nature portfolio

https://doi.org/10.1038/s43587-021-00138-z

Supplementary information

Endophenotype-based in silico network medicine discovery combined with insurance record data mining identifies sildenafil as a candidate drug for Alzheimer's disease

In the format provided by the authors and unedited

Supplementary Information

Endophenotype-based in-silico network medicine discovery combined with insurance records data mining identifies sildenafil as a candidate drug for Alzheimer's disease

Fang et al., *Nature Aging* 2021

*Correspondence to:
Feixiong Cheng, Ph.D.
Lerner Research Institute, Cleveland Clinic
Email: chengf@ccf.org
Tel: +1-216-4447654; Fax: +1-216-6361609

Supplementary Information includes Supplementary Materials and Methods, Supplementary Results and Discussion, 9 Supplementary Figures, and 15 Supplementary Tables.

Table of Contents

Supplementary Materials and Methods

1. Differential gene expression analysis for Alzheimer's disease

We collected a total of 34 transcriptomics datasets consisting of original brain microarray data (n=17) and bulk RNA-seq data (n=17) from seven types of transgenic AD mouse models, including 5XFAD [1], APP/PS1, HO-TASTPM [2], Tg2576 [3], TgCRND8 [3], Tau P301L [2], and rTg4510 [4] (**Supplementary Table 14**). All the seven types of AD transgenic mouse models overexpress human APP mutants or human MAPT mutants, which can develop Aβ aggregation or neurofibrillary tangles (NFTs), resulting in cognitive impairment. For example, the rTg4510 mouse is a tauopathy model that overexpresses the human tau P301L mutation associated with familial frontotemporal dementia, which develops progressive age-related NFTs, neuronal loss, and behavioral impairments[5]. All the samples in datasets were derived from total brain, specific brain regions (hippocampus, cortex, or cerebellum), or brain-derived microglial cell.

The original brain microarray datasets were obtained from Gene Expression Omnibus (https://www.ncbi.nlm.nih.gov/geo), including GSE64398 (number of datasets [n]=9), GSE74437 (n=2), GSE74438 (n=2), GSE65067 (n=1), GSE53480 (n=1), GSE56772 (n=1), and GSE57583 (n=1) according to different mouse age and brain region. The detailed information of these GEO datasets is provided in **Supplementary Table 14,** including number of identified differentially expressed genes (DEGs), platform, significance of differential expression, genetic background, and age (month), etc. All raw expression data were $\log_2$ transformed, and all samples were quantile normalized together. Probe IDs in each dataset were mapped to NCBI entrez IDs and probes mapping to multiple genome regions or without corresponding entrez IDs were removed. The items were filtered if the gene have opposite $\log_2$ Fold change (FC) value (one positive and the other negative), otherwise the items with higher absolute value of $\log_2$ FC was kept. The remaining items were imported to R statistical processing environment using a LIMMA/Bioconductor package.[6] The differentially expressed genes

(DEGs) were identified using a conservative statistical threshold of False Discovery Rate (FDR) < 0.05 and $|\log_2 FC| > 1.5$ for differential expression. All the mouse DEGs were further converted into unique human-orthologous genes using Mouse Genome Informatics (MGI) database[7]. All analyses were performed using R software.

2. Differential protein expression (DEP) analysis for Alzheimer's disease

The ten sets of DEPs were assembled from 3 AD transgenic models in two recent publications[8,9]. The first publication by Savas et al. performed global quantitative proteomic analysis in hAPP and hAPP-PS1 mouse models at young (3 month [M]) and old ages (12 M)[8]. The samples were derived from frontal cortex (FC), hippocampal (HIP), and cerebellar (CB) extracts in mouse brain. The statistical significance of differential expression of all proteins was assessed using a two-tailed one-sample t-test on their corresponding peptide quantification ratios between both conditions. The obtained *P* values were False Discovery Rate (FDR)-adjusted for multiple hypothesis testing using the Benjamini–Hochberg correction [10]. DEPs for each brain region in mouse were determined with the threshold of FDR < 0.05 and converted to homologous human gene [7]. We obtained four sets of DEPs [hAPP_3M (number of DEP [n] =363), hAPP_12M (n=624), hAPP-PS1_3M (n=262) and hAPP-PS1_12M (n=476)] after merging the DEPs from different brain regions (**Supplementary Table 14**).

The other publication by Kim et al. performed quantitative proteomics to uncover molecular and functional signatures in HIP of two transgenic mouse models, including $ADLP_{APT}$ that carry three human transgenes (APP, PS1, and MAPT), and hAPP-PS1(5XFAD) mouse [9]. The $ADLP_{APT}$ mice, generated by crossing the 5XFAD strain with JNPL3 tau animals, could exhibit amyloid plaques, accelerated neurofibrillary tangle formation, neuronal loss in the CA1 area, and memory deficit at an early age. The 5XFAD mouse model develops early plaque formation, intraneuronal Aβ aggregation, neuron loss, and behavioral deficits.[1] Three different ages of mouse were used, including young (4 M),

middle (7 M), and old (10 M). The cut-off value for statistical significance was set to P<0.05 for the Student's t-test and $|\log_2 FC|>1.25$. After we converted mouse genes to homologous human genes[7], we obtained six sets of DEPs, including $ADLP_{APT}$_4M (n=53), $ADLP_{APT}$_7M (n=124), $ADLP_{APT}$_10M (n=299), hAPP-PS1_4M (n=54), hAPP-PS1_7M (n=168), and hAPP-PS1_10M (n=237).

3. Permutation test for largest connected component

A permutation test, also called a randomization test, is a standard test to compute statistical significance[11,12]. We extracted the same number of randomly selected genes (or drug targets) with similar connectivity (degree) as the original seed genes in the human interactome. Then we calculated the largest connected component [LCC] for this randomly gene set. We performed the permutation testing as below:

$$P = \frac{\#\{S_m(p) > S_m\}}{\#\{total\ permutations\}} \quad (1)$$

A nominal P was computed by counting the number of permutations in which the LCC of the random gene set ($S_m(p)$) is greater than observed LCC (S_m). We repeated the permutation test 1000 times.

4. Functional enrichment analysis

To check how our AD seed genes (N=144) relates to amyloidosis and tauopathy in AD, we collected 9 amyloidosis-related gene sets and 4 tauopathy-related gene sets from Molecular Signature Database (MSigDB)[13], and calculated their statistical significance of overlap with 144 AD seed genes. The results revealed that AD seed genes were significantly enriched in all of 13 gene sets using Fisher's exact test (**Supplementary Table 1**). This implies that the curated AD seed genes can strongly reflect phenotypic manifestations such as amyloidosis and tauopathy in AD.

5. The proximal pathway enrichment analysis (PxEA)

PxEA[14] was used to prioritize the 100 candidate drugs with the lowest z-score by AD seed module. Here the 13 gene sets related to amyloidosis and tauopathy used in functional enrichment analysis were regarded as AD-related pathway genes. In total, 1554 Reactome pathway gene sets from the MSigDB[13] were used for pathway background. Among the 100 candidate drugs, 65 out of them obtained favorable PxEA scores (ES) and significant p values ($p<0.05$). After adopting the same inclusion criteria (excluding the nutraceutical drugs, metal drugs, and radioligand diagnostic agents), 56 drug candidates were kept (**Supplementary Table 4**). Finally we found that 16 drugs had reported clinical or preclinical evidence in AD. Compared with 21 drugs with AD evidence from our human interactome-based network proximity, our approach performed slightly better than PxEA[14].

6. Blood-brain barrier (BBB) property

To highlight the drug candidates with good BBB property, we further collected the computational and literature-reported BBB penetrable evidence for the top 25 predicted drug candidates by our network approach. Specifically, we manually curated the literature-reported BBB penetrable evidence from PubMed database, and also used admetSAR to predict BBB property[15].

7. Pharmacoepidemiologic validation

7.1. Data description. The IBM® MarketScan® Medicare Supplemental Database was one of the first in the US to profile the healthcare experience of retirees with Medicare supplemental insurance paid by employers. The MarketScan Medicare Supplemental Database provides detailed cost, use and outcomes data for healthcare services performed in both inpatient and outpatient settings.

