PLOS ONE

RESEARCH ARTICLE

Addressing overlapping sample challenges in genome-wide association studies: Meta-reductive approach

Farid Rajabli[1,2]*, **Azra Emekci**[3]

1 John P. Hussman Institute for Human Genomics, University of Miami Miller School of Medicine, Miami, FL, United States of America, **2** Dr. John T Macdonald Foundation Department of Human Genetics, University of Miami Miller School of Medicine, Miami, FL, United States of America, **3** Pioneer High School, San Jose, CA, United States of America

* Fxr213@med.miami.edu

Citation: Rajabli F, Emekci A (2024) Addressing overlapping sample challenges in genome-wide association studies: Meta-reductive approach. PLoS ONE 19(8): e0296207. https://doi.org/10.1371/journal.pone.0296207

Editor: Maria Ines Fariello Rico, Universidad de la Republica Uruguay: Facultad de Ingeniería, URUGUAY

Received: December 6, 2023

Accepted: June 10, 2024

Published: August 1, 2024

Data Availability Statement: The data used for real data validation in this study are publicly available and have been properly cited in the paper (doi: 10.1038/s41588-019-0358-2, doi: 10.1001/jamaneurol.2020.3536, 10.1038/s41588-022-01024-z, doi: 10.1016/j.jalz.2019.06.4950). Data are available through the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) Data Sharing Service (DSS) under accession numbers ng00100 and ng00075 (https://dss.niagads.org/datasets/), and through the European Bioinformatics Institute (EBI)

Abstract

Polygenic risk scores (PRS) are instrumental in genetics, offering insights into an individual level genetic risk to a range of diseases based on accumulated genetic variations. These scores rely on Genome-Wide Association Studies (GWAS). However, precision in PRS is often challenged by the requirement of extensive sample sizes and the potential for overlapping datasets that can inflate PRS calculations. In this study, we present a novel methodology, Meta-Reductive Approach (MRA), that was derived algebraically to adjust GWAS results, aiming to neutralize the influence of select cohorts. Our approach recalibrates summary statistics using algebraic derivations. Validating our technique with datasets from Alzheimer disease studies, we showed that the summary statistics of the MRA and those derived from individual-level data yielded the exact same values. This innovative method offers a promising avenue for enhancing the accuracy of PRS, especially when derived from meta-analyzed GWAS data.

Introduction

Polygenic risk scores (PRS) have emerged as an essential tool in the field of genetics [1, 2]. These scores offer a unique insight into an individual's genetic predisposition to a wide array of diseases and traits, capturing the cumulative effects of multiple genetic variants [3]. The Genome-Wide Association Studies (GWAS) serve as the base for creating PRS [4]. GWAS investigates the entire genetic makeup of individuals to identify genetic variations associated with specific diseases or traits. The predictive accuracy and precision of PRS are enhanced when the base GWAS summary statistics come from a sizeable sample, and the population in the GWAS matches the population where the PRS is being applied [4, 5]. Due to this need for a substantial sample size, studies often aim to meta-analyze all available genetic datasets to achieve the statistical power necessary for identifying genetic markers linked to the trait or disease. However, this approach presents a challenge in securing independent datasets for training, testing, and validating PRS performance [6]. The use of overlapping samples can inflate the PRS calculations, resulting in imprecise risk predictions.

GWAS Catalog DSS under accession numbers GCST90027158 and GCST009019 (https://www.ebi.ac.uk/gwas/). Summary statistics for APOE4 allele validation across populations are available in S1 Table. The simulation script, including the MRA function, is accessible at: https://github.com/hihg-um/MRA.

Funding: National Institute on Aging, AG070864, Dr. Farid Rajabli; BrightFocus Foundation, A2018556F, Dr. Farid Rajabli.

Competing interests: The author declares no competing interests.

