Keeping pace – progress in dementia research capacity

Methodology and Data Sources Appendix

1. Introduction

This annex outlines the detailed methodology for the bibliometric analysis undertaken by Thomson Reuters. It describes the data sources and methods used to assess dementia research in the UK and several other countries around the world. Methods used in bibliometric analyses, data collection and estimates of research papers are also described.

2. Definitions and descriptions of data sources

Papers/publications: Thomson Reuters abstracts publications including research journal articles, editorials, meeting abstracts and book reviews. The terms "paper" and "publication" are often used interchangeably to refer to printed and electronic outputs of many types. In the analyses presented here, we have used the term publication to refer to substantive journal articles, review and some proceedings papers and excludes editorials, meeting abstracts or other types of publication.

Web of Science: For the bibliometric analysis, data were sources from the database underlying the Thomson Reuters Web of Science™, which gives access to conference proceedings, patent, website, and chemical structures, compounds and reactions in addition to journals. It has a unified structure that integrates all data and search terms together and therefore provides a level of comparability not found in other databases. It is widely acknowledged to be the world's leading sources of citation and bibliometric data. The Web of Science Core Collection is part of the Web of Science, and focuses on research published in journal and conferences in science, medicine, arts, humanities and social sciences. The authoritative, multidisciplinary content covers over 27,000 of the highest impact journals worldwide, including Open Access journals and over 161,000 conference proceedings. Coverage is both current and retrospective in the sciences, social sciences, art and humanities, in some cases back to 1900. Within the research community, these data are often still referred to by the acronym "ISI". Thomson Reuters has extensive experience with databases on research inputs, activity and outputs and has developed innovative analytical approaches for benchmarking and interpreting international, national and institutional research impact.

Co-authorship of publications: the metadata associated with every research publication include the addresses of the authors. It is thus possible to develop an analysis of the organisations that co-author publications by extracting and examining these data. Co-authorship is generally accepted as an indicator of collaboration, although there are collaborations that do not result in co-authored publications and co-authored publications which involve limited collaboration. Conceivably other indicators of collaboration such as co-funding and international exchanges could be used but comprehensive and consistent data are not available.

Internationally collaborative publications: The number of internationally collaborative research publications is increasing rapidly. This is because such collaboration provides access to a wider range of resources, including intellectual resources, and accelerates the rate of discover as well as increasing the intellectual content and therefore the impact of

individual outputs. For this reason, internationally collaborative publications tend to be more highly cited than those that are solely domestic. In this analysis, publication will be considered to be international if more than one country included in the addresses associated with the authors of a paper.

3. Identification of disease-specific research publications

Preliminary searches were based on interpretation of the disease area from definitions in the 10th Revision of the International Classification of Diseases (ICD-10) [http://www.who.int/classifications/icd/en/] using the following codes:

- Dementia: F00 – F03 and G30 which includes Alzheimer's disease, vascular and unspecified dementia, as well as dementia in other diseases such as Parkinson's;
- Cancer: C00 – D48 which includes malignant neoplasms, lymphomas and leukaemias;
- Coronary heart disease: I20 – I25 basically classified as ischaemic heart disease including angina;
- Stroke: I60 – I69 including cerebral haemorrhage and stenosis of cerebral arteries.

While the United States only fully transitioned to ICD-10 in 2015, the codes themselves only serve as a guide for the composition search strings.

Article titles, abstracts and keywords were searched using search strings created from combinations of text words and text strings identified from these definitions to describe the disease area and using the following rules.

- The OR operand works by searching for the words in any combination from any of the searchable text;
- The SAME operand selects only when the words appear in the same sentence (or title or set of keywords);
- The AND operand selects when all search terms appear in anywhere in the text (abstract, title or set of keywords)
- The $ symbol denotes a single character e.g. utili$ation finds both utilisation and utilization;
- The * symbol is used to denote any number of characters e.g. network* finds networks, networking, networkers as well as network;
- Using text within "" finds the text string e.g. "health delivery" would not find 'delivery of health services' whereas (health SAME delivery) would find both phrases.