For most of the population, the medical claims are linked to outpatient prescription drug claims and person-level enrollment data through the use of unique patient or enrollee identifiers. Beneficiaries in the MarketScan Medicare Supplemental Database have drug

coverage; therefore, drug data are available and provide additional, often valuable, information. This feature makes the database a robust tool for pharmacoeconomic and outcomes research and helps provide insight into the drug use and spending patterns of older Americans.

In this analysis, the pharmacoepidemiology study utilized the MarketScan Medicare Claims database from 2012 to 2017. The dataset included individual-level diagnosis codes, procedure codes and pharmacy claim data for 7.23 million patients. Pharmacy prescriptions of sildenafil, diltiazem, losartan, glimepiride and metformin were identified by using RxNorm and National Drug Code (NDC).

7.2. Propensity score estimation. We define NE = north east, NC = north central, S = south, W = west, T2D = type 2 diabetes, HT = hypertension, and CAD = coronary artery disease. The propensity score of taking sildenafil vs. a comparator drug was estimated by the following logistic regression model:

$$\text{logit}[\Pr(\text{Drug} = \text{Sildenafil})] = \beta_0 + \beta_1 Age + \beta_2 Sex + \beta_3 1(Location = NE) + \beta_4 1(Location = NC) + \beta_5 1(Location = S) + \beta_6 T2D + \beta_7 HT + \beta_8 CAD$$

Further, among the subgroups defined by sex, the propensity score of taking sildenafil vs. a comparator drug was estimated by the following logistic regression model:

$$\text{logit}[\Pr(\text{Drug} = \text{Sildenafil})] = \beta_0 + \beta_1 Age + \beta_2 1(Location = NE) + \beta_3 1(Location = NC) + \beta_4 1(Location = S) + \beta_5 T2D + \beta_6 HT + \beta_7 CAD$$

7.3. Survival analysis. Stratified Cox models were used to compare the AD risks. For sildenafil vs. a comparator drug, the analyses were stratified (n strata = 10) by the estimated propensity score. For sildenafil vs. non-sildenafil, the analyses were stratified based on the subgroups defined by sex, T2D diagnose, HT diagnoses, or CAD diagnoses (n strata = 16). Additionally, we conduct analysis within each of the subgroup. All analyses use the following univariate Cox model:

$$\log[h(t)] = \log[h_0(t)] + \theta 1[\text{Drug} = \text{Sildenafil}]$$

Supplementary Results and Discussion

1. Proof-of-concept of Alzheimer's disease module

A disease module is composed of disease-associated proteins in the human interactome that form one or several connected subgraphs rather than being scattered randomly[16]. The main hypothesis of AD network medicine paradigm is that the underlying AD endophenotype network module should reflect both amyloid and tau pathology. Here we used the manually curated AD seed gene set for generating endophenotype network model (AD seed module) to characterize pathogenesis and therapeutic development for AD (**Extended Data Fig. 1**), which includes both amyloidosis and tauopathy seed genes, and genes characterizing other pathology hypothesis (i.e., neuroinflammation, and vascular dysfunction). The AD seed module includes 227 protein-protein interactions (PPIs) (edges or links) connecting 102 unique proteins (nodes). The network analysis shows that 12 proteins have degree (*K*) number higher than 10, and 11 of them belong to amyloid, tau, and amyloid & tau seed gene. Importantly, APP (*K*=39) and MAPT (*K*=18), encoding amyloid-beta precursor protein and microtubule-associated protein tau, have the highest degrees within the module, further highlighting the importance of amyloid and tau hypothesis in the pathogenesis of AD.

Yet, whether the manually curated AD seed gene set could form statistically significant modules in the human protein-protein interactome remains unknown. To test this hypothesis, we perform the largest connected component (LCC) analysis for the manually curated AD seed gene set. We found the AD seed module form significantly LCC (102/144 [70.8%], $P<0.001$, permutation test) compared to the same number of randomly selected genes with similar connectivity (degree) as the original seed genes in the human interactome (**Supplementary Fig. 1**). Collectively, we demonstrated the strong modularity of AD seed module based on AD seed gene set.

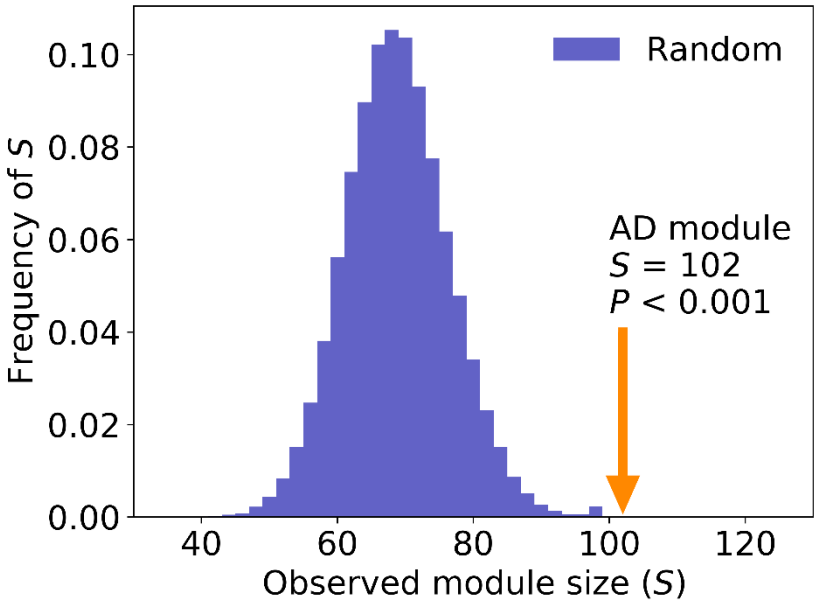
2. Clinical efficacy of endophenotype network-based drug repurposing

There are over 30 FDA-approved drugs being repurposed for AD[17]. To test the clinical efficacy of endophenotype hypothesis, we first collected the ongoing repurposable AD drugs from ClinicalTrials.gov. After excluding drugs without known physical drug-target interactions and drugs whose mechanism were not related to reducing amyloid or tau pathology, 21 drugs were obtained (**Supplementary Table 2**). We classified these drugs into three categories: (a) anti-amyloid, (b) anti-tau, and (c) dual-targeting both amyloid & tau (an endophenotype network model). Then we applied network proximity to quantify network distance between 21 drugs' molecular targets (drug modules) and AD network modules, including 3 network modules by assembling experimentally validated (seed) genes in amyloidosis (amyloid), tauopathy (tau), and AD (characterized by both Amyloid and Tau), and 34 disease modules from DEG sets derived from transcriptomics data (including microarray and bulk RNA-sequencing) from AD genetic mouse. As shown in **Extended Data Figs. 2** and **3a**, we found the molecular targets of drugs targeting amyloid & Tau had significantly closer network distance with 37 AD disease modules, compared with those targeting single amyloid (Wilcoxon test, $P=1.67\times10^{-15}$) or tau (Wilcoxon test, $P=6.87\times10^{-42}$). For example, riluzole, an FDA-approved drug for amyotrophic lateral sclerosis (ALS), is now in Phase II recruiting for patients for mild AD (NCT01703117). Riluzole could ameliorate learning and memory deficits in AD-like amyloid and tau pathology mouse models[18, 19]. Network proximity demonstrated that riluzole had an average Z-score of -1.30 for 37 AD disease modules, of which 16 modules ($P<0.05$) had significant close distance with molecular target of riluzole. Meanwhile, the molecular targets of drugs targeting amyloid also had significantly closer network distance with disease modules, compared with those targeting tau alone (Wilcoxon test, $P=1.88\times10^{-15}$).