A logical approach might be to exclude a specific cohort of interest and then rerun meta-analyses with the remaining datasets. However, given the significant computational resources needed and the difficulties in accessing detailed summary statistics for all cohorts, this isn't always viable. Nonetheless, we do have access to the cohort-level data for the specific dataset we aim to employ as a training and testing set. Recognizing this advantage, we formulated an alternative technique that incorporates the cohort-level result of our chosen dataset along with the meta-analysis GWAS findings. The goal is to neutralize the impact of the overlapping cohort of interest on the meta-analysis GWAS summary statistics, thus producing a PRS that avoids the inflationary tendencies arising from overlapping samples.

In this study, we derived equations to adjust GWAS results, effectively eliminating the impact of selected cohorts in inverse variance-based fixed effect meta-analysis (FEMA) studies. Through comprehensive simulations and real data analysis, we demonstrated that our methodology effectively updates the base data's summary statistics, thereby addressing the challenge.

Materials and methods

Derivation of adjusted summary statistics: Meta-Reductive Approach (MRA)

We analyzed two distinct sets of summary statistics:

1. A compilation from n datasets meta-analyzed using an inverse variance-based approach [7].
2. A specific dataset of interest that was also part of the meta-analysis.

For these datasets:

- B and SE symbolize the effect size and standard error, respectively, from the aggregate meta-analysis across n datasets.
- β_i and se_i specify the effect size and standard error for the individual cohort i.

Our primary aim was to compute a summary statistic that eliminates the influence of the dataset of interest, providing a clearer perspective on the overarching genetic structure.

i. **Inverse-variance-weighted effect-size estimation**. The inverse variance method gives more weight to studies with smaller variance because they offer more precise estimates. The weight, w_i, is the inverse of the variance, or squared standard error, of the effect size, β_i. Given,

$B = \frac{\sum_i^n \beta_i w_i}{\sum_i^n w_i}$ where the $w_i = \frac{1}{se_i^2}$

Expanding this:

$$Bw_1 + Bw_2 + Bw_3 + \cdots + Bw_{n-1} + Bw_n = \beta_1 w_1 + \beta_2 w_2 + \beta_3 w_3 + \cdots + \beta_{n-1} w_{n-1} + \beta_n w_n$$

This is the weighted sum of the effect sizes across all datasets, including the one of interest. Now, to remove the effect of the specific dataset, β_n, we rearrange:

$$Bw_1 + Bw_2 + Bw_3 + \cdots + Bw_{n-1} + Bw_n - \beta_n w_n = \beta_1 w_1 + \beta_2 w_2 + \beta_3 w_3 + \cdots + \beta_{n-1} w_{n-1}$$

Which yields:

$$B + \frac{Bw_n - \beta_n w_n}{w_1 + w_2 + w_3 + \cdots + w_{n-1}} = \frac{\beta_1 w_1 + \beta_2 w_2 + \beta_3 w_3 + \cdots + \beta_{n-1} w_{n-1}}{w_1 + w_2 + w_3 + \cdots + w_{n-1}}$$

This equation essentially adjusts the overall effect size, B, by subtracting the influence of the dataset of interest.

ii. **Standard error derivation**. The standard error (SE) offers a measure of the statistical accuracy of an estimate. Here, we adjust the SE based on the weights of all datasets excluding the one of interest.
Using:

$$SE^2 = \frac{1}{w_1 + w_2 + w_3 + \cdots + w_{n-1} + w_n}$$

We derive:

$$w_1 + w_2 + w_3 + \cdots + w_{n-1} = \frac{1 - SE^2 w_n}{SE^2}$$

This equation gives the combined weight of all datasets, excluding the dataset of interest.

iii. **Adjusted effect size and standard error**. Post removing the influence of the dataset of interest, the modified effect size is given by:

$$B_{adj} = \frac{\beta_1 w_1 + \beta_2 w_2 + \beta_3 w_3 + \cdots + \beta_{n-1} w_{n-1}}{w_1 + w_2 + w_3 + \cdots + w_{n-1}} = B + \frac{SE^2(Bw_n - \beta_n w_n)}{1 - SE^2 w_n}$$

This adjusted beta, B_{adj}, having nullified the contribution of the specific dataset n. Additionally, the adjusted standard error is:

$$SE^2_{adj} = \frac{SE^2}{1 - SE^2 w_n}$$

This adjustment ensures that the standard error reflects the precision of our new effect size estimate, free from the influence of the specific dataset.

