

Prioritization of drug repurposing hypotheses through network classification and electronic health record analysis

Anjali Sivanandan, Jennifer L. Wilson
University of California, Los Angeles

Background

Protein-protein interaction (PPI) networks are an increasingly popular method for predicting drug effects and can increase understanding of disease mechanisms^{1,2}, predict drug-drug interactions³, and identify drug repurposing opportunities^{4,5}. However, the translational impact of PPI networks is limited due to their tendency to over-predict drug effects. Lacking sufficient gold standards, the field considers most of these predictions to be false positives^{6,7}. The development of standardized tools for observational studies using the electronic health record (EHR) provides an opportunity to validate PPI network predictions. In our own work, we validated predicted drug effects using observational studies in the EHR⁸. The approach used PPI-predicted proteins downstream of druggable targets to generate network-based drug classes. Additionally, we were inspired by the class based comparisons conducted in the OHDSI LEGEND studies^{9,10} which motivated our comparison of network-based drug classes in the EHR. We hypothesize that PPI network classification coupled with large-scale observational studies could efficiently estimate drug true positive and negative effects. We also saw an opportunity to create a framework for large-scale assessment of drug repurposing hypotheses. We first piloted our study on drugs predicted to affect diabetes.

Methods

We used PathFX to predict drug effects for all drugs in DrugBank¹¹. Briefly, PathFX uses the available evidence supporting PPIs around drug targets to prioritize downstream proteins and uses statistical enrichment to predict pathway phenotypes, “effects”, for a drug’s target(s)⁶. We manually curated a list of 44 approved diabetes drugs and assessed their PathFX associations to 103 diabetes-related pathway phenotypes (“approved diabetes pathways” i.e. PathFX pathways of approved diabetes drugs). Approved drugs were connected through unique PPIs and we used hierarchical clustering to generate 3 pathway-based drug groups. We characterized each drug cluster using their pathway genes: (1) insulin cluster, (2) glucagon cluster and (3) apolipoprotein cluster. Gene Ontology (GO) enrichment of genes in each drug cluster was conducted using GOrilla¹² with all unique PathFX nodes as the background set.

We conducted an EHR study to assess the average treatment effect on A1C difference between drug clusters. Preliminary analysis found that the insulin cluster had insufficient patient counts for individual comparison so we combined the insulin and apolipoprotein clusters for EHR analysis as hierarchical clustering indicated that they were more similar to each other than the glucagon cluster. We used the University of California Health Data Warehouse (UCHDW), a de-identified health record, in the standard OMOP Common Data Model format. We do not have access to OHDSI tools and packages so a key challenge we address through this study is the implementation of observational studies on the UCHDW, drawing from OHDSI best practices. We used a standard treatment-comparator study design with 1:1 propensity score (PS) matching with no duplicates (Fig. 1a). We were motivated to understand

