Modeling Alzheimer's Disease by Combining Knowledge Extracted from Biomedical Literature with Biomedical Ontologies

Scott A. Malec, PhD,[1] Sanya B. Taneja, MS,[2] Steven M. Albert, PhD, MS,[4] Helmet T. Karim, PhD,[6] Arthur S. Levine, MD,[3,7] Paul W. Munro, PhD,[5] C. Elizabeth Shaaban, PhD, MPH,[4] Jonathan S. Silverstein, MD,[1] Kailyn F. Witonsky, BS,[3] Tiffany J. Callahan, MPH,[8] Richard D. Boyce, PhD[1,2]

[1]DBMI, School of Medicine, University of Pittsburgh (Pitt); [2]Intelligent Systems Program (Pitt); [3]School of Medicine (Pitt); [4]Department of Epidemiology, School of Public Health (Pitt); [5]School of Computing and Information (Pitt); [6]Department of Psychiatry, School of Medicine (Pitt); [7]Brain Institute (Pitt); [8]Computational Bioscience Program, University of Colorado Anschutz Medicine Campus

Introduction

- Alzheimer's Disease (AD) is a progressive neurodegenerative disease and the most common cause of dementia with a multifactorial etiology
- Routinely collected health data may hold clues to causes
- These data pose challenges such as confounding[1]
- Adjusting for common causes (confounders[1,2]) reduces bias, while adjusting for common effects (colliders[3,4]) or intermediate variables (mediators[5]) is harmful, per the examples in Figure A

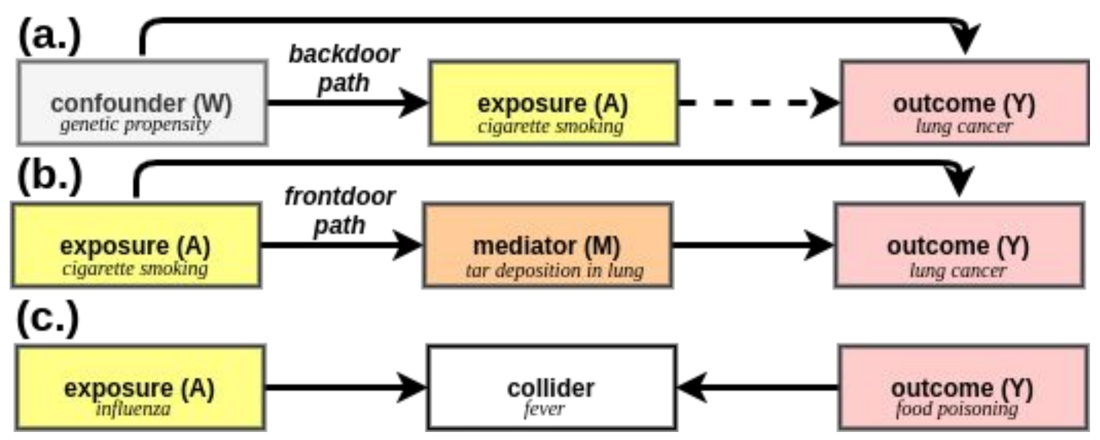

Figure A. Illustration depicts examples of confounders, mediators, and colliders. In (c.), fever (a collider) induces bias between influenza and food poisoning

- Literature may hold clues about which variables to control for[6,7]
- Literature is incomplete[8], machine reading has low recall
- This paper describes a pipeline to address these obstacles, investigating depression as a risk factor for AD[9,10]

Methods & Materials

- Figure B illustrates the workflow for refining and knowledge mined from the literature using two machine reading systems[11,12]
- We use a PubMed query developed by health sciences librarian
  - AD-related literature published 2010 to 7/2021 from clinical studies
- We use a Knowledge Graph framework called PheKnowLator[13] developed by computational biologists to combine ontology-based resources after performing graph completion
- We search the KG for confounders, mediators, and colliders for the depression to AD relationship
- We translate standard epidemiological definitions of causal roles into SPARQL queries that identify potential covariates fulfilling the definitions for these variables