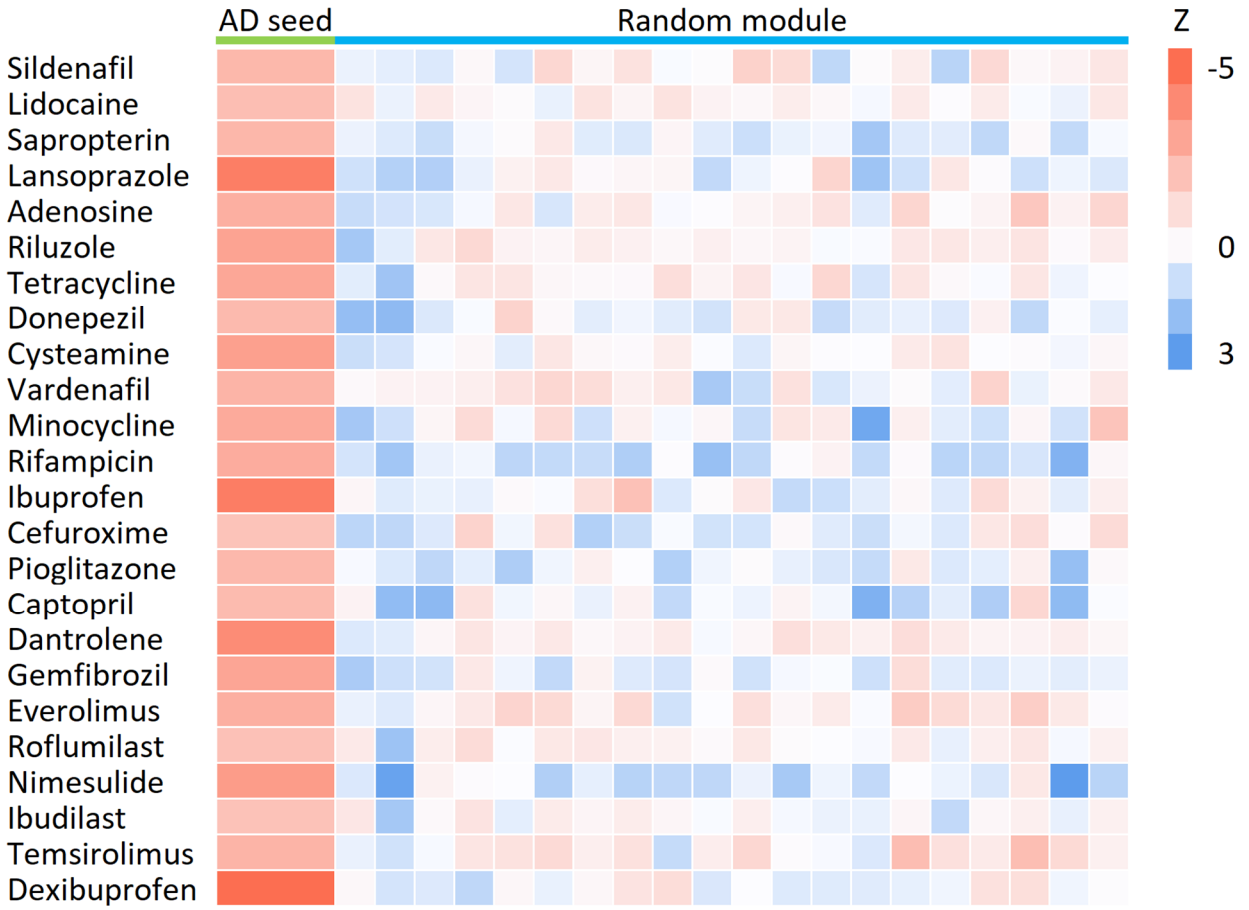
To confirm this hypothesis, we further examined the network distance between these 21 drug modules and 10 AD disease modules derived from proteomics data of AD

genetic mouse (**Extended Data Fig. 3b**). The similar conclusion was drawn that the drug modules targeting amyloid & tau had significantly closer network distance with proteomics modules, compared with those targeting amyloid (P=0.028, Wilcoxon test) or tau module alone (P=2.18×10^{-16}, Wilcoxon test). We also found that proteomics disease modules have closer network distance with drug modules, compared with transcriptomics disease modules. For example, the 12 drugs targeting amyloid & tau had an average Z-score of -0.16 for 34 transcriptomic modules, compared to an average Z-score of -0.99 for 10 proteomics modules. This implied that the proteomic data-derived disease modules are more suitable for network proximity prediction in AD than transcriptomic data-derived endophenotype disease modules.

