Quantifying Representation Bias In Alzheimer's Disease Datasets: A Cross-dataset Risk Factor Analysis

Mansi Singhal[1], Megan Delehanty[2], Ali Karimi[3], Leticia Rittner[4], Mariana Bento[1]

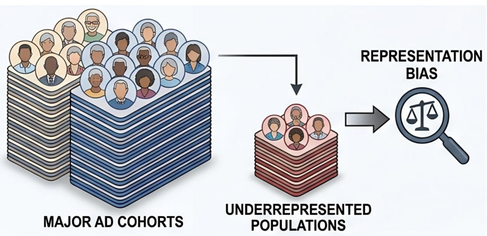
[1]Department of Biomedical Engineering, University of Calgary, Calgary, AB, Canada, [2]Department of Philosophy, University of Calgary, Calgary, AB, Canada, [3]Department of Communication, University of Calgary, Calgary, AB, Canada, [4]School of Electrical and Computer Engineering, University of Campinas, Campinas, Brazil.

KEY TAKEAWAYS

- Analyzed **10 Alzheimer's disease (AD) datasets** using a novel bias score (−100% to +100%) across **19 risk factors.**
- Bias ranged from −66% (PPMI) to +70% (MIRIAD), with most cohorts showing strong under-representation (DLBS 100%, ADNI 83%, DBPC 75%) and MIRIAD showing over-representation (50%).
- Non-modifiable and modifiable factors were consistently under-represented, and no dataset was balanced. Biased data produce inequitable AI predictions, reducing clinical utility.

INTRODUCTION

- **The AI Promise in AD:** Machine learning offers powerful tools for early Alzheimer's disease (AD) detection and personalized prognostic tracking.
- **Representation bias:** Most models rely heavily on homogeneous datasets (e.g., ADNI), leading to a stark drop in diagnostic accuracy for minority populations, which undermines the generalizability of AI models for AD.
- **Demographic Disparities:** Current biomedical datasets fail to reflect the true epidemiological diversity of AD, embedding historical healthcare inequities into algorithmic predictions.
- **The Risk Factor Gap:** While demographic gaps are widely acknowledged, existing AI pipelines lack a unified framework to balance non-modifiable (genetic) and modifiable (lifestyle/socioeconomic) risk factors across diverse cohorts.



This study introduces a novel bias score to measure the over- and under-representation of **19 risk factors** across **10 major AD cohorts**.

- Dataset summaries revealed substantial heterogeneity, with average bias ranging from **−66% (PPMI)** to **+70% (MIRIAD)**.
- Most cohorts showed strong negative bias (under-representation):
- **DLBS:** 100% under-representation
- **ADNI:** 83% under-representation
- **DBPC:** 75% under-representation
- **MIRIAD** demonstrated a unique positive bias (50% over-representation).
- **Non-modifiable factors** (age, sex, genetics) – raising fairness concerns
- **Modifiable factors** (lifestyle, comorbidities, pollution) – often below −70% bias
- **No single dataset** achieved balanced representation across all factors.

Table 1. Dataset characteristics and bias summary across 10 AD cohorts: sample size, risk factors analyzed (n), and % over/under-represented.

Dataset	Sample Size	Risk Factors (n)	Over-represented (%)	Under-represented (%)
ADNI (Alzheimer's Disease Neuroimaging Initiative)	2100	12	11%	83%
BrainLat (Latin American Brain Health Institute)	638	6	33%	67%
DBPC (DementiaBank Pitt Corpus)	1100	8	12%	75%
DLBS (Dallas Lifespan Brain Study)	950	3	0%	100%
HABS (Harvard Aging Brain Study)	1800	10	20%	80%
MIRIAD (Minimal Interval Resonance Imaging in Alzheimer's Disease)	460	4	50%	50%
NACC (National Alzheimer's Coordinating Center)	45000	16	25%	75%
OASIS (Open Access Series of Imaging Studies)	1098	9	25%	75%
PPMI (Parkinson's Progression Markers Initiative)	1700	14	29%	71%
SEA-AD (Seattle Alzheimer's Disease Brain Cell Atlas)	310	5	40%	60%