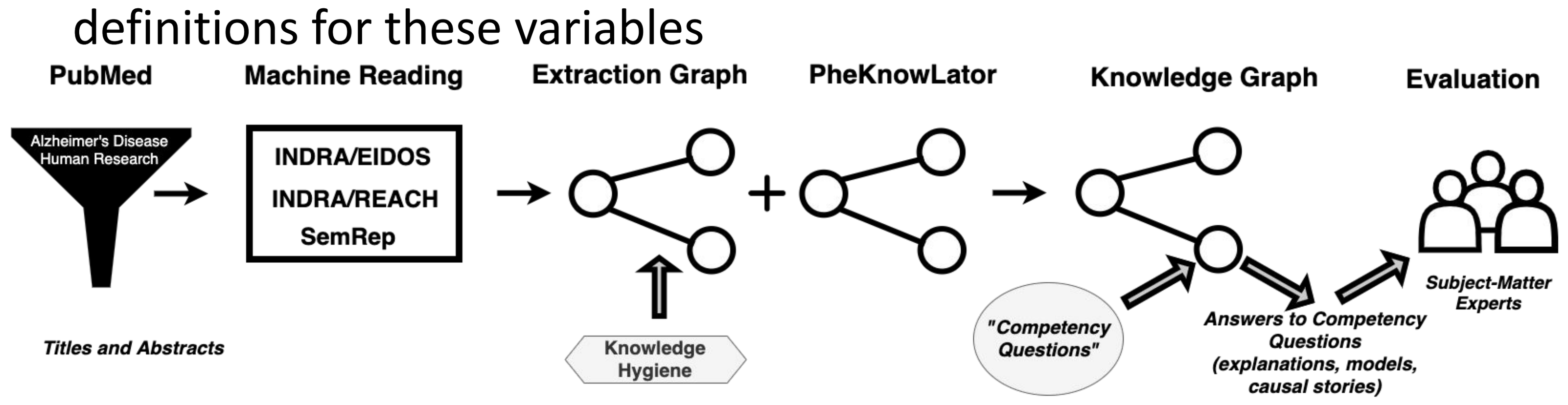

Figure B. Workflow of the study

Results

- A total of 13,365 PubMed-indexed articles were returned from PubMed from which 226,997 subject-predicate-object triples were extracted by the machine readers, including 10,020 unique UMLS concepts, and 2504 concepts were mapped to the merged ontologies in PheKnowLator
- Variable search methods identified 43 confounders, 16 mediators, and 23 colliders that were clinical phenotypes
- 31 of the phenotype variables were solely confounders, while the other 27 conditions also fulfilled the criteria for other roles, as per Table A
- Curiously, phenotypes relating to hypersensitivity to blood glucose levels, e.g., hypoglycemia and T2DM, were identified in all three categories of causal variables

Table A. Example confounders (common causes), mediators (intermediate causes), and colliders (common effects) identified by searching the knowledge graph.

Causal Role	Conditions
Confounder only	amyloidosis, atrial fibrillation, cerebral atrophy, chronic infections disease, COPD, encephalopathies, hypercholesterolemia, hyperglycemia, hyperinsulinism, hypertension, hypoglycemia, hypotension, inflammation, leukoencephalopathy, low tension glaucoma, migraines, myocardial infarc, non-alcoholic fatty liver disease, obesity, overweight, periodontal disease, Rickets, sleep apnea, vitamin D deficiency
Mediators only	anemia
Colliders only	apraxias, encephalitis, falls, frontotemporal dementia, immune response, neurofibrillary degeneration, pneumonia, psychotic disorders, senile plaques,
Confounders + Colliders	atrophic lateral sclerosis, congestive heart failure, deglutition disorders, tauopathies
All three roles	atherosclerosis, brain hemorrhage, cerebrovascular accident, diabetes mellitus, insulin resistance, ischemic stroke, malnutrition, obesity, Parkinsonian disorders

Discussion and Future Work

- The many identified confounders, mediators, and colliders confirm the complexity of third-factor variables
- Examples of potential confounders missed by the strategy include adverse childhood experiences, e.g., neglect by or death or mental illness of a parent[14]

Conclusion and Future Work

- The existence of problematic variables that fulfill multiple causal roles strongly suggests the value of a combined machine-human strategy
- The next steps include:
  - IRB-approved validation study using survey of AD experts
  - Using the KG-derived adjustment sets to answer hypothetical causal questions about AD from EHR-derived data
  - Comparing KG-derived adjustments sets with traditional data-driven feature selection methods

Selected References

1. VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat. 2013 Feb;41(1):196–220.
2. Shpitser I, VanderWeele T, Robins JM. On the validity of covariate adjustment for estimating causal effects. In 2010. p. 527–36.
3. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. Annu Rev Sociol [Internet]. 2014 Jul;40:31–53.
4. Cole SR et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol [Internet]. 2010 Apr;39(2):417–20.
5. Wang T et al. Sensitivity analysis for mistakenly adjusting for mediators in estimating total effect in observational studies. BMJ Open [Internet]. 2017 Nov [cited 2020 Jan 22];7(11):e015640.
6. Malec SA et al. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. medRxiv [Internet]. 2021 Jan 1;2020.07.08.20113035.
7. Malec SA et al. Literature-Based Discovery of Confounding in Observational Clinical Data. AMIA Annu Symp Proc AMIA Symp. 2016;2016:1920–9.
8. Malec SA, Boyce RD. Exploring Novel Computable Knowledge in Structured Drug Product Labels. AMIA Summits Transl Sci Proc [Internet]. 2020 May 30 [cited 2020 Jun 14];2020:403–12.
9. Ownby RL et al. Depression and Risk for Alzheimer Disease. Arch Gen Psychiatry. 2006 May;63(5):530–8.
10. Butters MA et al. Pathways linking late-life depression to persistent cognitive impairment and dementia. Dialogues Clin Neurosci. 2008 Sep;10(3):345–57.
11. Gyori BM et al. From word models to executable models of signaling networks using automated assembly. Mol Syst Biol [Internet]. 2017 Nov 24 [cited 2020 Aug 14];13(11).
12. Kilicoglu H. et a. Broad-coverage biomedical relation extraction with SemRep. BMC Bioinformatics. 2020 May 14;21(1):188.
13. Callahan TJ et al.Framework for Automated Construction of Heterogeneous Large-Scale Biomedical Knowledge Graphs. bioRxiv. 2020 May 2;2020.04.30.071407.
14. Tani Y, Fujiwara T, Kondo K. Association Between Adverse Childhood Experiences and Dementia in Older Japanese Adults. JAMA Netw Open. 2020 Feb 5;3(2):e1920740.