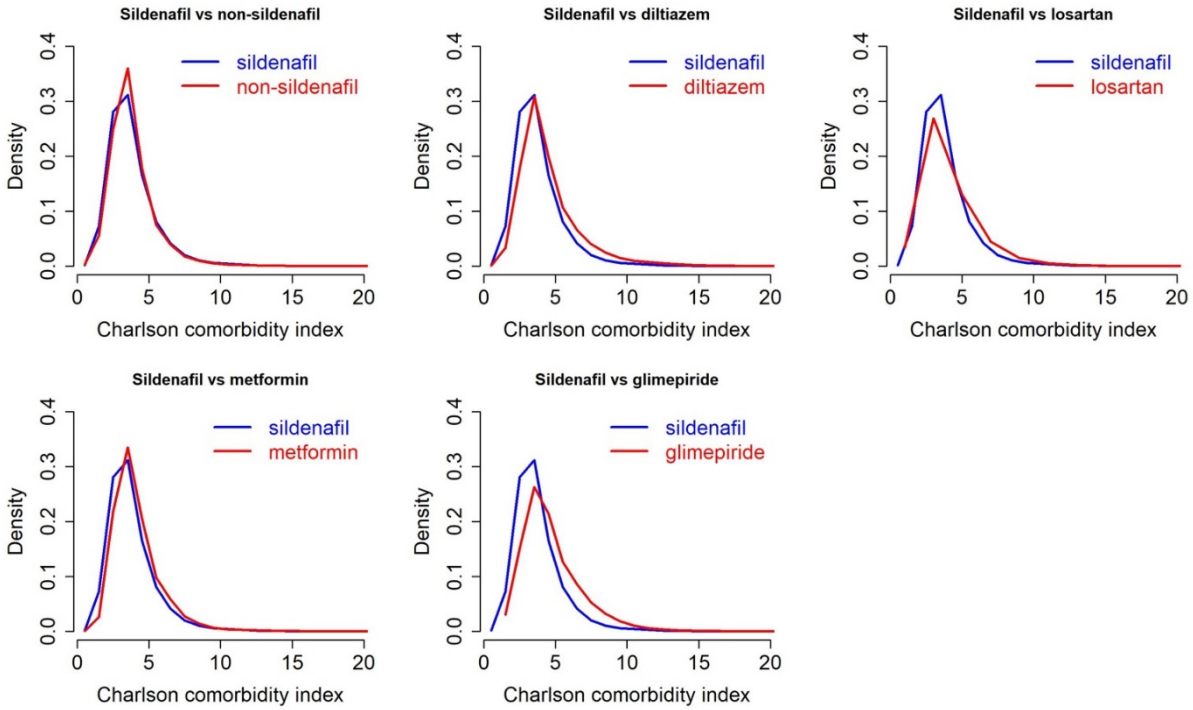
Supplementary Figures



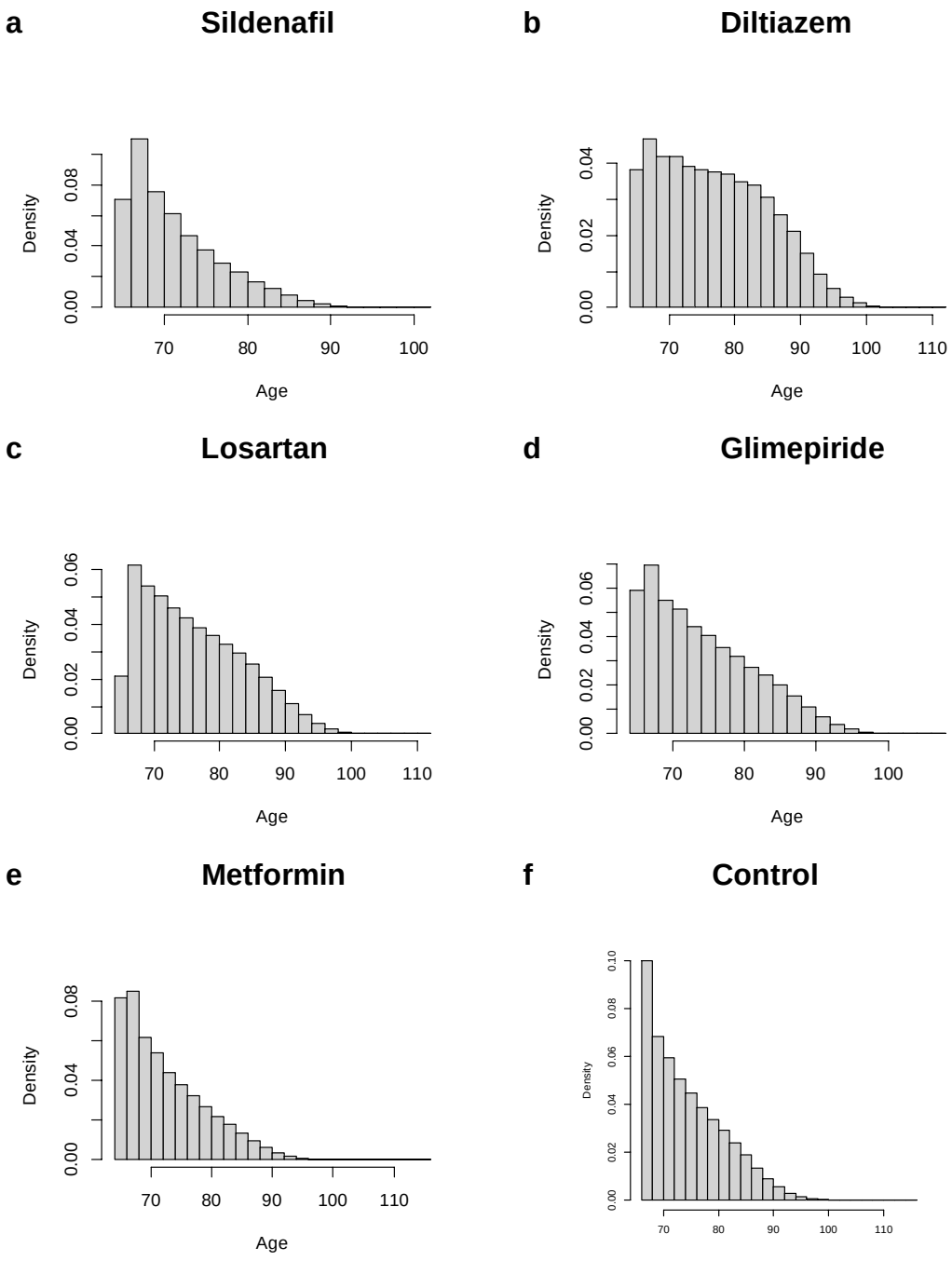
Supplementary Figure 1. The largest connected component (LCC) analysis for Alzheimer's disease seed module. Experimentally identified AD seed genes form significantly large disease module in the human interactome (one-sided, $P < 0.001$, 1,000 permutation test). Experimentally validated gene products (proteins) are likely to cluster in the same network neighborhood or disease module within the human protein-protein interactome. The observed module size (S) 102 (seed genes, orange line) is significantly larger than the random background.



Supplementary Figure 2. A heat map highlighting a low bias of network proximity scores for drug candidates shown in Figure 2. We compared Z-score (Z, network proximity measure) between AD seed module and randomly generated modules (module size of 144, 20 times of randomized experiments).



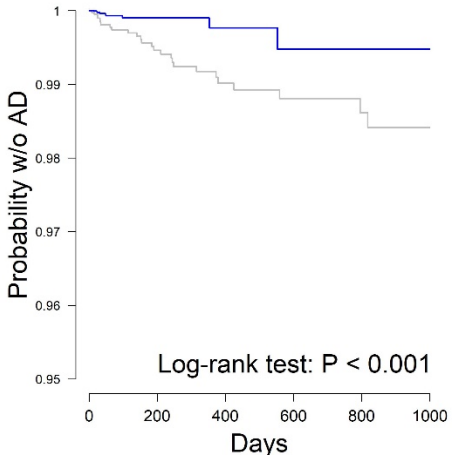
Supplementary Figure 3. Charlson comorbidity index distribution between sildenafil usage group and the comparator treatment groups (receiving comparator Rxs).



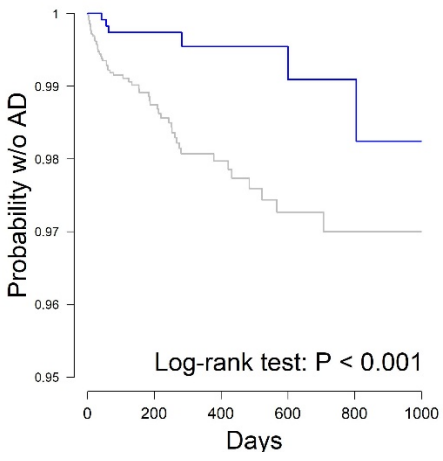
Supplementary Figure 4. The age distribution of each drug population group. (a) Sildenafil; **(b)** Diltiazem; **(c)** Losartan; **(d)** Glimepiride; **(e)** Metformin; **(f)** propensity score matching non-sildenafil cohort (control).