In order to identify publications in the selected disease areas and build the publication dataset for the analyses in this report, the same search terms used for the 2012 Alzheimer's Research Defeat Dementia report were used in this study. These terms are listed below: Final search terms used for the collation of research publications were:

- Dementia: Alzheimer* OR dementia*
- Cancer: cancer OR neoplasm$ OR neoplastic OR carcinoma$ OR melanoma$ OR lymphoma$ OR myeloma$ OR leukaemia OR leukemia OR (Hodgkin* SAME disease)
- Coronary heart disease: "coronary heart disease" OR angina OR "myocardial infarction" OR atherosclerosis
- Stroke: stroke OR (cerebrovascular AND disease) OR ((cerebral OR brain OR subdural) AND (aneurysm OR haemorrhage OR hemorrhage OR ischaemi* OR infarction)).

The datasets used for this project were limited to articles and reviews (excluding conference proceedings, letters, editorials etc.) published in 2014-15 and indexed in Thomson Reuters Web of Science™ Core Collection: Science Citation Index-Expanded, Social Sciences Citation Index and Arts & Humanities Citation Index. Dementia papers published in 2008-9 were also included.

A further restriction was used to identify only those papers with at least one author from one of the selected countries: the UK, Frances, Germany, Sweden or the USA. Author affiliations were identified through author-address link data in Web of Science.

4. Disambiguation

Organisation unification: Correctly associating research outputs with institutions is an essential component of the analysis process, given that authors present their institutional affiliation in a variety of ways. The identification of institutions is performed using the author addresses from the Web of Science Core Collection.

Names and entries are updated to reflect organisational changes, and this unification is applied to addresses in new articles published in the Web of Science. This method relies on the accuracy of this information is provided by the authors: reported institutions may have variant names that are not unified. Wherever possible, Thompson Reuters used existing institution name unifications (of author-provided address variants) available on the Thomson Reuters Incites™ platform of the Web of Science Core Collection please refer to the InCites Indicators Handbook under the institution section (http://researchanalytics.thomsonreuters.com/m/pdfs/indicators-handbook.pdf). All analyses within the report are based on the author-address link data (the only data available which links the author with their affiliated country).

Researcher disambiguation: In order to assess the number of dementia researchers in the UK, rules had to be established to distinguish one researcher from another. Author-address link data provide a source for information to allow identification of distinct authors. In this report, the number of researchers was estimated using two rules. The first rule counts each distinct combination of an author's last name and fist initial (LNFI) as one author. The first initial is used instead of the full first name because many author names listed on publications are incomplete (i.e., the first column in Table 1 shows that many author names listed include the last name and first initials only). This rule is likely to underestimated the total number of researchers and is thus used as a lower-bound. For example, using this rule, all of the entries in Table 1 would be attributed to one author. "Williams, G". However, closer examination of the author-address link data reveals that there are two authors, Guy Williams and Gareth Williams. Since the full first name is not always available, a second rule was developed that identifies individual researchers using a combination of last name, first initial and the first part of the author address, which usually corresponds to the abbreviated institution name (LNFI + ADDR). This rule may overestimate the number of researchers. For example, using this rules, the entries in Table 1 would be attributed to three authors, "Williams, G from King's College London", "Williams, G from Univ Cambridge", and "Williams, G from Oxford Univ". It is not possible to determine whether "Williams, G from King's College London" and "Williams, G from Oxford Univ" were correctly identified as two authors or whether "Williams, G [Gareth]" moved from one university to another without manual review, confirmation from the authors or large-scale author disambiguation. Thus, the LNFI + ADDR rule was used as an upper-bound for the estimate of the number of researchers.

Author	**Address**	**Researcher (LNFI)**	**Institution (ADDR)**
Williams, G. B.	Univ Cambridge, Wolfson Brain Imaging Ctr, Dept Clin Neurosci, Addenbrookes Hosp, Cambridge CB2 0QQ, England	Williams, G	Univ Cambridge
Williams, Gareth	King's Coll London, Wolfson Ctr Age Related Dis, London SE1 1UL	Williams, G	King's Coll London

Williams, Gareth	Oxford University	Williams, G	Oxford Univ
Williams, Guy	Univ Cambridge, Addenbrookes Hosp, Wolfson Brain Imaging Ctr, Cambridge CB2 0QQ, England	Williams, G	Univ Cambridge
Williams, Guy B.	Univ Cambridge, Addenbrookes Hosp, Sch Clin Med, Dept Clin Neurosci,Wolfson Brain Imaging Ctr, Cambridge CB2 0QQ, England	Williams, G	Univ Cambridge
Williams, Guy B.	Univ Cambridge, Sch Clin Med, Dept Clin Neurosci, Cambridge, England	Williams, G	Univ Cambridge