatments is more effective. Methodologically, the method geometric deep learning model is fine-tuned on the efficacy prediction task by using information about drug efficacy from the completed clinical trials. **(c)** Performance comparison of PlaNet with disease-drug-outcome (DDO) classifier and transformer-based language model BERT [24, 25]. PlaNetLM is obtained by augmenting PlaNet with the text embedding of the trial arm protocol [29] (Methods). Performance is measured as the mean area under receiver operating characteristic curve (AUROC) score across 10 runs of each model on different test data samples. Error bars are 95% bootstrap confidence intervals. **(d)** Effect of the training set size on the performance. With more training data, PlaNet substantially improves performance strongly indicating that further improvements can be expected by increasing the size of the training set. Performance is measured as the mean AUROC score across 10 runs on different test data samples. Error bars are 95% bootstrap confidence intervals. **(e)** PlaNet predicts efficacy of novel, experimental drugs that have never been seen in a clinical trial before. Bars represent the mean AUROC score for drugs that have been seen in the labeled training data (left; blue color), and never-before-seen drugs (right; grey color). Mean performance is computed across 10 runs of different test data samples and error bars are 95% bootstrap confidence intervals. **(f, g)** Examples of correct predictions. PlaNet outputs probabilities that a particular treatment will lead to higher overall survival of the population. **(f)** PlaNet correctly predicted higher overall survival of melanoma patients in paclitaxel arm compared to tasisulam-sodium arm. The model has never before seen any effect (labeled example) of the tasisulam-sodium drug. **(g)** PlaNet correctly predicted higher progression free survival of melanoma patients when given combination of dabrafenib and trametinib drugs compared to trametinib drug alone. The model has never before seen any effect of dabrafenib or trametinib drugs.

