Understanding ethnic diversity in open dementia neuroimaging datasets

Nicholas Heng¹, Timothy Rittman¹

¹Department of Neurosciences, University of Cambridge, Herchel Smith building, Cambridge Biomedical Campus, Robinson Way, Cambridge, CB2 0SZ

Address for correspondence: Dr Nicholas Heng, Department of Clinical Neurosciences, University of Cambridge, Herchel Smith building, Cambridge Biomedical Campus, Robinson Way, Cambridge, CB2 0SZ

Electronic mail: Nicholas.Heng@nhs.net

Keywords

Neuroimaging; Ethnicity; Neurodegenerative disorders; Dementia

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Abstract

Introduction

Ethnic differences in dementia are increasingly recognised in epidemiological measures and diagnostic biomarkers. Nonetheless, ethnic diversity remains limited in many study populations. Here we provide insights into ethnic diversity in open access neuroimaging dementia datasets.

Methods

Datasets comprising dementia populations who underwent neuroimaging assessment with available data on ethnicity were included. Statistical analyses of sample and effect sizes were based on the Cochrane Handbook.

Results

14 databases were included, with 12 studies of healthy and MCI groups, and 11 of dementia groups. Combining all studies, the largest ethnic group was Caucasian (21,512 participants) with the next most common being Afro-Caribbean (1,960), followed by Asian (780). The smallest effect size detectable within the Caucasian group was 0.03, compared to Afro-Caribbean (0.1) and Asian (0.16).

Discussion

Our findings quantify the lack of ethnic diversity in openly available neuroimaging dementia datasets. More representative data would facilitate the development and validation of neuroimaging biomarkers relevant across ethnicities.

(150 words)
Introduction

The past few decades have seen growing interest in the field of biomarkers for neurodegenerative conditions. The neuroimaging community has led the way in open data, facilitating an explosion of research in neuroimaging biomarkers for dementia. This interest is in the context of an increasing global burden of neurodegenerative disorders, particularly in relation to the impact of Alzheimer’s disease and other dementias on an increasingly ageing population. Crucially, it has been estimated that the prevalence of dementia will increase from 57.4 million cases globally in 2019 to 152.8 million cases in 2050, posing a considerable risk to global healthcare and society in the near future.

There has been emerging evidence of ethnic differences amongst dementia populations, not only in prevalence and incidence, but also in cerebrospinal fluid (CSF) and imaging biomarkers. Nonetheless, many studies remain homogenous in the ethnicity of participants. This may hinder the translation of results to real world applications. As such, we aimed to provide insights into the ethnic diversity of currently available open neuroimaging dementia databases worldwide.

Materials and methods

We compiled and analysed demographic data reported by open access neuroimaging dementia databases. Databases were included if they consisted of (a) patients with a diagnosis of dementia or mild cognitive impairment, (b) had available neuroimaging data, and (c) demographic data including the breakdown of ethnicities. Datasets were identified through online research platforms including...
the Global Alzheimer’s Association Interactive Network (GAAIN) (https://www.gaain.org/), individual database repositories, and via peer-reviewed journal articles. We excluded datasets of solely genetic forms of dementia since these may be associated with specific ethnicities, or include large families that might bias the estimate of the distribution of ethnicities. A total of 46 databases were found but 32 were subsequently excluded as they either only included healthy controls, had no available data on demographics, or did not include neuroimaging data. Given the different definitions of ethnicities available, we took a pragmatic approach using the most widely used terms in the literature that permitted comparison between studies.

Statistical analyses on combined mean and standard deviation was performed as laid out by the Cochrane Handbook, and effect sizes calculations were using the pwr package in R (version 4.2.2) with. To compare samples of presumed equal sizes, we performed a power calculation for a two sample t test, estimating the effect size or sample size detectable with 90% power at a significance level (p value) of 0.05. Sample sizes were initially computed by setting a range of effect sizes, while minimum detectable effect sizes for single ethnic groups were then calculated using the aggregated dementia patient populations of different ethnicities from the open access dementia databases.