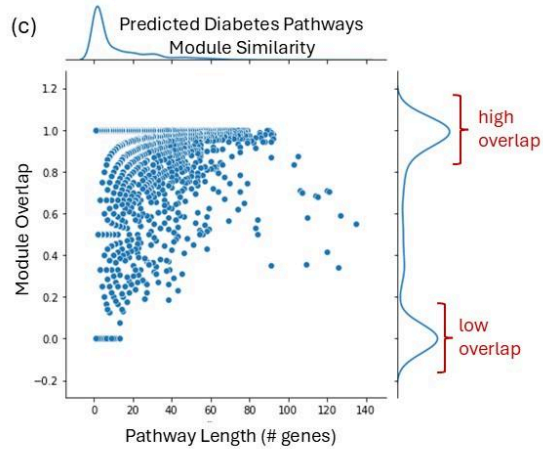
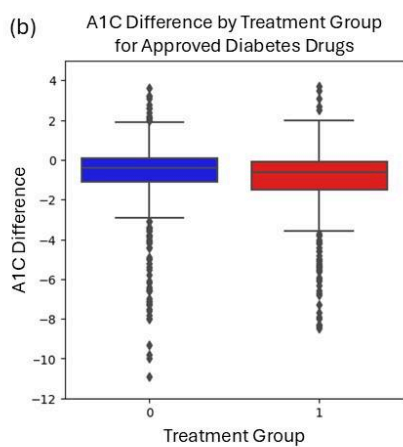
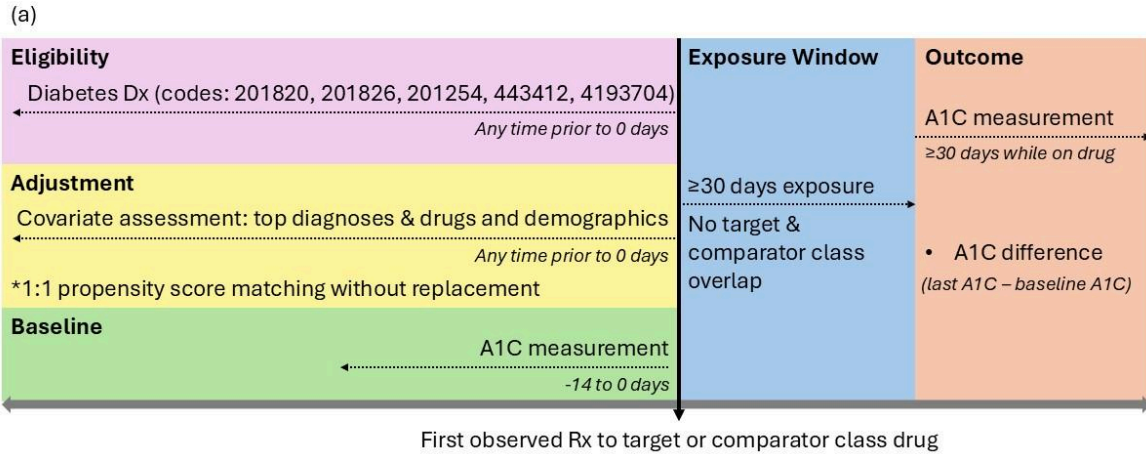
the extent of a patient's altered biology so we emphasized the patients' other diagnoses and drugs as covariates (top 50 non-study diagnoses, top 50 non-study drugs, and patient demographics of gender, race, ethnicity, and year of birth). We used logistic regression to predict treatment assignment and assigned PS based on the 104 covariates. A1C measurements are routinely used to monitor diabetes progression and were well captured in our dataset so we defined A1C difference as the outcome of interest (difference between an A1C measurement taken ≤ 2 weeks before the drug exposure and one after ≥ 30 days on the drug). This resulted in a final study group of 950 patients (475 target; 475 comparator).

We then looked at PathFX pathways connecting non-diabetes drugs to diabetes phenotypes ("predicted diabetes pathways"). We grouped predicted diabetes drugs based on similarity with approved drugs and derived two new classes: "high" and "low" overlap predicted drugs. We assessed incidence of a diabetes diagnosis after drug exposure to the high or low overlap drug classes and used a Fisher's exact test to assess enrichment of diagnoses relative to the entire EHR database.

Results

We generated PathFX networks for 40 of the anti-diabetics with sufficient data and discovered 216 pathway predictions associating 32/40 approved anti-diabetics to diabetes pathways. Hierarchical clustering of anti-diabetics based on pathway genes resulted in 3 drug clusters: (1) insulin cluster, (2) glucagon cluster and (3) apolipoprotein cluster. Despite having unique PPIs, all 3 drug clusters were enriched for GO terms related to the regulation of lipids, hormones and metabolic processes. We then measured clinical outcomes in diabetes for patients on different drug clusters in the EHR using A1C difference before and after drug exposure. This revealed no significant difference between patients on the glucagon versus insulin or apolipoprotein cluster drugs (Fig 1b), which suggested equivalence between drug clusters and motivated us to create a diabetes disease module consisting of all genes in approved diabetes pathways. We considered this 216 gene disease module to be our "known-effect" diabetes pathway genes.

PathFX predicted 10,205 predicted diabetes pathways connecting 2,492 non-diabetes drugs to diabetes phenotypes. We evaluated pathway similarity between predicted diabetes pathways and the diabetes disease module and reclassified drugs based on "high" or "low" similarity with approved drugs (Fig 1c). We hypothesized that similarity to approved diabetes pathways would indicate true or false connections to diabetes. Our EHR study revealed that 0 out of 21,027 patients on low similarity drugs had a diabetes diagnosis (odds ratio = 0.0, p-value = 0.0) compared to 506 of 21,027 diabetes diagnoses in patients on high similarity drugs (odds ratio = 0.449, p-value = 2.432e-91), suggesting that low similarity drugs have minimal associations to diabetes and are likely true negatives (i.e. unlikely candidates for drug repurposing). We then used these drug effect examples to prioritize 1,359 high similarity drugs from our initial 2,492 predicted diabetes drugs for further investigation for potential true connections to the disease (Fig 1d).