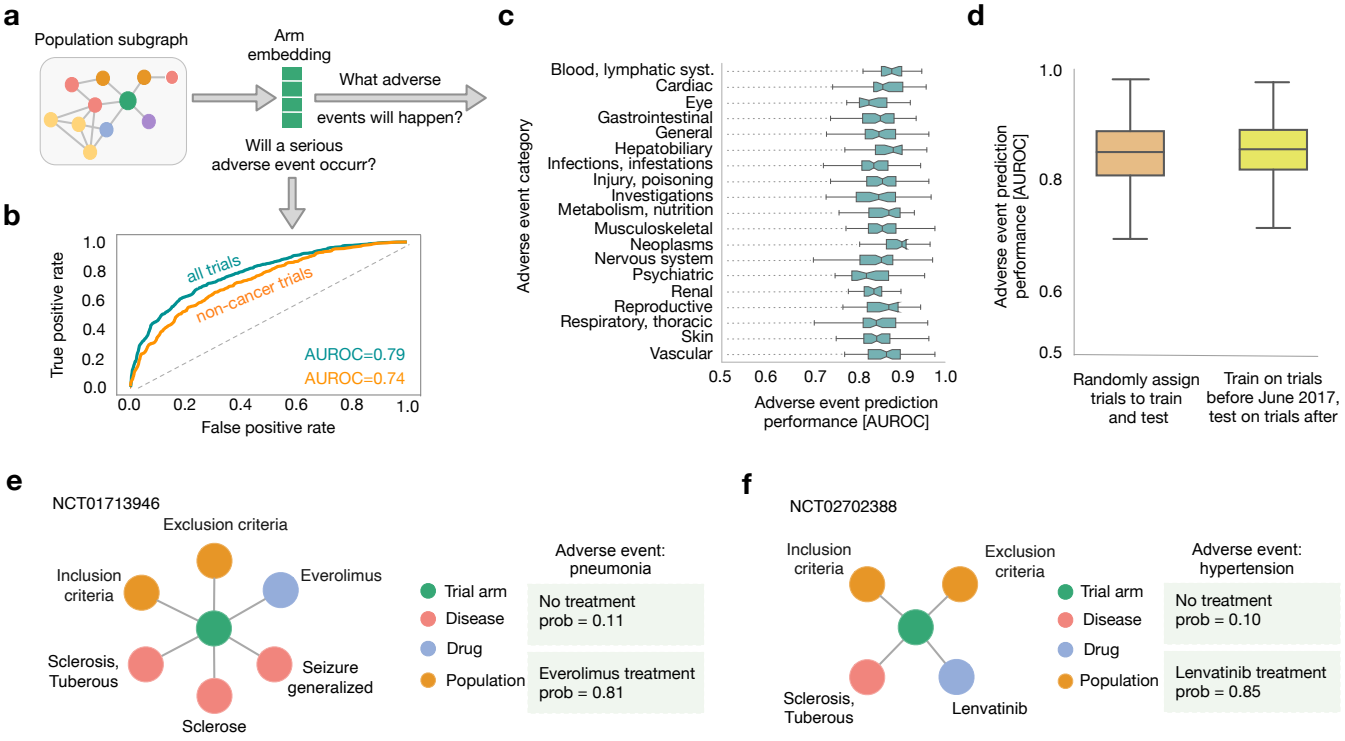

Figure 3: PlaNet reasons about safety of clinical trials. **(a)** Given a trial arm embedding, PlaNet predicts (b) whether a serious adverse event will occur and (c) what adverse event will happen. Methodologically, the methodolog geometric deep learning model is fine-tuned on the safety task by using information about drug safety from the completed clinical trials. **(b)** Performance of PlaNet on predicting occurrence of serious adverse events. PlaNet achieves AUROC score of 0.79 on predicting whether serious adverse event will occur. Green curve shows performance on all trials, while orange curve shows performance on on trials that do not investigate cancer diseases. **(c)** Performance of PlaNet on predicting exact category of adverse events measured as AUROC score. We consider 554 adverse events defined as preferred terms (PT) in MedDRA hierarchy [39] and group them according to the organ level categories. We consider organ level categories with at least 20 PT terms. The boxes show the quartiles of the performance distribution across different adverse events. Whiskers show the rest of the distribution. **(d)** Performance of PlaNet on predicting adverse events of future clinical trials. PlaNet achieves similar performance on predicting outcome of future clinical trials when compared to trials that are randomly split into train and test dataset independent of the year in which they were conducted. The performance is measured using AUROC and boxes show quartiles of the AUROC distribution across different adverse events. Whiskers show the rest of the distribution. **(e, f)** Examples of individual predictions of adverse events. Model assigns probability that an adverse event will be enriched in a given arm compared to no-treatment arm (Methods). **(e)** In an everolimus safety trial for tuberous sclerosis complex with refractory partial-onset seizures, PlaNet correctly predicted pneumonia as an adverse event with a high confidence. Although pneumonia is a very rare adverse event of everolimus [40], in this trial pneumonia was reported as a very common adverse event with one patient dying from pneumonia, which was suspected to be treatment-related [41]. **(f)** In a lenvatinib safety trial for thyroid cancer patients, PlaNet correctly predicted uncontrolled hypertension as an adverse event. Uncontrolled hypertension was reported as the most frequent adverse event in that trial [42].