Ethical approval was not required for this study as it utilized publicly available summary statistics.

Results

Validation using real data

To validate our methodological approach, we utilized summary statistics from four publicly accessible Alzheimer disease studies: Kunkle et al. [8], Kunkle et al. [9] AA, Bellinguez et al. [10], and Moreno-Grau S. et al. [11] From these studies, 100,000 markers were selected to conduct a meta-analysis using the METASOFT software [12].

Following the initial meta-analysis, we applied a systematic "leave-one-out" strategy. For each iteration, we excluded the summary statistics from one dataset and conducted a meta-

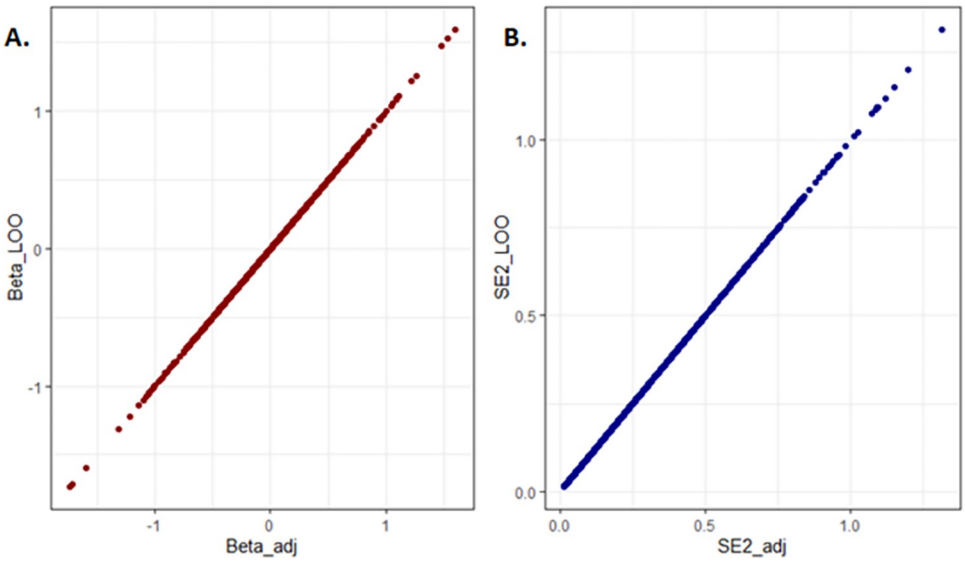

Fig 1. Comparison between the adjusted results from the Meta-Reductive Analysis (MRA) (Beta_adj and SE2_adj) approach and the "leave-one-out" inverse variance-based fixed effect meta-analysis (FEMA) (Beta_LOO and SE2_LOO). The MRA-adjusted values show identical results with the FEMA calculation for both Beta values (A) and Standard Error (B).

https://doi.org/10.1371/journal.pone.0296207.g001

analysis of the remaining three. The results from this procedure served as our individual-level data for the three datasets in question.

For the final step of validation, we calculated the adjusted B_{adj} and SE^2_{adj} values based on MRA and compared them against the individual-level data summary statistics derived from the "leave-one-out" FEMA. Our results showed that the summary statistics of the FEMA and MRA approaches yielded the exact same values. To demonstrate this, we plotted the betas and standard errors (Fig 1). The graphical representation illustrates that both beta and standard error values from the "leave-one-out""FEMA and MRA give the same results.

