Evaluating early life risk factors for late life health outcome using probabilistic matching of early and late life cohorts.

Adina Zeki Al Hazzouri[1], Katrina Kezios[1], Scott Zimmerman[2], Sebastian Calonico[3], and M. Maria Glymour[2]

Author's affiliations:

[1]Department of Epidemiology, Mailman School of Public Health, Columbia University, NYC 10032

[2]Department of Epidemiology and Biostatistics, University of California, San Francisco, San Francisco, CA 94158

[3]Department of Health Policy and Management, Mailman School of Public Health, Columbia University, NYC 10032

Word count: 1901

Abstract:

Research on Alzheimer's Disease and Related Dementias (ADRD) is hampered by the absence of studies including prospective follow-up from early life through older ages when ADRD is diagnosed. This is a notable gap in the United States and impedes research on lifecourse determinants of ADRD and ADRD disparities, many of which appear attributable to early life experiences. In this simulation project, we evaluate a matching method to create a synthetic lifecourse cohort by merging early and late life cohorts on a set of harmonized covariates. We evaluate performance under several causal scenarios for the association between our exposure and outcome, and varying characteristics of the matching method. In scenarios when a measure is available along all pathways linking exposure and outcome, the synthetic cohort performs well, with bias approaching null as the number of matching levels increases. This approach may create novel opportunities to rigorously evaluate early- and mid-life determinants of ADRD and ADRD disparities.

Introduction:

Understanding and eliminating racial/ethnic and socioeconomic disparities in Alzheimer's disease and other dementias (ADRD) is a national priority.[1,2] Many important risk factors for dementia are thought to occur in early or mid- life, and understanding these determinants offers a window of opportunity to modify the course of the preclinical phase. Yet, evidence about the timing and influence of early and mid-life exposures on ADRD is limited because cohorts intended for the study of ADRD typically enroll older adults. Although some countries have birth cohort studies[3-6] that are now reaching later life when ADRD emerges, there is no such cohort available in the United States. However, a patchwork of studies have collected longitudinal information on individuals from early childhood through late adolescence or from early adulthood through midlife. Combining these studies affords opportunities to evaluate lifecourse trajectories of risk factors and could potentially help evaluate early- and mid-life risk factors for ADRD.

A small number of initiatives have pooled studies from across the lifecourse to create synthetic cohorts.[7,8] Theoretically, these cohorts can be used to describe the arc of changes in social or cardiovascular risk factors across the lifecourse. However, whether pooled cohorts can provide an unbiased estimate of causal effects of a variable measured in an early life cohort on an outcome measured in a late life cohort has not been systematically evaluated. Such an analysis requires not just pooling data but matching individual observations from the late life cohort to observations from the early life cohort. It is impossible to identify one-to-one matches, so these matches must be conducted on a probabilistic basis for observations within the same multivariate covariate stratum. We evaluate a method of creating a synthetic cohort by combining a prospectively followed early life cohort with a prospectively followed late life cohort, based on

probabilistic matching across a set of covariates shared between the two cohorts. We evaluate causal structures under which such a probabilistically matched pooled cohort can be used to estimate causal effects of risk factors on outcomes originally measured in different cohorts.

Methods:

While most cohorts do not overlap, they often share some variables in common which become the basis for the linking and matching process that we propose (i.e. matching of older subjects to their younger counterparts). We first created a simulated population with 7 variables: U [e.g., parental SES], V representing the predictor of interest [e.g. early life elevated blood pressure], W representing a shared variable measured in both cohorts [e.g. midlife life elevated blood pressure], W2 representing another shared variable [e.g. midlife elevated cholesterol], X [e.g. cerebrovascular disease], a modifier M [e.g. sex] and Y representing the outcome of interest [dementia]). We evaluated several causal scenarios for the relations between these variables, represented as Directed Acyclic Graphs (Figure 1). From a population size of 30,000, we then sampled a hypothetical early life cohort (N=2,500), in which only U, V, W, W2 and M were measured, and a late life cohort (N=2,500), in which only W, W2, M, X, and Y were measured. We categorized values of the variables measured in both the early and late life cohorts (i.e. shared variables) and created a *many-to-many* match where each observation in the late life cohort for which W=w was matched with each observation in the early life cohort for which W=w. In other scenarios, the *many-to-many match* of the early and late life cohorts was done using not only values of W but also values of W2 and M, all of which are shared between the two cohorts. We then evaluated the estimated effect of V on Y (the relationship of interest) in the synthetic merged cohort and compared this to the effect of V on Y that would be estimated in the

original simulated population if all variables were measured on the same individuals, using linear regression models corrected for repeated measures on the same observation. We considered performance under alternative causal structures in the simulated population which we believe are common structures for the relationship between V (early life BP) and Y (dementia). We do not anticipate that the synthetic cohort will out-perform the causal inferences that would be possible with the original full data set, so our benchmark for comparison of the probabilistically matched synthetic cohort is the effect estimate derived from the full population. In other words, if the causal effect would not be identifiable in the full population, we do not expect it to be identifiable in the synthetic cohort, but we evaluate whether the synthetic cohort delivers the same (potentially biased) effect estimate as the full population.