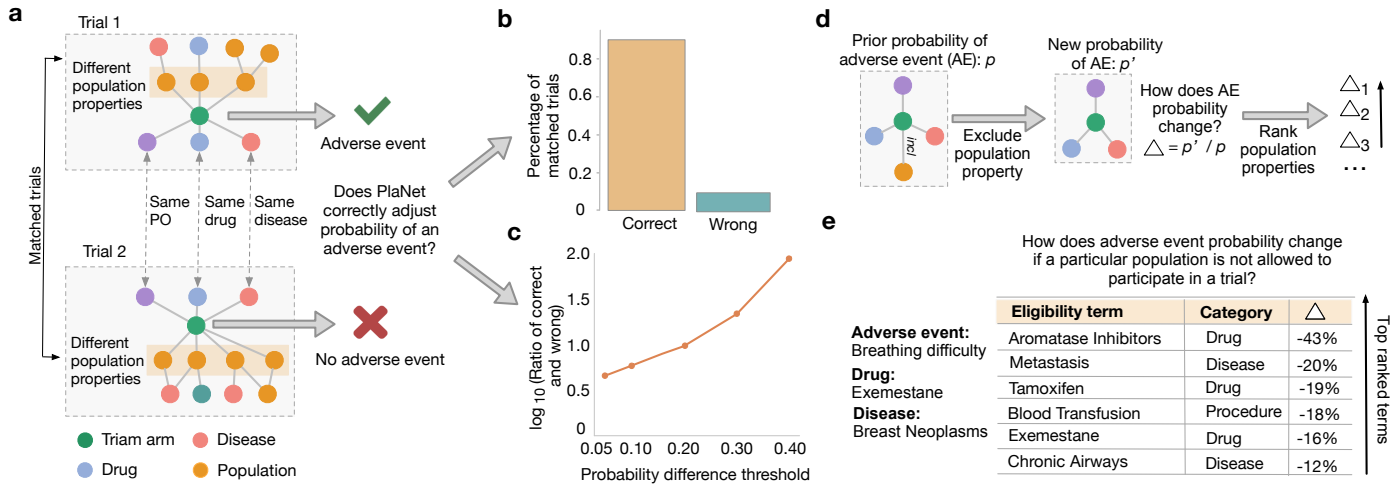

Figure 4: PlaNet identifies characteristics of populations that are at risk of developing adverse events. (a) We match clinical trials that study same drug, same disease and have same primary outcome (PO), but differ in the characteristics of the eligible population and result in different adverse events, *i.e.*, adverse event was observed in one trial, but not in the other. For pairs of such clinical trials, we assess whether model correctly adjusted prediction of an adverse event and predicted higher probability of an adverse event in one trial compared to the other. **(b)** Percentage of matched trials on which PlaNet correctly adjusted the probability of an adverse event (orange color; left) and percentage on which the adjustment was wrong (green color; right). PlaNet makes 10 times more correct adjustments than wrong. We count pairs only if the difference between probability of adverse event occurrence of two matched trials is at least 0.2. **(c)** The effect of the probability difference threshold on the ratio of correct and wrong probability adjustments. Even with smaller difference in probabilities (at least 0.05), the number of correct adjustments is more than 4 times higher than the number of wrong adjustments. With the difference of at least 0.4 the number of correct adjustments is 90 times higher than the number of wrong adjustments. For each probability threshold p, we count matched trials as correct or wrong only if the difference between probabilities is at least p. **(d)** PlaNet identifies population characteristics whose exclusion can reduce probability of adverse events. Given a population property, we estimate prior probability of an adverse event when population with a given property is included in the trial. We then change the trial by excluding population with that property, and observe the change in adverse event probability Δ. By ranking terms according to probability score, we can identify population properties whose exclusion can increase safety of clinical trials. **(e)** Use case of (d) for a trial that tests exemestane drug for breast neoplasms and in which breathing difficulty was observed as an adverse event. PlaNet finds population properties that have the highest effect on causing breathing difficulty. By excluding that population from the trial, PlaNet suggests that the probability of breathing difficulty can be significantly reduced. We rank terms that belong to drug, disease and procedure categories.

Data availability

We made all data including the clinical knowledge graph available at https://snap.stanford.edu/planet/data.zip.

Code availability

PlaNet was written in Python using the PyTorch library. The source code is available on Github at https://github.com/snap-stanford/planet.

Acknowledgements

We thank Camilo Ruiz and Michael Moor for their feedback on our manuscript and Marinka Zitnik for her feedback on the project. We gratefully acknowledge the support of DARPA under Nos. HR00112190039 (TAMI), N660011924033 (MCS); ARO under Nos. W911NF-16-1-0342 (MURI), W911NF-16-1-0171 (DURIP); NSF under Nos. OAC-1835598 (CINES), OAC-1934578 (HDR), CCF-1918940 (Expeditions), NIH under No. 3U54HG010426-04S1 (HuBMAP), Stanford Data Science Initiative, Wu Tsai Neurosciences Institute, Amazon, Docomo, GSK, Hitachi, Intel, JPMorgan Chase, Juniper Networks, KDDI, NEC, and Toshiba.

Author Contributions

M.B. and J.L. conceived the study. M.B., M.Y., P.A. and J.L. performed research, contributed new analytical tools, designed algorithmic framework, analyzed data and wrote the manuscript.

References

1. Ramamoorthy, A., Pacanowski, M., Bull, J. & Zhang, L. Racial/ethnic differences in drug disposition and response: review of recently approved drugs. *Clinical Pharmacology & Therapeutics* **97**, 263–273 (2015).