(d) Identification of TN drug effects helps prioritize predictions for future investigations

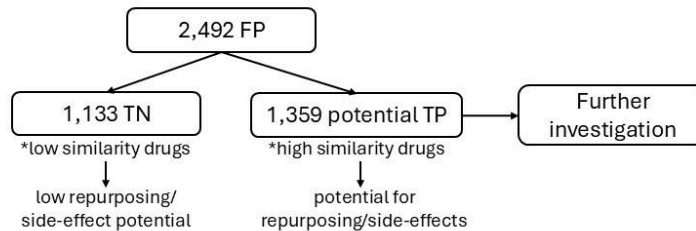


Figure 1. Network classification with EHR analysis allows for prioritization of drug repurposing hypotheses. (a) Schematic of EHR study design for comparison of network-based drug classes. (b) Boxplot distribution of A1C difference before and after drug exposure in patients on glucagon cluster drugs (0, blue) and patients on insulin or apolipoprotein cluster drugs (1, red). (c) Module overlap of predicted diabetes pathways across pathway size. Each point represents a predicted diabetes pathway connecting a non-diabetes drug to a diabetes phenotype. (d) Drug effect examples obtained through network classification and EHR analysis can be used to separate predicted diabetes drugs (currently classified as false positives (FP)) into likely true negatives (TN) and potential true positives (TP) that can be prioritized for further investigation.

Conclusion

In this study, we developed a framework for drug repurposing hypothesis prioritization through network classification and EHR analysis. We first studied diabetes-associated predictions in PathFX.

While approved anti-diabetics clustered into distinct pathway-based groups, EHR studies suggested equivalence of drug clusters on diabetes outcomes (i.e., A1C difference). We re-classified predicted diabetes drugs into groups with “high” or “low” overlap with approved pathways and discovered patients on “low” overlap drugs never experienced a diabetes diagnosis, suggesting they are true negatives and are unlikely candidates for drug repurposing. Our approach leverages protein-protein interaction data to assess similarity of unvalidated, predicted pathways and approved, known-effect pathways to generate drug repurposing hypotheses which can then be prioritized for further evaluation using EHR analyses. Investigating prescribing patterns in the EHR will additionally narrow down our list of drugs to investigate and influence the design of future studies. We hope to enable prioritization of drug repurposing hypotheses and increase the translational impact of drug-effect prediction models in drug development. Towards this goal, future work will focus on further developing our pipeline with additional EHR studies and expanding this analysis framework to other disease areas such as hypertension and cancer.

References

1. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. Rev. Genet.* **12**, 56–68 (2011).
2. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome networks and human disease. *Cell* **144**, 986–998 (2011).
3. Li, H., Li, T., Quang, D. & Guan, Y. Network propagation predicts drug synergy in cancers. *Cancer Res.* **78**, 5446–5457 (2018).
4. Gordon, D. E. *et al.* A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature* **583**, 459–468 (2020).
5. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for COVID-19. *Proc. Natl. Acad. Sci. U. S. A.* **118**, (2021).
6. Wilson, J. L. *et al.* PathFX provides mechanistic insights into drug efficacy and safety for regulatory review and therapeutic development. *PLoS Comput. Biol.* **14**, e1006614 (2018).
7. Ruiz, C., Zitnik, M. & Leskovec, J. Identification of disease treatment mechanisms through the multiscale interactome. *Nat Commun* **12**, 1796 (2021).
8. Wilson, J. L. *et al.* A network paradigm predicts drug synergistic effects using downstream protein-protein interactions. *CPT Pharmacometrics Syst Pharmacol* **11**, 1527–1538 (2022).
9. Suchard, M. A. *et al.* Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* **394**, 1816–1826 (2019).
10. Hripcsak, G. *et al.* Research Protocol: Large-scale evidence generation and evaluation across a network of databases for type 2 diabetes mellitus. Preprint at <https://ohdsi-studies.github.io/LegendT2dm/Protocol.html#References>.
11. Wishart, D. S. *et al.* DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34**, D668–72 (2006).
12. Eden, E., Navon, R., Steinfeld, I., Lipson, D. & Yakhini, Z. GOzilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* **10**, 48 (2009).