Sildenafil Vs. Diltiazem

erectile dysfunction

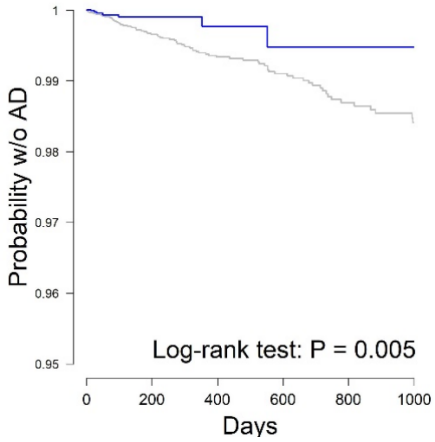


pulmonary hypertension

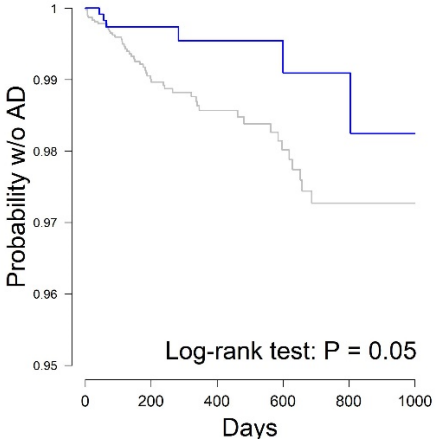


Sildenafil Vs. Losartan

erectile dysfunction

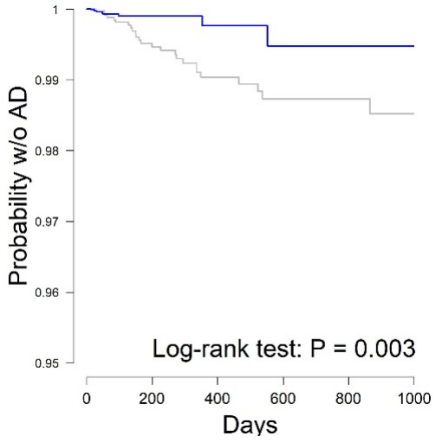


pulmonary hypertension

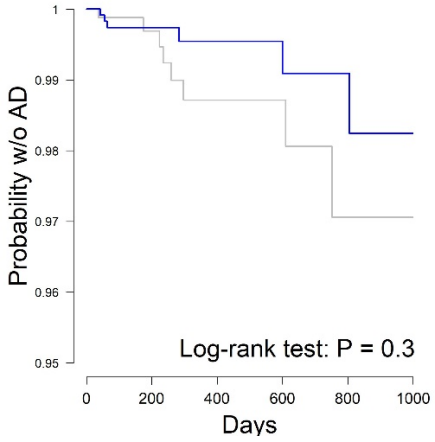


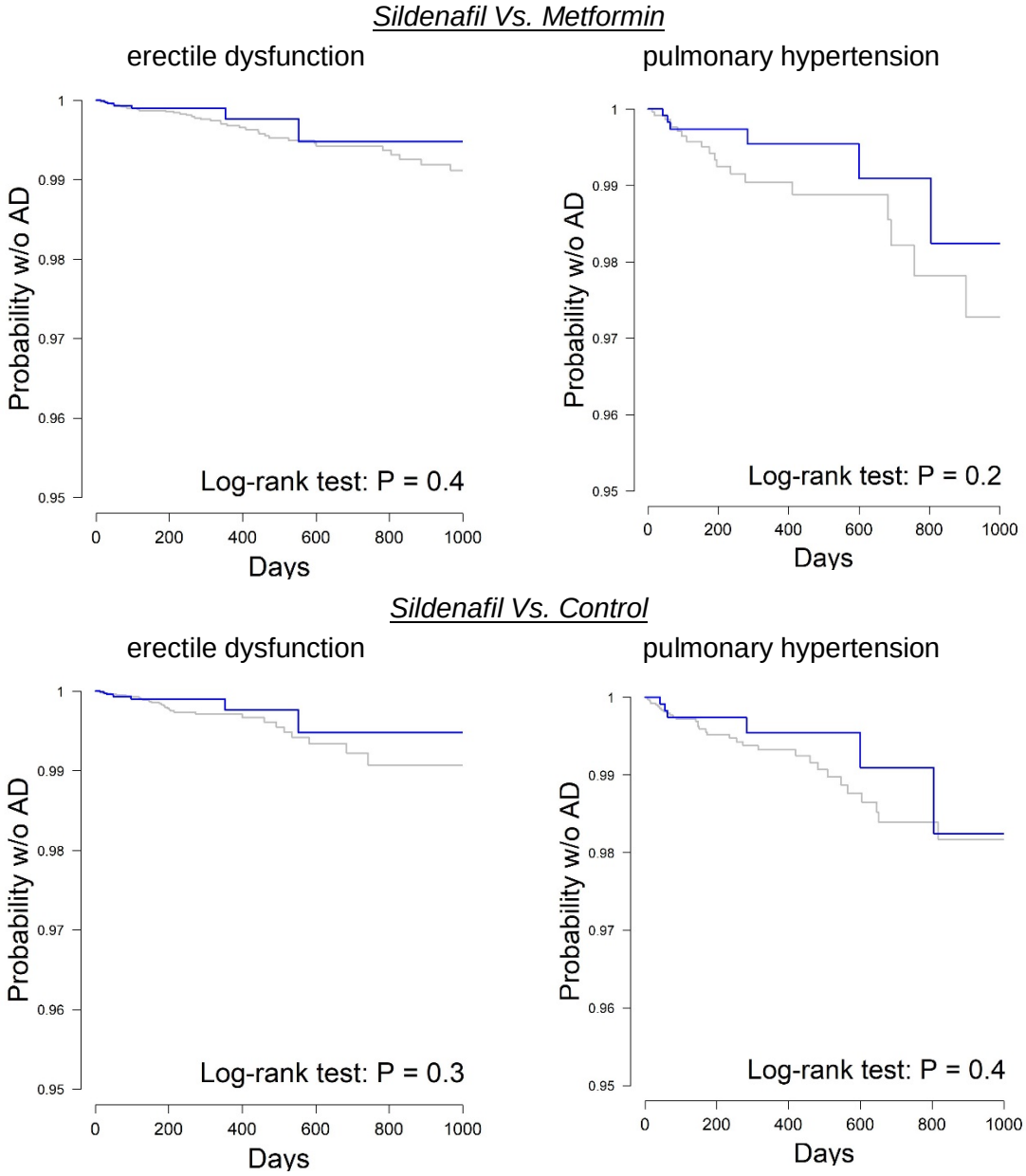
Sildenafil Vs. Glimepiride

erectile dysfunction

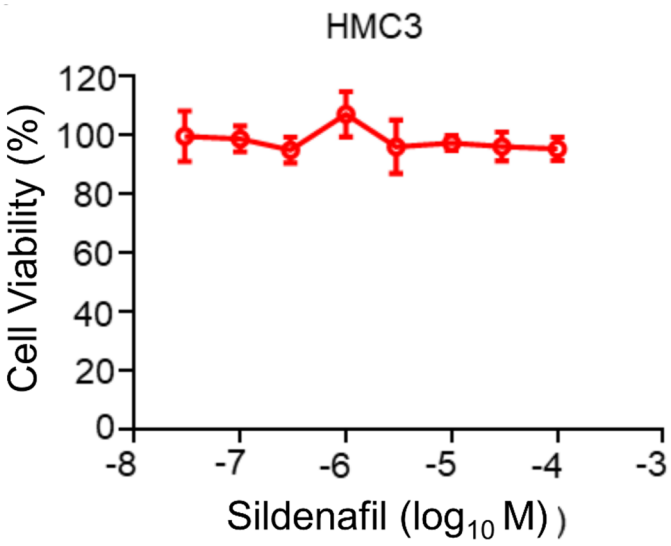


pulmonary hypertension

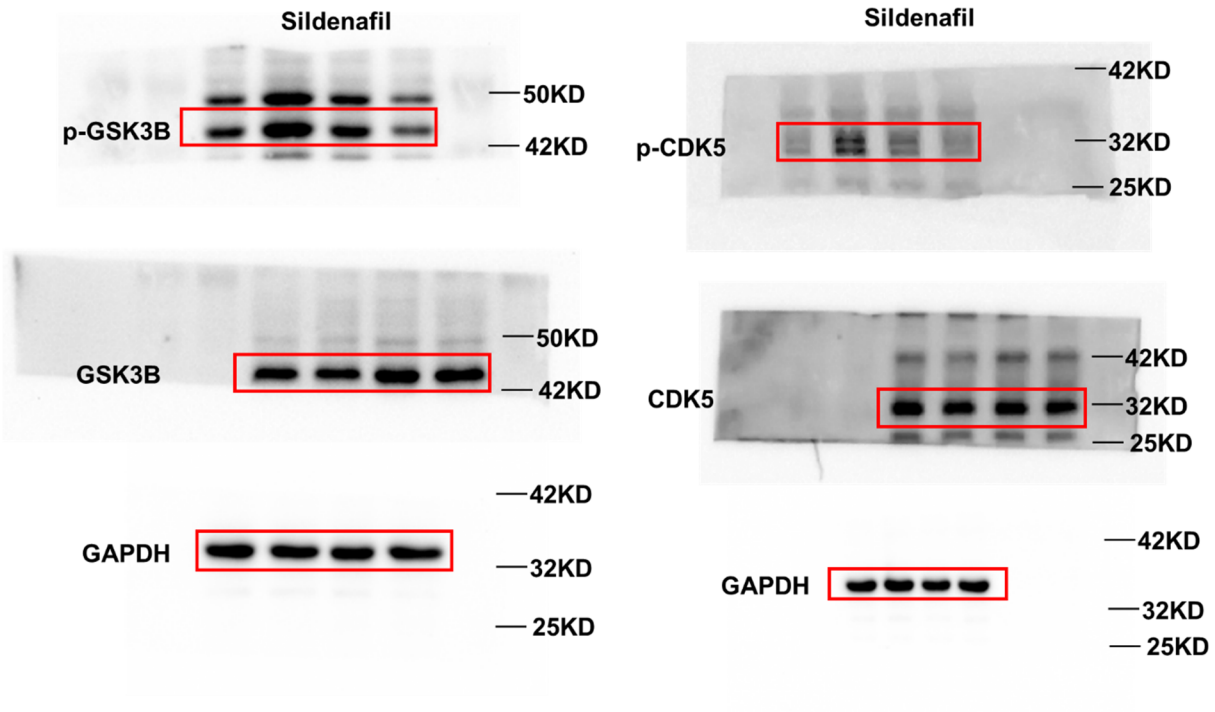




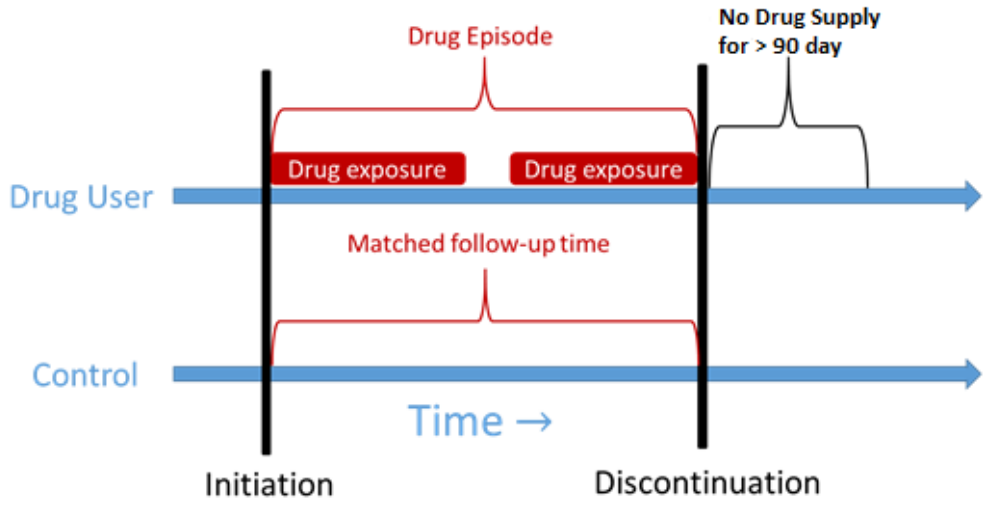
Supplementary Figure 5. Kaplan-Meier plot and P-value under log-rank test (blue line = Sildenafil, gray line = comparator drug or control)**.** All p-values in the Kaplan Meier plots are based on two-sided log-rank test.



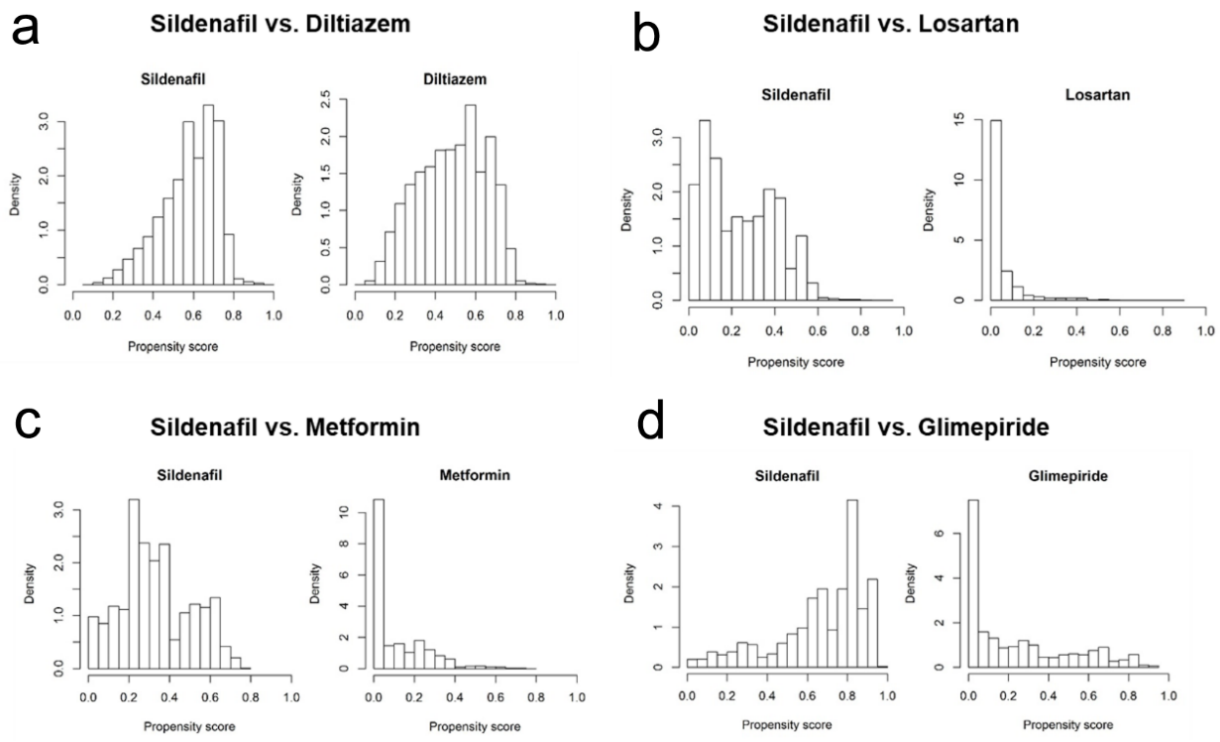
Supplementary Figure 6. Effects of sildenafil on the cell viability of HMC3 cells. HMC3 cells were treated with indicated concentrations of sildenafil for 48 h and cell viability was determined using MTT. Data are represented as mean ± standard error of mean (SEM, n = 3) and each experiment was performed at least three times in duplicate.



Supplementary Figure 7. Effects of sildenafil on LPS-induced activation of GSK3β and CDK5 in human microglia HMC3 cells. HMC3 cells were pretreated with sildenafil and followed LPS treatment (1 μg/mL, 30 min). The total cell lysates were collected and subjected to Western blot analysis.



Supplementary Figure 8. The definition of a drug episode. A drug episode is defined as the time between drug initiation and drug discontinuation. Specifically, drug initiation is defined as the first day of drug supply (i.e., 1st prescription date). Drug discontinuation is defined as the last day of drug supply (i.e., last prescription date + days of supply) and without drug supply for the next 90 days.