2. Schork, N. J. Personalized medicine: time for one-person trials. *Nature* **520**, 609–611 (2015).

3. Charles, H., Good, C. B., Hanusa, B. H., Chang, C.-C. H. & Whittle, J. Racial differences in adherence to cardiac medications. *Journal of the National Medical Association* **95**, 17 (2003).

4. Siegel, K., Karus, D. & Schrimshaw, E. Racial differences in attitudes toward protease inhibitors among older HIV-infected men. *AIDS care* **12**, 423–434 (2000).

5. Liu, K. A. & Dipietro Mager, N. A. Women's involvement in clinical trials: historical perspective and future implications. *Pharmacy Practice (Granada)* **14**, 0–0 (2016).

6. Franconi, F., Brunelleschi, S., Steardo, L. & Cuomo, V. Gender differences in drug responses. *Pharmacological Research* **55**, 81–95 (2007).

7. Knepper, T. C. & McLeod, H. L. When will clinical trials finally reflect diversity? *Nature* (2018).

8. Liu, R. *et al.* Evaluating eligibility criteria of oncology trials using real-world data and AI. *Nature* **592**, 629–633 (2021).

9. Jin, C. *et al.* Predicting treatment response from longitudinal images using multi-task deep learning. *Nature Communications* **12**, 1851 (2021).

10. Xu, Y. *et al.* Deep learning predicts lung cancer treatment response from serial medical imaginglongitudinal deep learning to track treatment response. *Clinical Cancer Research* **25**, 3266–3275 (2019).

11. Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences* **115**, E2970–E2979 (2018).

12. Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nature Communications* **12**, 1–15 (2021).

13. Cheng, F. *et al.* Network-based approach to prediction and population-based validation of in silico drug repurposing. *Nature Communications* **9**, 1–12 (2018).

14. Luo, Y. *et al.* A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature Communications* **8**, 1–13 (2017).

15. Cheng, F., Kovács, I. A. & Barabási, A.-L. Network-based prediction of drug combinations. *Nature Communications* **10**, 1–11 (2019).

16. Santos, A. *et al.* A knowledge graph to interpret clinical proteomics data. *Nature Biotechnology* **40**, 692–702 (2022).

17. Piñero, J. *et al.* Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* gkw943 (2016).

18. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* (2019).

19. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

20. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29 (2000).

21. Feunang, Y. D. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **8**, 1–20 (2016).

22. Davis, A. P. *et al.* Chemical-induced phenotypes at CTD help inform the predisease state and construct adverse outcome pathways. *Toxicological Sciences* **165**, 145–156 (2018).

23. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software* **3**, 861 (2018).

24. Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics (NAACL)* (2019).

25. Gu, Y. *et al.* Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare* **3**, 1–23 (2021).

26. Chalabi, M. *et al.* Efficacy of chemotherapy and atezolizumab in patients with non-small-cell lung cancer receiving antibiotics and proton pump inhibitors: pooled post hoc analyses of the oak and poplar trials. *Annals of Oncology* **31**, 525–531 (2020).

27. Leonard, J. P. *et al.* Augment: a phase iii study of lenalidomide plus rituximab versus placebo plus rituximab in relapsed or refractory indolent lymphoma. *Journal of Clinical Oncology* **37**, 1188 (2019).

28. Yasunaga, M., Ren, H., Bosselut, A., Liang, P. & Leskovec, J. Qa-gnn: Reasoning with language models and knowledge graphs for question answering. In *North American Chapter of the Association for Computational Linguistics (NAACL)* (2021).

29. Yasunaga, M. *et al.* Deep bidirectional language-knowledge graph pretraining. In *Advances in Neural Information Processing Systems* (2022).

30. Prayle, A. P., Hurley, M. N. & Smyth, A. R. Compliance with mandatory reporting of clinical trial results on clinicaltrials. gov: cross sectional study. *BMJ* **344** (2012).

31. Ross, J. S., Mulvey, G. K., Hines, E. M., Nissen, S. E. & Krumholz, H. M. Trial publication after registration in clinicaltrials. gov: a cross-sectional analysis. *PLoS Medicine* **6**, e1000144 (2009).

32. Ershler, W. B. Capecitabine monotherapy: safe and effective treatment for metastatic breast cancer. *The Oncologist* **11**, 325–335 (2006).