Additionally, we conducted a validation analysis for the APOE4 allele, utilizing data from a multi-ancestry study by Rajabli et al. [13], which included four population-based studies: African American, East Asian, Hispanic, and non-Hispanic Whites (S1 Table). We applied a "leave-one-out" strategy by removing one population at a time and performing the validation on the remaining studies. We followed the same steps as described previously, and the results were exact same across all tests, as detailed in Table 1.

Simulation

We simulated Beta coefficients and their corresponding SEs across ten studies, each containing 10,000 markers (using R programming language.) We generated random Beta coefficients

Table 1. Validation analysis for APOE4 allele using summary statistics from Rajabli et al. study. "MRA-Beta" and "MRA-SE" denote beta and standard error values derived using the MRA approach, respectively. "Traditional-Beta" and "Traditional-SE" refer to beta and standard error values obtained from the meta-analysis of three studies.

Study removed	MRA-beta	MRA-SE	Traditional-Beta	Traditional-SE
African American	1.189192599	0.01671915	1.189192599	0.01671915
East Asian	1.09385844	0.017205134	1.09385844	0.017205134
Hispanic	1.215580078	0.016903311	1.215580078	0.016903311
non-Hispanic White	1.177899334	0.026321213	1.177899334	0.026321213

https://doi.org/10.1371/journal.pone.0296207.t001

utilizing the "**rnorm"** function, under the assumption of a normal distribution, characterized by a mean of zero and a standard deviation of one. We used "**runif"** function to produce random SEs values from a uniform distribution, with specified minimum and maximum limits of 0.1 and 0.5, respectively. The simulated Betas and SEs for each marker within a study were then organized into dedicated columns within a data frame.

We applied the "**rma"** function from the **metaphor** [14] package to facilitate a fixed-effects meta-analysis on the generated Betas and SEs. Then we implemented "leave-one-out" strategy, mirroring the methodology applied to real data. We calculated the adjusted B_{adj} and SE^2_{adj} values employing our proposed method and compared it with individual-level data derived from the "leave-one-out" meta-analyses. The outcomes revealed that the summary statistics were identical, similar to the findings from real data analysis. The simulation script is provided with the MRA function here: https://github.com/hihg-um/MRA.

Discussion

This study employs algebraic adjustments to GWAS summary statistics to eliminate the influence of specific datasets in meta-analyses. The algebraic solutions applied to real and simulated data consistently matched our expectations of achieving identical results. The validation confirms the robustness and reliability of derived equations, emphasizing the effectiveness of our methods in addressing the challenges associated with sample overlap in meta-analyses.

Furthermore, our approach utilizes the widely recognized inverse-variance method for fixed-effect meta-analysis. This choice ensures that our adjustments are based on a widely accepted framework, enhancing the general applicability and relevance of our findings. While our study focuses on inverse-variance method fixed-effects models, the foundational principles of our approach could potentially be adapted for random-effects meta-analyses, which would be useful in situations where variability between studies is significant.

In summary, our research highlights the practicality of excluding specific datasets to refine effect estimates in inverse-variance method meta-analysis. We provide a method that enables researchers to neutralize the impact of overlapping cohorts on meta-analysis GWAS summary statistics, thereby producing a PRS that avoids the possible inflations associated with overlapping samples. This approach is important for enhancing the accuracy and reliability of PRS in genetic studies.

Supporting information

S1 Table. Summary statistics for APOE e4 allele association with Alzheimer disease across populations.
(XLSX)

Author Contributions

Conceptualization: Farid Rajabli.

Data curation: Farid Rajabli.

Formal analysis: Farid Rajabli, Azra Emekci.

Funding acquisition: Farid Rajabli.

Investigation: Farid Rajabli.

Methodology: Farid Rajabli.

Resources: Farid Rajabli.

Software: Farid Rajabli, Azra Emekci.

Supervision: Farid Rajabli.

Validation: Farid Rajabli, Azra Emekci.

Visualization: Farid Rajabli.

Writing – original draft: Farid Rajabli.

Writing – review & editing: Farid Rajabli.