We examined the estimated effect of V on Y under causal structures represented in the DAGs in Figure 1. For each DAG, we conducted several scenarios (A, B and C) changing the number of categories of the matching/linking variables W, W_2, and M. We anticipated that more finely grained categorization of the matching variables would improve performance of the simulation, so we considered categorizing the matching variables into 3 values, 10 values, or 30 values. In DAG1 or simulation 1, the causal structure imposes a relationship between V and Y via W. In Simulation 2.1, the causal structure imposes a relationship between V and Y that includes a direct path and an indirect path via W. Simulation 2.2. is like simulation 2.1 but instead of a direct path from V to Y, this path is now indirect and includes W2. Simulations 2.3 and 2.4 are similar to simulation 2.1 but include a binary effect modifier M of the V-W path. Simulation 2.5 is similar to simulation 2.3/2.4 but the V-Y path is indirect and includes W2. Simulation 3.1 imposes a relationship between V and Y since it is confounded by U (i.e. V and Y are connected

through a backdoor path), but when adjusting for U in simulation 3.2 then we expect no relationship between V and Y. In simulations 4.1 and 4.2, the causal structure suggests no causal effect of V on Y but a potential association induced by survival (i.e. collider bias) because analyses are restricted to individuals who survived into older age.

For each causal scenario and instance, we ran a 1000 replications. We then calculated the bias in the estimated effect of V on Y as the difference in the simulated/true versus synthetic/borrowed coefficient for V; we evaluated both empirical standard error (i.e., the standard deviation of the estimated effects across simulations) and used the standard error calculated within each simulation to estimate statistical power. All simulations were done using STATA 16/MP.

Results:

Simulation 1: In a simple setting with a causal sequence from V to Y, we find that the synthetic cohort performs well (Table 1). Bias is small when the matching variable has only 3 categories on which to link the early-life and late-life cohorts, and becomes even smaller when the number of matching categories is increased to 30. Increasing the number of categories does not reduce the standard deviation of the bias.

Simulation 2: In the setting in which V influences Y via a mechanism unrelated to the matching variable (i.e. there is a direct path from V to Y not via W), substantial bias in the estimated effect occurs in the synthetic cohort (Simulation 2.1). When another linking variable, W2, was added to the direct path from V to Y, the bias was reduced but not eliminated (Simulation 2.2). Again, increasing the number of matching categories reduces bias but does not impact the standard

deviation of the bias in either Simulation 2.1 or 2.2. In the context of an effect modifier of the V to W path when the modifier was not available as a matching variable (Simulation 2.3), the bias is again substantial, and the bias remains substantial even when matching on both W and M (Simulation 2.4). In both Simulations 2.3 and 2.4, there is a direct effect of the exposure on the outcome via a path not captured in the matching variables. In the setting where another linking variable, W2, is introduced on the direct path from V to Y, and then matching on the combination of W, W2 and M, the bias becomes small (Simulation 2.5).

Simulation 3: in the setting where U is a confounder of the V to Y relationship of interest, the synthetic cohort delivers effect estimates very similar to those based in the full population both without (Simulation 3.1) and with (Simulation 3.2) adjustment for U, and the effect estimates are closer to the null when adjusting for U (Simulation 3.2).

Simulation 4: Simulation 4 explores settings where there is a variable influenced by both V and other causes of Y. In simulation 4.1, analyses are conditioned on a S (e.g., restricted to individuals who survive so S=1), the estimated association is spurious but the bias compared to estimation in the full population is negligible when the matching variable has many categories (Simulation 4.1). When we do not condition on S, the estimated association between V and Y is close to the null in both the true population and the synthetic cohort (Simulation 4.2).