**Data availability**

Data sharing is not applicable to this article as no new data were created or analysed in this study.
Results

Demographics of dementia databases

A total of 14 dementia neuroimaging datasets were included, separated into the three diagnostic groups, with 12 including healthy participants (Supplementary Table 1), 12 including patients with mild cognitive impairment (MCI) (Supplementary Table 2) and 11 including patients with dementia (Table 1). In these tables, two entries for the Alzheimer’s Disease Neuroimaging Initiative (ADNI) dataset were made due to the separation of the ADNI-1 from ADNIGO and ADNI-2 cohorts. The majority of patients were from North America and Europe, with the two largest databases being from the National Alzheimer’s Coordinating Center (NACC) and UK Biobank respectively, in which there were a considerably higher percentage of Caucasians compared to other ethnicities.

Effect size analyses

To understand how the breakdown of ethnicity in these datasets could affect research studies, we calculated the sample sizes required for a range of effect sizes. For example, based on a recent systematic review and meta-analysis on fluid biomarkers for Alzheimer’s disease, it was found that CSF p-tau and t-tau levels were significantly higher in the Caucasian population compared to African Americans with MCI, with a standard mean difference of -0.50 (95% CI -0.73 to -0.28) and -0.52 (95% CI -0.75 to -0.30) respectively – though bearing in mind these did not necessarily inform the effect size in other biomarkers or ethnicities. Therefore, using an estimated effect size of 0.50 and basing off a power calculation of 90% and significance level of 0.05, the
number of patients required to detect a difference was N=86 each for two groups of patients of different ethnicities. We went on to calculate sample sizes for a range of effect sizes to obtain a better idea of the sample size to consider when planning future studies. In addition, we assessed whether the available data were sufficient to make comparisons between the Caucasian population and other ethnic groups (Table 2).

In an alternate approach, using available data for patients with dementia in those datasets combined with similar power calculation of 90% and significance of 0.05, we determined the smallest detectable effect size given currently available data (Table 3). The Caucasian population had the smallest minimum detectable effect size at 0.03 due to its size.

**Discussion**

With the increasing number of studies focusing on ethnic differences in dementia, there is little doubt that more emphasis needs to be placed on the role that ethnic differences play in biomarker research. Our findings suggest that despite the vast amount of comprehensive and high-quality data available worldwide, most participants come from a Caucasian background, limiting comparison to other populations. Considerable numbers of patients are required for assessing small magnitude effect sizes – which becomes particularly important when trying to identify potentially subtle differences between ethnicities. The minimum detectable effect size can therefore act as a guide or threshold towards that end. In fact, the majority of the population sizes were made up of two large databases in the UK and US. We hope these findings can act as a starting point into deciding how to expand representation of different ethnic groups in future.
studies on dementia.

Understanding the limitations of currently available data can provide an opportunity to uncover and
tackle the challenges associated with ensuring ethnic diversity in studies – stemming from a range
of barriers to participation including literacy and language barriers, lack of understanding or
misperceptions of the condition and services available, and cultural factors leading to mistrust of
services or stigma surrounding the condition. There are some encouraging initiatives to
address these challenges, such as those by the UK Dementia Research Institute, which are
designed not only to facilitate recruitment of diverse patient cohorts by increasing awareness and
broadening inclusion criteria, but also to improve access to healthcare through collaborations with
local teams.

We were only able to obtain data for openly available datasets. We know from published data and
from personal contacts that many studies use local cohorts, and some large national cohorts are
not shared with the wider community. We advocate exploring the barriers to sharing those data,
including the concerns of those who have collected and curate those datasets.

There are several limitations to our study – the first being that we were unable to comment on the
representativeness (as opposed to heterogeneity) of the combined characteristics of the
populations included in the database. Data on global ethnicity is not readily available and
classifications differ between different countries, making it difficult to draw comparisons. Secondly,
a considerable number of studies were excluded due to the lack of available demographic data,
and those that were included were mainly based in the Western hemisphere, which may mean we
have underestimate the non-Caucasian ethnicities actually available.

In our analysis we assume that datasets can easily be combined. In fact, harmonisation between
datasets presents a significant methodological challenge given that protocols differ and site effects
need to be modelled. This is particularly a challenge for combining neuroimaging data
despite the increasing availability of tools for this purpose such as ComBat.

Conclusion

With increasing awareness of the differences between ethnicities in dementia, it is imperative that
we begin to prioritise and broaden research to better understand underlying mechanisms, to
address the challenges associated with ethnic diversity in studies, and ultimately to pave the way
for reliable translation into clinical practice.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial,
or not-for-profit sectors. This research was supported by the NIHR Cambridge Biomedical
Research Centre (NIHR203312). The views expressed are those of the author(s) and not
necessarily those of the NIHR or the Department of Health and Social Care.
Competing Interests

The authors report no competing interests.