33. Hamid, O. *et al.* A randomized, open-label clinical trial of tasisulam sodium versus paclitaxel as second-line treatment in patients with metastatic melanoma. *Cancer* **120**, 2016–2024 (2014).

34. Long, G. *et al.* Dabrafenib plus trametinib versus dabrafenib monotherapy in patients with metastatic BRAF V600E/K-mutant melanoma: long-term survival and safety analysis of a phase 3 study. *Annals of Oncology* **28**, 1631–1639 (2017).

35. Atias, N. & Sharan, R. An algorithmic framework for predicting side effects of drugs. *Journal of Computational Biology* **18** (2011).

36. Liu, M. *et al.* Large-scale prediction of adverse drug reactions using chemical, biological, and phenotypic properties of drugs. *Journal of the American Medical Informatics Association* **19**, e28–e35 (2012).

37. Zitnik, M., Agrawal, M. & Leskovec, J. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* **34**, i457–i466 (2018).

38. Galeano, D., Li, S., Gerstein, M. & Paccanaro, A. Predicting the frequencies of drug side effects. *Nature Communications* **11**, 1–14 (2020).

39. Brown, E. G., Wood, L. & Wood, S. The medical dictionary for regulatory activities (MedDRA). *Drug Safety* **20**, 109–117 (1999).

40. Saito, Y. *et al.* A case of pneumocystis pneumonia associated with everolimus therapy for renal cell carcinoma. *Japanese Journal of Clinical Oncology* **43**, 559–562 (2013).

41. Curatolo, P. *et al.* Adjunctive everolimus for children and adolescents with treatment-refractory seizures associated with tuberous sclerosis complex: post-hoc analysis of the phase 3 exist-3 trial. *The Lancet Child & Adolescent Health* **2**, 495–504 (2018).

42. Giani, C. *et al.* Safety and quality-of-life data from an Italian expanded access program of lenvatinib for treatment of thyroid cancer. *Thyroid* **31**, 224–232 (2021).

43. Sudre, C. H. *et al.* Attributes and predictors of long COVID. *Nature medicine* **27**, 626–631 (2021).

44. Patell, R. *et al.* Postdischarge thrombosis and hemorrhage in patients with COVID-19. *Blood* **136**, 1342–1346 (2020).

45. McCoy, J. *et al.* Proxalutamide reduces the rate of hospitalization for COVID-19 male outpatients: A randomized double-blinded placebo-controlled trial. *Frontiers in Medicine* 1043 (2021).

46. Geels, P., Eisenhauer, E., Bezjak, A., Zee, B. & Day, A. Palliative effect of chemotherapy: objective tumor response is associated with symptom improvement in patients with metastatic breast cancer. *Journal of Clinical Oncology* **18**, 2395–2405 (2000).

47. Peters, A. & Tadi, P. Aromatase inhibitors. *StatPearls [Internet]* (2021).

48. Lee, J. *et al.* BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240 (2020).

49. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and medical phenotypes. *Cell* **148**, 1293–1307 (2012).

50. Ashley, E. A. Towards precision medicine. *Nature Reviews Genetics* **17**, 507–522 (2016).

51. Schlichtkrull, M. *et al.* Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, 593–607 (Springer, 2018).

52. Yang, B., Yih, W., He, X., Gao, J. & Deng, L. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations* (2015).

53. Sun, Z., Deng, Z., Nie, J. & Tang, J. RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations* (2019).

54. Yasunaga, M., Leskovec, J. & Liang, P. LinkBERT: Pretraining language models with document links. In *Association for Computational Linguistics (ACL)* (2022).

55. Vaswani, A. *et al.* Attention is all you need. *Advances in neural information processing systems* **30** (2017).

56. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE International Conference on Computer Vision*, 1026–1034 (2015).

57. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**, 1929–1958 (2014).

58. Ying, R., Bourgeois, D., You, J., Zitnik, M. & Leskovec, J. GNNExplainer: Generating explanations for graph neural networks. In *Advances in Neural Information Processing Systems* (2019).

59. Lipscomb, C. E. Medical subject headings (MeSH). *Bulletin of the Medical Library Association* **88**, 265 (2000).

60. Bodenreider, O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* **32**, D267–D270 (2004).