DATA NOTE

Generation of a cleaned dataset listing Avon Longitudinal Study of Parents And Children peer-reviewed publications to 2015 [version 1; referees: 2 approved]

Oliver Butters [1,2], Amran Ismail [2], Sue Thompson[1], Rebecca Wilson [1,2]

[1]Institute of Health and Society, Newcastle University, UK, Newcastle upon Tyne, UK
[2]Social and Community Medicine, University of Bristol, Bristol, UK

v1 **First published:** 19 Dec 2018, **3**:161 (https://doi.org/10.12688/wellcomeopenres.14986.1)
Latest published: 19 Dec 2018, **3**:161 (https://doi.org/10.12688/wellcomeopenres.14986.1)

Abstract

Birth cohort studies generate huge amounts of data, and as a consequence are a source of many peer reviewed publications. We have taken the list of publications from the Avon Longitudinal Study of Parents and Children UK birth cohort, filtered, de-duplicated and cleaned it to generate a bibliographic research data set. This dataset could be used for accurate reporting and monitoring of the impact of the study as well as bibliometric research.

Keywords

Birth cohort, Bibliography, ALSPAC

This article is included in the Avon Longitudinal Study of Parents and Children (ALSPAC) gateway.

Open Peer Review

Referee Status: ✓ ✓

	Invited Referees 1	2
version 1 published 19 Dec 2018	✓ report	✓ report

1 **Carly Strasser**, Fred Hutchinson Cancer Research Center (Fred Hutch), USA

2 **Dylan Kneale**, University College London (UCL), UK

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Oliver Butters (olly.butters@newcastle.ac.uk)

Author roles: Butters O: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Software, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Ismail A**: Data Curation, Investigation; **Thompson S**: Data Curation, Investigation; **Wilson R**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Supervision, Validation, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This project is funded by CLOSER, whose mission is to maximise the use, value and impact of longitudinal studies. CLOSER is funded by the Economic and Social Research Council (ESRC) and Medical Research Council (MRC) (grant reference: ES/K000357/1). The UK Medical Research Council and Wellcome (Grant ref: 102215) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors they will serve as guarantors for the contents of this paper. RW is also supported by the UK Medical Research Council (MRC) (award reference: MR/S003959/1). AI was funded in part by the Nuffield Foundation research placement program.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Butters O, Ismail A, Thompson S and Wilson R. **Generation of a cleaned dataset listing Avon Longitudinal Study of Parents And Children peer-reviewed publications to 2015 [version 1; referees: 2 approved]** Wellcome Open Research 2018, **3**:161 (https://doi.org/10.12688/wellcomeopenres.14986.1)

First published: 19 Dec 2018, **3**:161 (https://doi.org/10.12688/wellcomeopenres.14986.1)

Introduction

Birth cohort studies in the U.K. generate and distribute huge amounts of longitudinal data for medical, social and economic research. Data is generally applied for and given out to researchers once the relevant governance conditions have been met[1]. It is often the case that these studies keep track of the publications that have arisen from the data they have given to researchers for project monitoring purposes and to report back to the funder(s). The size of these lists of publications is sometimes used as a crude metric of the the research outputs or impact for the study.

Most modern academic journals will assign a unique persistent identifier to new publications. This persistent identifier may be unique and resolvable by the journal, but may be meaningless outside of the journal's ecosystem. The Digital Object Identifier (DOI) is the de facto persistent identifier which is used as an independent external reference to publications, posters, data, software etc. DOI resolving services exist to refer users (human and machine) to the journal web page for a given DOI, CrossRef holds over 100 million such records. These resolving services also host a wealth of metadata themselves. The DOI data model outlines the format of DOI data. In addition to CrossRef there exists other resolving and metadata services that are domain-specific. These may have more in depth metadata about their domain than the generality that the DOI data model can offer. In this work we also make use of the persistent identifiers that the National Center for Biotechnology Information (NCBI) PubMed generates (PubMed IDs, PMID), and the metadata their resolving service provides[2]. This offers extra metadata over and above that available from CrossRef, but only on medical focused publications, i.e. a subset of all available publications in birth cohort studies.