RESULTS

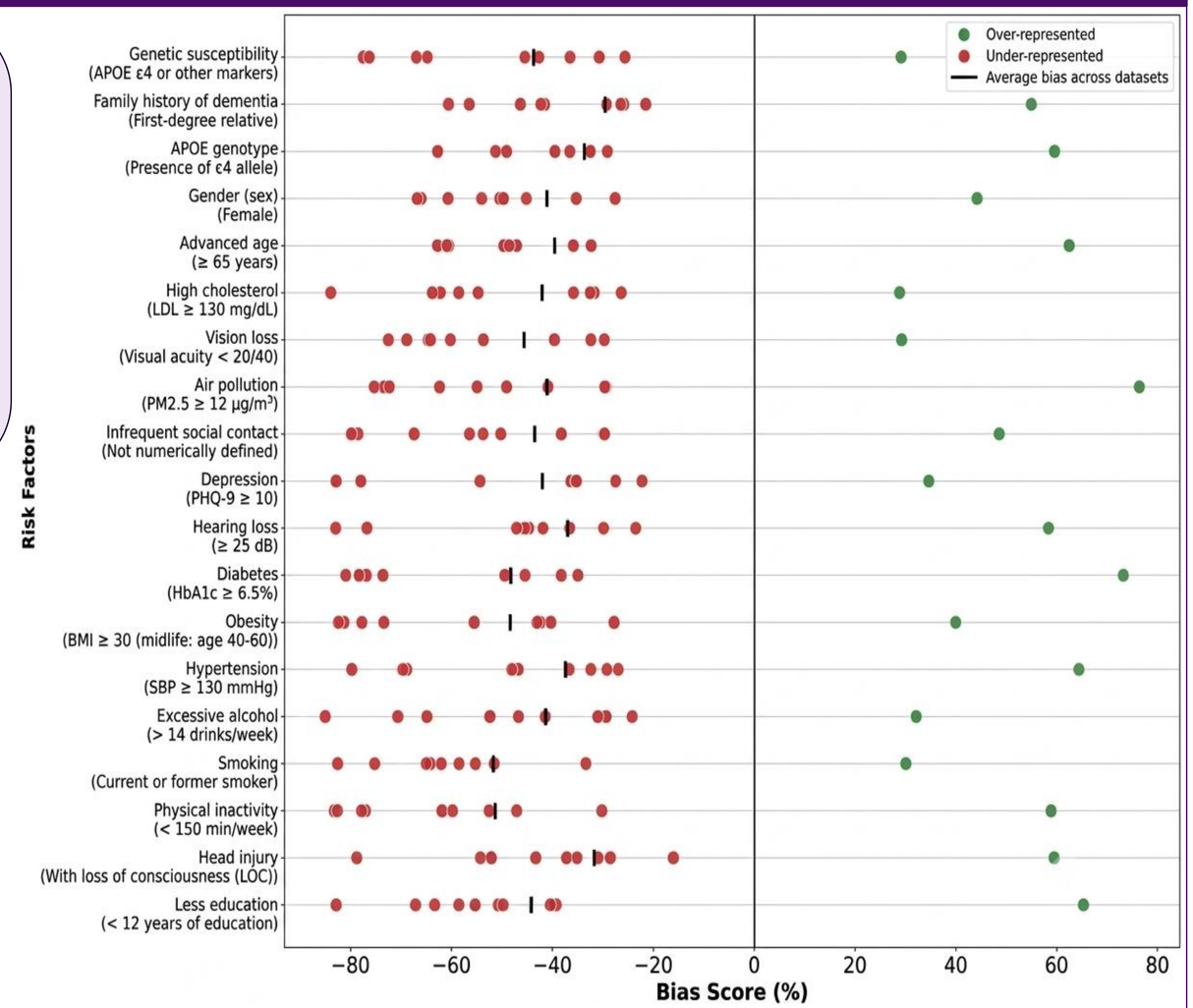
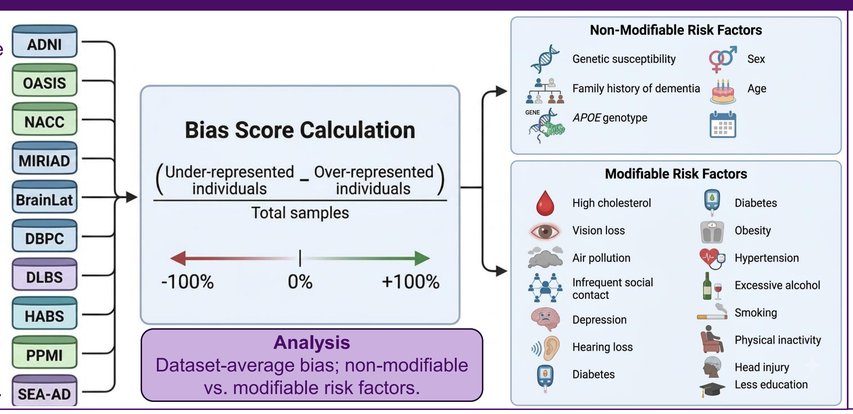


Figure 1 displays bias scores per dataset across 20 risk factors, with red indicating under-representation (negative bias), green indicating over-representation (positive bias), black showing the average bias per factor, and the dashed line representing the zero-bias reference.

METHODS

- **Dataset Selection:** Analyzed ten diverse Alzheimer's disease datasets: ADNI, OASIS, NACC, MIRIAD, BrainLat, DBPC, DLBS, HABS, PPMI, and SEA-AD.
- **Risk Factor Categorization:** Evaluated risk variables by partitioning them into two distinct categories:
- **Metric Boundaries:** Standardized scores across a spectrum ranging from **-100%** (complete under-representation) to **+100%** (complete over-representation).



CONCLUSIONS

- **Under-representation** dominates most cohorts, while **MIRIAD** shows an opposing **over-representation** bias.
- This is **critical for the field** because biased datasets directly produce **inequitable AI predictions**, reducing clinical applicability.
- Develop **fairness-aware models** that account for representation biases.
- Implement **stratified data curation strategies** to balance risk factor distributions.
- Design **prospective collection protocols** that explicitly target under-represented risk groups.
- Aim to achieve **equitable AD research** and improve generalizability of AI models.

REFERENCES

[1] Petersen RC, et al. ADNI: clinical characterization. *Neurology*. 2010;74(3):201-209.
[2] Dagley A, et al. Harvard aging brain study: dataset and accessibility. *NeuroImage*. 2017;144:255-258.
[3] Malone IB, et al. MIRIAD: public release of Alzheimer's MR imaging dataset. *NeuroImage*. 2013;70:33-36.
[4] Beekly DL, et al. NACC database: the uniform data set. *Alzheimer Dis Assoc Disord*. 2007;21(3):249-258.

CONTACT

mansi.singhal@ucalgary.ca

ADVANCED IMAGING & ARTIFICIAL INTELLIGENCE LAB

ALZHEIMER'S ASSOCIATION | AAIC 26