Supplementary Figure 9. The estimated propensity score among the users of sildenafil and the comparator drugs. (a) Sildenafil vs. Diltiazem; **(b)** Sildenafil vs. Losartan; **(c)** Sildenafil vs. Metformin; **(d)** Sildenafil vs. Glimepiride.

Supplementary Table legends

Supplementary Table 1. Functional enrichment analysis result for the AD seed gene set (N=144). (xlsx)

Supplementary Table 2. The detailed information of 21 existing drugs in AD clinical trials to test the clinical efficacy of endophenotype hypothesis. (xlsx)

Supplementary Table 3. The anti-AD clinical, *in vivo*, and blood–brain barrier (BBB) properties from publicly available database and *in silico* prediction for the 66 potential AD drugs. (xlsx)

Supplementary Table 4. Prioritizing 100 candidate drugs with the lowest z-score by AD seed module using PxEA approach and obtaining 56 candidate drugs. (xlsx)

Supplementary Table 5. 120 FDA-approved drugs with reported clinical or preclinical (*in vivo*) evidence in AD from the AlzGPS database. (xlsx)

Supplementary Table 6. Network proximity scores between molecular targets of 12 ongoing repurposable AD drugs (dually targeting amyloid & tau) and AD seed module. (xlsx)

Supplementary Table 7. Hazard ratios (HR) and 95% confidence intervals (CI) across five drug cohort studies after adjusting mild cognitive impartment (MCI). All p-values associated with confidence interval are based on Cox model.

Drug Cohorts	Hazard Ratio	95% CI	P-value
Sildenafil vs. Non-sildenafil	0.14	0.11 – 0.17	$<10^{-8}$
Sildenafil vs. Diltiazem	0.34	0.28 – 0.42	$<10^{-8}$
Sildenafil vs. Losartan	0.44	0.36 – 0.55	$<10^{-8}$

Sildenafil vs. Glimepiride	0.35	0.28 – 0.45	$<10^{-8}$
Sildenafil vs. Metformin	0.36	0.30 – 0.45	$<10^{-8}$

Supplementary Table 8. Age stratified analysis for all comparisons of sildenafil. (xlsx)

Supplementary Table 9. Sex-based subgroup analyses for sildenafil. (xlsx)

Supplementary Table 10. Average daily dosage of Sildenafil between male and female individuals.