In this paper we describe how we created a cleaned, de-duplicated list of peer-reviewed publications arising from the Avon Longitudinal Study of Parents and Children (ALSPAC). ALSPAC began in 1990 (see the cohort profiles for an overview[3,4]), and has publications within the biomedical research domain. ALSPAC reports to have over 1800 publications as of August 2018[5]. The study website contains details of all the data that is available through a fully searchable data dictionary and variable search tool.

Methods

Data cleaning

The ALSPAC master list of publications at the time this project started (2014), consisted of a large table in a Microsoft Word document. This table was imported into a spreadsheet containing a reference to the publication, a DOI and a PMID. Given the amount of time that has passed since the original master list was parsed we have merged this list with the list of publications on the ALSPAC website as at 12/9/18. One pertinent point is that there exists a small number of publications in the original Microsoft Word document that are not present on the website; we include these here for completeness.

Each publication was audited manually to ensure it was a peer reviewed publication i.e. that the journal had a defined peer-review process and/or that it appeared in Ulrichs Web Global Serials Directory with a "refereed" status. Non-peer-reviewed articles were removed from the publications list. Examples of non-peer-reviewed publications included theses, book chapters, published abstracts, opinion articles, comments on other articles, working papers and technical reports.

The DOI and PMID for each entry were also audited to validate the identifier and ensure they corresponded to the correct article. A common error was the truncation of a PMID, which due to the numerical nature of PMIDs was itself a valid PMID albeit referring to the wrong publication. If a DOI or PMID was missing from a publication, wherever possible this was sourced from the journal or PubMed directly. The DOI and PMID fields from the publications spreadsheet were used to import the publications lists into a bibliographic library in Zotero. Zotero uses NCBI PubMed to resolve PMIDs and CrossRef to resolve DOIs.

We then further cleaned the list of publications by deduplicating the list using Zotero's native de-duplicate feature. Duplicates often arose in the bibliography when a publication was accepted in one year and then appeared online the next, or when it was listed with a DOI in one case and a PMID in another. Another common source of duplicates was having both the pre-print and the final published paper marked as separate items. In this case we disregarded the pre-print.

Given that publications are not necessarily reported to ALSPAC on acceptance to a journal, and some journals have a long turn around in publication time, we chose to have a cut-off of the end of 2015 for this data set. Given the misclassification of years of some publications, we added all publications up to the end of 2016 (as defined by the list on the ALSPAC website), but disregarded any that had a publication date after the end of 2015. This criteria left us with 1300 peer reviewed publications claimed by ALSPAC to the end of 2015. Table 1 shows a summary of the data.

Data description

To make this list of publications available to others in as useful way as possible we exported it from our Zotero library in two different formats: BibTeX format to be able to import into any reference manager and comma separated variable (CSV) to allow import into analysis tools to do bibliometric analysis with. Both of these formats are described in Table 2 and Table 3, respectively. Zotero v5.0.56 was used to export the data.

Table 1. Data coverage. Percentages rounded down in each case.

Date range	1989–2015
Publication count	1300
DOIs (%)	97
PMIDs (%)	95
Publication title (%)	100
Year published (%)	100

Table 2. A data description of the BibTeX ALSPAC peer reviewed publications list to 2015.

Variable	Description
citation key	A unique identifier
title	Article title
author	Name(s) of author(s)
abstract	Article abstract
journal	Journal title
volume	Journal volume
number	Journal issue
pages	Article page numbers in the journal
year	Year published
month	Month published
keywords	Article keywords
issn	International Standard Serial Number
doi	Digital Object Identifier
pmid	PubMed identifier
pmcid	PubMed Central identifier

Table 3. A data description of the CSV file of ALSPAC peer reviewed publications list to 2015.

Variable	Description
Year	Year published
Author	Name(s) of author(s)
Title	Article title
Publication title	Journal title
ISSN	International Standard Serial Number
DOI	Direct Object Identifier
Abstract Note	Article abstract
Date	Date article published
Pages	Article page numbers in the journal
Issue	Journal issue
Volume	Journal volume
Extra	PubMed and/or PubMed Central ID;
Manual tags	Article keywords

Data availability

The cleaned BibTeX and CSV data described here are available at Zenodo. DOI: https://doi.org/10.5281/zenodo.2276785[6].