Discussion:

In this work, we used simulation mirroring a range of possible causal structures to establish the feasibility of evaluating early life risk factors for late life outcomes by creating a synthetic cohort

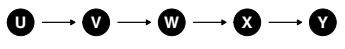
pooling early and late life studies that share some but not all variables. Using the shared variables as a basis for a many-to-many match, we were able to recover the effect estimates from the full population using the synthetic cohort in many but not all causal scenarios.

Overall, our findings suggest that in the causal scenarios we propose, i.e. matching older and younger cohorts based on shared variables, the bias comparing the true/simulated estimate versus the synthetic cohort estimate was largest in scenarios that included a direct effect of V on Y without an intermediate variable to match on (i.e., W2). When every path from V to Y (the relation of interest) included an intermediate variable to match on and link the young and old cohorts, then the bias was relatively small and in some cases close to zero. We also found bias to be relatively small in common causal structures such as in the presence of confounding or in the presence of collider bias due to conditioning on survivors. Most of the settings we evaluated included an effect of V on Y, but in the settings of no effect of V on Y, the simulations also performed well.

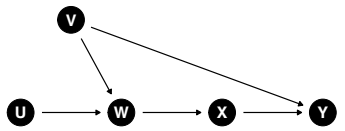
Important limitations include the fact that we would not expect our approach to outperform results in the full population. Therefore, all of the conventional assumptions for causal inference must hold (i.e., exchangeability, consistency, positivity). We explored only a limited number of causal settings and other causal structures should be considered, as well as other estimands, e.g., direct effects. We found matching to be computationally time-consuming when there were numerous matching variables with multiple possible values.

Evaluating early and mid-life determinants of ADRD to identify risk and protective factors is a high priority for reducing the population burden of ADRD in the future.[1] This research is often stymied by the limited follow-up or lack of detailed early life measures available in ADRD cohorts. The synthetic cohort approach we propose affords opportunities to evaluate lifecourse risk factor effects on late life outcomes. Key domains for future research include evaluating additional causal scenarios, causal estimands, and characteristics of linking variables.

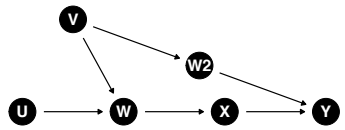
Figure 1. Directed Acyclic Graphs representing alternative causal structures considered. We adopt the informal convention of showing an arrow into an arrow to emphasize effect modification. Formally, the modifying arrows point to the same node as the modified arrows, e.g., in simulation 2.4, M influences W.



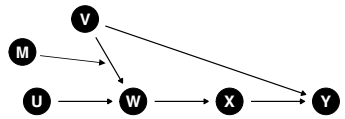
DAG 1: matching on W



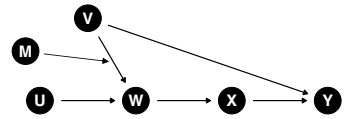
DAG 2.1: matching on W



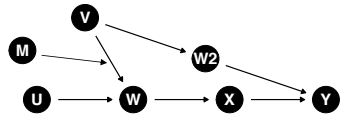
DAG 2.2: matching on W and W2



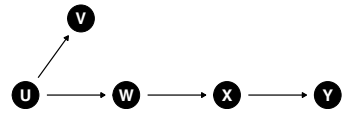
DAG 2.3: matching on W



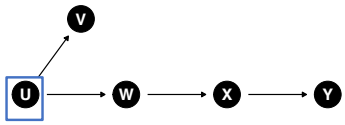
DAG 2.4: matching on W and M



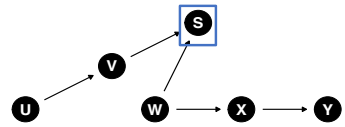
DAG 2.5: matching on W, W2, M



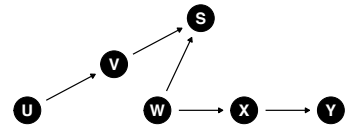
DAG 3.1: matching on W



DAG 3.2: matching on W, conditioning on U



DAG 4.1: matching on W, conditioning on S ("selection")



DAG 4.2: matching on W

Table 1. Simulation results pertaining to each causal structures considered, as depicted in Figure 1.