Cohorts	Average daily dosage of Sildenafil in mg (±standard deviation [SD])	P value
Male	75.9 (32.4)	0.03
Female	22.1 (15.0)	

Supplementary Table 11. Subgroup analyses for sildenafil vs. comparator drugs in both erectile dysfunction (ED) cohort and pulmonary hypertension cohort.

Referent group	erectile dysfunction		pulmonary hypertension	
	N Total (AD Case)	Hazard Ratio (95% CI)	N Total (AD Case)	Hazard Ratio (95% CI)
Sildenafil	17,889 (10)	NA	1,606 (7)	NA
Diltiazem	3,983 (27)	0.32 (0.14 – 0.76)	4,191 (62)	0.39 (0.18 - 0.87)
Losartan	13,327 (87)	0.49 (0.23 – 1.02)	4,707 (52)	0.59 (0.27 – 1.33)
Glimepiride	3,599 (26)	0.50 (0.15 – 1.69)	961 (8)	0.63 (0.16 – 2.43)
Metformin	12,681 (43)	0.66 (0.27 – 1.59)	2,370 (21)	0.87 (0.35 – 2.21)
Control	71,028 (57)	0.70 (0.36 – 1.36)	6,290 (39)	0.70 (0.31 – 1.57)

Supplementary Table 12. Average annual cost between sildenafil users and non-sildenafil users.

Drug Cohort	Average annual cost per 1,000 US dollars (±standard deviation [SD])	P-value
Sildenafil users	32.6 (57.6)	0.36
Other drug users	27.2 (64.7)	

Supplementary Table 13. List of experimentally validated (seed) genes in amyloidosis (amyloid), tauopathy (tau) and Alzheimer's disease (AD). (xlsx)

Supplementary Table 14. Transcriptomic and proteomic datasets from transgenic AD mouse models. (xlsx)

Supplementary Table 15. Phenotype definitions for Alzheimer's disease, Mild cognitive impartment (MCI), Type 2 diabetes, Hypertension, and Coronary artery disease.

Alzheimer's disease [20, 21]
3310 (Alzheimer's disease), F00 (Dementia in Alzheimer disease), F000A (Dementia in Alzheimer disease with early onset), F001A (Dementia in Alzheimer disease with late onset), F002A (Dementia in Alzheimer disease, atypical or mixed type), F009A (Dementia in Alzheimer disease, unspecified), G30 (Alzheimer's disease), G300 (Alzheimer disease with early onset), G301 (Alzheimer disease with late onset), G308 (Other Alzheimer disease), G309 (Alzheimer disease, unspecified)
Mild cognitive impartment (MCI) [20, 21]
ICD-9 331.83; ICD-10 G31.84
Type 2 diabetes [22, 23]
25000 (Diabetes mellitus without mention of complication, type II or unspecified type, not stated as uncontrolled), 25050 (Diabetes with ophthalmic manifestations, type II or unspecified type, not stated as uncontrolled), 25002 (Diabetes mellitus without mention of complication, type II or unspecified type, uncontrolled), 25052 (Diabetes with ophthalmic manifestations, type II or unspecified type, uncontrolled), 25010 (Diabetes with ketoacidosis, type II or unspecified type, not stated as uncontrolled), 25060 (Diabetes with neurological manifestations, type II or unspecified type, not stated as uncontrolled), 25012 (Diabetes with ketoacidosis, type II or unspecified type, uncontrolled), 25062 (Diabetes with neurological manifestations, type II or unspecified type, uncontrolled), 25020 (Diabetes

with hyperosmolarity, type II or unspecified type, not stated as uncontrolled), 25070 (Diabetes with peripheral circulatory disorders, type II or unspecified type, not stated as uncontrolled), 25022 (Diabetes with hyperosmolarity, type II or unspecified type, uncontrolled), 25072 (Diabetes with peripheral circulatory disorders, type II or unspecified type, uncontrolled), 25030 (Diabetes with other coma, type II or unspecified type, not stated as uncontrolled), 25080 (Diabetes with other specified manifestations, type II or unspecified type, not stated as uncontrolled), 25032 (Diabetes with other coma, type II or unspecified type, uncontrolled), 25082 (Diabetes with other specified manifestations, type II or unspecified type, uncontrolled), 25040 (Diabetes with renal manifestations, type II or unspecified type, not stated as uncontrolled), 25090 (Diabetes with unspecified complication, type II or unspecified type, not stated as uncontrolled), 25042 (Diabetes with renal manifestations, type II or unspecified type, uncontrolled), 25092 (Diabetes with unspecified complication, type II or unspecified type, uncontrolled), E089 (Diabetes mellitus due to underlying condition without complications), E1100 (Type 2 diabetes mellitus with hyperosmolarity without nonketotic hyperglycemic-hyperosmolar coma), E1101 (Type 2 diabetes mellitus with hyperosmolarity with coma), E1110 (Type 2 diabetes mellitus with ketoacidosis without coma), E1121 (Type 2 diabetes mellitus with diabetic nephropathy), E1122 (Type 2 diabetes mellitus with diabetic chronic kidney disease), E1129 (Type 2 diabetes mellitus with other diabetic kidney complication), E11311 (Type 2 diabetes mellitus with unspecified diabetic retinopathy with macular edema), E11319 (Type 2 diabetes mellitus with unspecified diabetic retinopathy without macular edema), E113219 (Type 2 diabetes mellitus with mild nonproliferative diabetic retinopathy with macular edema, unspecified eye), E113291 (Type 2 diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, right eye), E113292 (Type 2 diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, left eye), E113293 (Type 2 diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, bilateral), E113299 (Type 2 diabetes mellitus with mild nonproliferative diabetic retinopathy without macular edema, unspecified eye), E113319 (Type 2 diabetes mellitus with moderate nonproliferative diabetic retinopathy with macular edema, unspecified eye), E113391 (Type 2 diabetes mellitus with moderate nonproliferative diabetic retinopathy without macular edema, right eye), E113392 (Type 2 diabetes mellitus with moderate nonproliferative diabetic retinopathy without macular edema, left eye), E113393 (Type 2 diabetes mellitus with moderate nonproliferative diabetic retinopathy without macular edema, bilateral), E113399 (Type 2 diabetes mellitus with moderate nonproliferative diabetic retinopathy without macular edema, unspecified eye), E113419 (Type 2 diabetes mellitus with severe nonproliferative diabetic retinopathy with macular edema, unspecified eye), E113491 (Type 2 diabetes mellitus with severe nonproliferative diabetic retinopathy with macular edema, right eye), E113492 (Type 2 diabetes mellitus with severe nonproliferative diabetic retinopathy with macular edema, left eye), E113493 (Type 2 diabetes mellitus with severe nonproliferative diabetic retinopathy without macular edema, bilateral), E113499 (Type 2 diabetes mellitus with severe nonproliferative diabetic retinopathy without macular edema, unspecified eye), E113519 (Type 2 diabetes mellitus with proliferative diabetic retinopathy with macular edema, unspecified eye), E113591 (Type 2 diabetes mellitus with proliferative diabetic retinopathy without macular edema, right eye), E113592 (Type 2 diabetes mellitus with proliferative diabetic retinopathy without macular edema, left eye), E113593 (Type 2 diabetes mellitus with proliferative diabetic