Supplementary Materials

Supplementary tables 1 and 2 will be available online.

References


Table 1. Table showing the breakdown of demographics data of patients with dementia in databases globally as separated by region.

<table>
<thead>
<tr>
<th>S/N</th>
<th>Database</th>
<th>Number of patients with dementia</th>
<th>Mean age (SD)</th>
<th>Gender</th>
<th>Ethnicity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Male (%)</td>
<td>Female (%)</td>
</tr>
<tr>
<td>North America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>ADNI-1 (USA) (10)</td>
<td>192</td>
<td>75.3 (7.5)</td>
<td>52.6</td>
<td>47.4</td>
</tr>
<tr>
<td>2</td>
<td>ADNIGO and ADNI-2 (USA) (11)</td>
<td>145</td>
<td>74.6 (8.1)</td>
<td>59.0</td>
<td>41.0</td>
</tr>
<tr>
<td>3</td>
<td>NACC (USA) (12)</td>
<td>20,053</td>
<td>75.9 (10.8)</td>
<td>48.0</td>
<td>52.0</td>
</tr>
<tr>
<td>South America</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>HABLE (USA) (13)</td>
<td>185</td>
<td>68.2 (9.9)</td>
<td>45.9</td>
<td>54.1</td>
</tr>
<tr>
<td>5</td>
<td>Argentina-ADNI (Argentina) (15)</td>
<td>12</td>
<td>77.9 (5.5)</td>
<td>41.7</td>
<td>58.3</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>I-ADNI (Italy) (16)</td>
<td>201</td>
<td>71.8 (8.4)</td>
<td>38.8</td>
<td>61.2</td>
</tr>
<tr>
<td>7</td>
<td>UK Biobank (UK) (17)</td>
<td>2778</td>
<td>64.7 (4.2)</td>
<td>45.3</td>
<td>54.7</td>
</tr>
<tr>
<td>8</td>
<td>ARWIBO (Italy) (18)</td>
<td>402</td>
<td>73.5 (8.5)</td>
<td>36.8</td>
<td>63.2</td>
</tr>
<tr>
<td>9</td>
<td>EDS (Italy, Germany, Netherlands) (19)</td>
<td>139</td>
<td>73.0 (8.0)</td>
<td>43.2</td>
<td>56.8</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>J-ADNI (Japan) (20)</td>
<td>149</td>
<td>73.7 (6.6)</td>
<td>43.0</td>
<td>57.0</td>
</tr>
<tr>
<td>11</td>
<td>WMH-AD (Taiwan) (from GAAIN)</td>
<td>43</td>
<td>77.2 (7.7)</td>
<td>25.6</td>
<td>74.4</td>
</tr>
<tr>
<td>12</td>
<td>KBASE (South Korea) (21)</td>
<td>87</td>
<td>73.0 (8.1)</td>
<td>31.0</td>
<td>69.0</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>24,368</td>
<td>74.4 (10.7)</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Table 2. Table showing the sample sizes required for specific effect sizes to be obtained based on power calculations of 90% and significance level of 0.05, with subsequent columns showing whether comparisons between ethnic groups can be performed based on currently available data.

<table>
<thead>
<tr>
<th>Effect Size</th>
<th>Sample size required</th>
<th>Caucasian vs Afro-Caribbean</th>
<th>Caucasian vs Asian</th>
<th>Caucasian vs Mixed</th>
<th>Caucasian vs Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>86</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>0.3</td>
<td>235</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>0.2</td>
<td>527</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>0.1</td>
<td>2103</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>0.05</td>
<td>8407</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>0.01</td>
<td>210150</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Table 3. Table showing the smallest effect sizes detected using population sizes based on available ethnicity data from dementia databases with similar parameters of 90% power and significance level of 0.05.

<table>
<thead>
<tr>
<th>Ethnicity group</th>
<th>Total number of patients with dementia</th>
<th>Smallest effect size detected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>21512</td>
<td>0.03</td>
</tr>
<tr>
<td>Afro-Caribbean</td>
<td>1960</td>
<td>0.10</td>
</tr>
<tr>
<td>Asian</td>
<td>780</td>
<td>0.16</td>
</tr>
<tr>
<td>Mixed</td>
<td>532</td>
<td>0.20</td>
</tr>
<tr>
<td>Others</td>
<td>138</td>
<td>0.39</td>
</tr>
</tbody>
</table>