References

1. Padilla-Martinez F., Collin F., Kwasniewski M., and Kretowski A. (2020). Systematic Review of Polygenic Risk Scores for Type 1 and Type 2 Diabetes. Int. J. Mol. Sci. 5, 10.3390/ijms21051703. https://doi.org/10.3390/ijms21051703 PMID: 32131491
2. Harrison J.R., Mistry S., Muskett N., and Escott-Price V. (2020). From Polygenic Scores to Precision Medicine in Alzheimer's Disease: A Systematic Review. Journal of Alzheimer's disease 4, 1271–1283. https://doi.org/10.3233/JAD-191233 PMID: 32250305
3. Gallagher S., Hughes E., Wagner S., Tshiaba P., Rosenthal E., Roa B.B., et al. (2020). Association of a Polygenic Risk Score With Breast Cancer Among Women Carriers of High- and Moderate-Risk Breast Cancer Genes. JAMA Netw. Open 7, e208501. https://doi.org/10.1001/jamanetworkopen.2020.8501 PMID: 32609350
4. Choi S.W., Mak T.S., and O'Reilly P.F. (2020). Tutorial: a guide to performing polygenic risk score analyses. Nat. Protoc. 9, 2759–2772. https://doi.org/10.1038/s41596-020-0353-1 PMID: 32709988
5. Anonymous Genome-wide association studies | Nature Reviews Methods Primers.
6. Choi S.W., Mak T.S.H., Hoggart C.J., and O'Reilly P.F. (2022). EraSOR: a software tool to eliminate inflation caused by sample overlap in polygenic score analyses. Gigascience. https://doi.org/10.1093/gigascience/giad043 PMID: 37326441
7. Willer C.J., Li Y., and Abecasis G.R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 17, 2190–2191. https://doi.org/10.1093/bioinformatics/btq340 PMID: 20616382
8. Kunkle B.W., Grenier-Boley B., Sims R., Bis J.C., Damotte V., Naj A.C., et al. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates A beta, tau, immunity and lipid processing. Nature genetics 3, 414–+.
9. Kunkle B.W., Schmidt M., Klein H.U., Naj A.C., Hamilton-Nelson K.L., Larson E.B., et al. (2021). Novel Alzheimer Disease Risk Loci and Pathways in African American Individuals Using the African Genome Resources Panel: A Meta-analysis. JAMA Neurol. 1, 102–113.
10. Bellenguez C., Kucukali F., Jansen I.E., Kleineidam L., Moreno-Grau S., Amin N., et al. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. Nat. Genet. 4, 412–436. https://doi.org/10.1038/s41588-022-01024-z PMID: 35379992
11. Moreno-Grau S., de Rojas I., Hernandez I., Quintela I., Montrreal L., Alegret M., et al. (2019). Genome-wide association analysis of dementia and its clinical endophenotypes reveal novel loci associated with Alzheimer's disease and three causality networks: The GR@ACE project. Alzheimers Dement. 10, 1333–1347. https://doi.org/10.1016/j.jalz.2019.06.4950 PMID: 31473137
12. Han B., and Eskin E. (2011). Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-wide Association Studies. American journal of human genetics 5, 586–598. https://doi.org/10.1016/j.ajhg.2011.04.014 PMID: 21565292
13. Rajabli F, Benchek P, Tosto G, Kushch N, Sha J, Bazemore K, et al. Multi-ancestry genome-wide meta-analysis of 56,241 individuals identifies *LRRC4C*, *LHX5-AS1* and nominates ancestry-specific loci *PTPRK*, *GRB14*, and *KIAA0825* as novel risk loci for Alzheimer's disease: the Alzheimer's Disease Genetics Consortium. medRxiv [Preprint]. 2023 Jul 8:2023.07.06.23292311. https://doi.org/10.1101/2023.07.06.23292311 PMID: 37461624; PMCID: PMC10350126.
14. Viechtbauer W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, 36(3), 1–48. https://doi.org/10.18637/jss.v036.i03.