retinopathy without macular edema, bilateral), E113599 (Type 2 diabetes mellitus with proliferative diabetic retinopathy without macular edema, unspecified eye), E1136 (Type 2 diabetes mellitus with diabetic cataract), E1139 (Type 2 diabetes mellitus with other diabetic ophthalmic complication), E1140 (Type 2 diabetes mellitus with diabetic neuropathy, unspecified), E1142 (Type 2 diabetes mellitus with diabetic polyneuropathy), E1143 (Type 2 diabetes mellitus with diabetic autonomic (poly)neuropathy), E1144 (Type 2 diabetes mellitus with diabetic amyotrophy), E1149 (Type 2 diabetes mellitus with other diabetic neurological complication), E1151 (Type 2 diabetes mellitus with diabetic peripheral angiopathy without gangrene), E1152 (Type 2 diabetes mellitus with diabetic peripheral angiopathy with gangrene), E1159 (Type 2 diabetes mellitus with other circulatory complications), E11610 (Type 2 diabetes mellitus with diabetic neuropathic arthropathy), E11618 (Type 2 diabetes mellitus with other diabetic arthropathy), E11620 (Type 2 diabetes mellitus with diabetic dermatitis), E11621 (Type 2 diabetes mellitus with foot ulcer), E11628 (Type 2 diabetes mellitus with other skin complications), E11630 (Type 2 diabetes mellitus with periodontal disease), E11641 (Type 2 diabetes mellitus with hypoglycemia with coma), E11649 (Type 2 diabetes mellitus with hypoglycemia without coma), E1165 (Type 2 diabetes mellitus with hyperglycemia), E1169 (Type 2 diabetes mellitus with other specified complication), E118 (Type 2 diabetes mellitus with unspecified complications), E119 (Type 2 diabetes mellitus without complications), E1300 (Other specified diabetes mellitus with hyperosmolarity without nonketotic hyperglycemic-hyperosmolar coma), E1310 (Other specified diabetes mellitus with ketoacidosis without coma), E1311 (Other specified diabetes mellitus with ketoacidosis with coma), E1322 (Other specified diabetes mellitus with diabetic chronic kidney disease), E1329 (Other specified diabetes mellitus with other diabetic kidney complication), E1339 (Other specified diabetes mellitus with other diabetic ophthalmic complication), E1349 (Other specified diabetes mellitus with other diabetic neurological complication), E1351 (Other specified diabetes mellitus with diabetic peripheral angiopathy without gangrene), E13628 (Other specified diabetes mellitus with other skin complications), E1365 (Other specified diabetes mellitus with hyperglycemia), E1369 (Other specified diabetes mellitus with other specified complication), E138 (Other specified diabetes mellitus with unspecified complications), E139 (Other specified diabetes mellitus without complications)
Hypertensions [24, 25]
4010 (Malignant essential hypertension), 4011 (Benign essential hypertension), 4019 (Unspecified essential hypertension), I10 (Essential (Primary) Hypertension), I169 (Sequelae of cerebrovascular disease)
Coronary artery disease [26, 27, 28]
410 (Acute myocardial infarction), 411 (Other acute and subacute forms of ischemic heart disease), 412 (Old myocardial infarction), 4131 (Prinzmetal angina), 414 (Other forms of chronic ischemic heart disease), I20 (Angina pectoris), I21 (ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction), I22 (Subsequent ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction), I23 (Certain current complications following ST elevation (STEMI) and non-ST elevation (NSTEMI) myocardial infarction (within the 28 day period)), I24 (Other acute ischemic heart diseases), I25 (Chronic ischemic heart disease)

References

1. Bouter, Y. et al. Deciphering the molecular profile of plaques, memory decline and neuron loss in two mouse models for Alzheimer's disease by deep sequencing. *Front. Aging Neurosci.* **6**, 75 (2014).

2. Matarin, M. et al. A genome-wide gene-expression analysis and database in transgenic mice during development of amyloid or tau pathology. *Cell Rep.* **10**, 633-644 (2015).

3. Rothman, S. M. et al. Human Alzheimer's disease gene expression signatures and immune profile in APP mouse models: a discrete transcriptomic view of Abeta plaque pathology. *J. Neuroinflammation.* **15**, 256 (2018).

4. Wang, H. et al. Genome-wide RNAseq study of the molecular mechanisms underlying microglia activation in response to pathological tau perturbation in the rTg4510 tau transgenic animal model. *Mol. Neurodegener.* **13**, 65 (2018).

5. Ramsden, M. et al. Age-dependent neurofibrillary tangle formation, neuron loss, and memory impairment in a mouse model of human tauopathy (P301L). *J. Neurosci.* **25**, 10637-10647 (2005).

6. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

7. Eppig, J. T. et al. Mouse Genome Informatics (MGI): Resources for Mining Mouse Genetic, Genomic, and Biological Data in Support of Primary and Translational Research. *Methods Mol. Biol.* **1488**, 47-73 (2017).

8. Savas, J. N. et al. Amyloid Accumulation Drives Proteome-wide Alterations in Mouse Models of Alzheimer's Disease-like Pathology. *Cell Rep.* **21,** 2614-2627 (2017).

9. Kim, D. K. et al. Molecular and functional signatures in a novel Alzheimer's disease mouse model assessed by quantitative proteomics. *Mol. Neurodegener*. **13,** 2 (2018).

10. Benjamini, Y. et al. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289-300 (1995).

11. Huang, Y. et al. A Systems Pharmacology Approach Uncovers Wogonoside as an Angiogenesis Inhibitor of Triple-Negative Breast Cancer by Targeting Hedgehog Signaling. *Cell Chem. Biol.* **26**, 1143-1158 (2019).
12. Cheng, F. et al. Network-based prediction of drug combinations. *Nat. Commun.* **10,** 1197 (2019).
13. Liberzon, A. et al. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Systems*, **1**, 417-425 (2015).
14. Aguirre-Plans, J. et al. Proximal Pathway Enrichment Analysis for Targeting Comorbid Diseases via Network Endopharmacology. *Pharmaceuticals (Basel).* **11,** 61 (2018).
15. Cheng, F. et al. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties. *J. Chem. Inf. Model.* 2012, **52,** 3099-3105 (2012).
16. Menche, J. et al. Disease networks. Uncovering disease-disease relationships through the incomplete interactome. *Science* **347,** 1257601 (2015).
17. Cummings, J. et al. Alzheimer's disease drug development pipeline: 2019. *Alzheimers Dement.* **5**, 272-293 (2019).
18. Hunsberger, H. C. et al. Riluzole rescues glutamate alterations, cognitive deficits, and tau pathology associated with P301L tau expression. *J. Neurochem.* **13**5, 381-394 (2015).
19. Mokhtari, Z. et al. Riluzole ameliorates learning and memory deficits in Abeta25-35-induced rat model of Alzheimer's disease and is independent of cholinoceptor activation. *Biomed. Pharmacother.* **87,** 135-144 (2017).
20. Wei, W. Q. et al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J. Am. Med. Inform. Assoc.* **23,** e20-27 (2016).
21. Wilkinson, T. et al. Identifying dementia cases with routinely collected health data: A systematic review. *Alzheimers Dement.* **14,** 1038-1051 (2018).
22. Kho, A. N. et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *J. Am. Med. Inform.*

Assoc. **19**, 212-218 (2012).

23. Wei, W. Q. et al. Impact of data fragmentation across healthcare centers on the accuracy of a high-throughput clinical phenotyping algorithm for specifying subjects with type 2 diabetes mellitus. *J. Am. Med. Inform. Assoc.* **19**, 219-224 (2012).

24. Federman, D. G. et al. Relationship between provider type and the attainment of treatment goals in primary care. *Am. J. Manag. Care* **11,** 561-566 (2005).

25. Banerjee, D. et al. Underdiagnosis of hypertension using electronic health records. *Am. J. Hypertens.* **25,** 97-102 (2012).

26. Federman, D. G. et al. Relationship between provider type and the attainment of treatment goals in primary care. *Am. J. Manag. Care.* **11**, 561-566 (2005).

27. Smilowitz, N. R. et al. Association Between Anemia, Bleeding, and Transfusion with Long-term Mortality Following Noncardiac Surgery. *Am. J. Med.* **129**, 315-323 (2016).

28. Abul-Husn, N. S. et al. Genetic identification of familial hypercholesterolemia within a single U.S. health care system. *Science* **354**, aaf7000 (2016).