Simulation scenario		Number of categories for matching variable(s)	Estimated effect of V on Y in the full population (i.e. Truth)		Estimated effect of V on Y in the merged synthetic population		BIAS: Difference in estimate in full population versus merged synthetic population	
			$\bar{B}_{full\ population}$	(SE)	$\bar{B}_{synthetic\ population}$	(SE)	$b_{mean\ difference}$	(SE)
Simulation 1	Scenario A	3	.342	(.016)	.272	(.012)	-.070	(.012)
	Scenario B	10	.343	(.016)	.329	(.013)	-.014	(.011)
	Scenario C	30	.343	(.016)	.340	(.013)	-.003	(.011)
Simulation 2.1	Scenario A	3	1.042	(.021)	.456	(.018)	-.586	(.021)
	Scenario B	10	1.043	(.021)	.553	(.021)	-.4901	(.021)
	Scenario C	30	1.044	(.021)	.571	(.023)	-.473	(.020)
Simulation 2.2	Scenario A	3	.904	(.024)	.774	(.022)	-.131	(.019)
	Scenario B	10	.954	(.028)	.926	(.026)	-.028	(.020)
	Scenario C	30	.978	(.032)	.969	(.031)	-.009	(.022)
Simulation 2.3	Scenario A	3	.921	(.021)	.254	(.015)	-.667	(.023)
	Scenario B	10	.921	(.021)	.316	(.018)	-.605	(.022)
	Scenario C	30	.920	(.023)	.329	(.021)	-.591	(.022)
Simulation 2.4	Scenario A	3	.924	(.021)	.281	(.017)	-.643	(.023)
	Scenario B	10	.932	(.022)	.355	(.021)	-.577	(.022)
	Scenario C	30	.936	(.023)	.376	(.023)	-.560	(.022)
Simulation 2.5	Scenario A	3	.780	(.025)	.657	(.022)	-.122	(.019)
	Scenario B	10	.836	(.031)	.810	(.028)	-.026	(.020)
	Scenario C	30	.867	(.038)	.859	(.037)	-.008	(.026)
Simulation 3.1	Scenario A	3	.247	(.022)	.194	(.011)	-.052	(.020)
	Scenario B	10	.245	(.020)	.235	(.012)	-.011	(.018)
	Scenario C	30	.245	(.020)	.243	(.013)	-.002	(.018)
Simulation 3.2	Scenario A	3	.001	(.028)	.000	(.010)	-.001	(.030)
	Scenario B	10	.000	(.028)	-.000	(.010)	-.000	(.029)
	Scenario C	30	.000	(.027)	.000	(.010)	.000	(.029)
Simulation 4.1	Scenario A	3	-.063	(.023)	-.053	(.010)	.010	(.025)
	Scenario B	10	-.064	(.022)	-.063	(.011)	.000	(.024)
	Scenario C	30	-.064	(.023)	-.063	(.012)	.001	(.024)
Simulation 4.2	Scenario A	3	.001	(.017)	-.000	(.007)	-.001	(.019)
	Scenario B	10	.000	(.017)	-.000	(.008)	-.000	(.018)
	Scenario C	30	.001	(.017)	-.000	(.008)	-.001	(.018)

Data shows empirical standard error which is the standard deviation of the estimates across the 1000 iterations of each respective simulation.

Data availability statement: the data is simulated and thus data sharing is not applicable.

References:

1. *National Plan to Address Alzheimer's Disease: 2019 Update. Retrieved from https://aspe.hhs.gov/report/national-plan-address-alzheimers-disease-2019-update.* 2015.
2. Montine TJ, Koroshetz WJ, Babcock D, et al. Recommendations of the Alzheimer's disease-related dementias conference. *Neurology.* 2014;83(9):851-860.
3. Elliott J, Shepherd P. Cohort profile: 1970 British Birth Cohort (BCS70). *Int J Epidemiol.* 2006;35(4):836-843.
4. Wadsworth M, Kuh D, Richards M, Hardy R. Cohort Profile: The 1946 National Birth Cohort (MRC National Survey of Health and Development). *Int J Epidemiol.* 2006;35(1):49-54.
5. Power C, Elliott J. Cohort profile: 1958 British birth cohort (National Child Development Study). *Int J Epidemiol.* 2006;35(1):34-41.
6. Lumey LH, Ravelli AC, Wiessing LG, Koppe JG, Treffers PE, Stein ZA. The Dutch famine birth cohort study: design, validation of exposure, and selected characteristics of subjects after 43 years follow-up. *Paediatr Perinat Epidemiol.* 1993;7(4):354-367.
7. Levine DA, Gross AL, Briceno EM, et al. Association Between Blood Pressure and Later-Life Cognition Among Black and White Individuals. *JAMA Neurol.* 2020.
8. Zeki Al Hazzouri A, Vittinghoff E, Zhang Y, et al. Use of a pooled cohort to impute cardiovascular disease risk factors across the adult life course. *Int J Epidemiol.* 2019;48(3):1004-1013.