All of the metadata presented here is publicly available in its raw form—the list of publications is available from the ALSPAC website and the individual publications' metadata from their respective publishers. PubMed and CrossRef have additional terms and conditions[1,2] on their aggregated metadata, but these are permissive and allow fair use.

Grant information

This project is funded by CLOSER, whose mission is to maximise the use, value and impact of longitudinal studies. CLOSER is funded by the Economic and Social Research Council (ESRC) and Medical Research Council (MRC) (grant reference: ES/K000357/1). The UK Medical Research Council and Wellcome (Grant ref: 102215) and the University of Bristol provide core support for ALSPAC. This publication is the work of the authors they will serve as guarantors for the contents of this paper. RW is also supported by the UK Medical Research Council (MRC) (award reference: MR/S003959/1). AI was funded in part by the Nuffield Foundation research placement program.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

We are extremely grateful to all the families who took part in ALSPAC, the midwives for their help in recruiting them, and the whole ALSPAC team, which includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists and nurses. We are particularly grateful to the ALSPAC secretaries who have maintained the list of publications.

References

1. Murtagh MJ, Blell MT, Butters OW, *et al.*: **Better governance, better access: practising responsible data sharing in the METADAC governance infrastructure.** *Hum Genomics.* 2018; **12**(1): 24.
PubMed Abstract | Publisher Full Text | Free Full Text
2. NCBI Resource Coordinators: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res.* 2018; **46**(D1): D8–D13.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Boyd A, Golding J, Macleod J, *et al.*: **Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children.** *Int J Epidemiol.* 2013; **42**(1): 111–127.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Fraser A, Macdonald-Wallis C, Tilling K, *et al.*: **Cohort Profile: the Avon Longitudinal Study of Parents and Children: ALSPAC mothers cohort.** *Int J Epidemiol.* 2013; **42**(1): 97–110.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Children of the 90s, (@CO90s): **#DidYouKnow know: 1800+ scientific papers have been published using Children of the 90s data, covering #Depression in #pregnancy, #obesity in children, childhood IQ & #bipolardisorder, #anxietyproblems and more http://www.bristol.ac.uk/alspac/news/.** Twitter, 2018.
Reference Source
6. Butters O, Ismail A, Thompson S, *et al.*: **ALSPAC peer reviewed publications 1989-2015 (Version 1.0) [Data set].** *Zenodo.* 2018.
http://www.doi.org/10.5281/zenodo.2276785

Open Peer Review

Current Referee Status: ✔ ✔

Version 1

Referee Report 22 February 2019

https://doi.org/10.21956/wellcomeopenres.16347.r34479

✔ **Dylan Kneale**
Evidence for Policy and Practice Information and Coordinating Centre, UCL Institute of Education, University College London (UCL), London, UK

Thank you for the opportunity of reviewing this data note. I think that this represents really exciting work and a huge effort in documenting studies using the data and cleaning these.

There are some aspects that I think could be better described to help support other similar exercises in the future:

1. I think the rationale around collating this information could be strengthened a little. In particular, I think the rationale should better make the case that understanding the scientific impact of these cohort studies is key in supporting the continuation of this study and funding future studies.
2. The source data are drawn from records held by the ALSPAC team, which has been keeping track of publications. These formed a ‘master list’ of publications which was then extensively cleaned and refined to form the dataset. However, the processes used to keep track of publications need to be better described – how are studies identified and to what extent do the researchers feel that these records represent the full range of peer-reviewed studies published using ALSPAC data?
3. The authors described that these are publications arising from ALSPAC data. Were any criteria imposed on what this ‘usage’ should look like? For example would a commentary that makes reference to the ALSPAC data (possibly alongside other studies) be included as a publication; would secondary analyses of studies using ALSPAC data be included (e.g. using an effect size from a study using ALSPAC data as part of a meta-analysis)? While the PMID and DOIs were cleaned, were the studies checked for their usage of these data? This seems important to clarify. While this data note describes a dataset of ALSPAC publications, I’m not clear if this is exclusively a dataset of primary studies using ALSPAC data in novel analyses, or if it also includes other publication types. If the dataset does include other publication types, does this have implications for the way in which the dataset should be used by future researchers?
4. As a minor suggestion, it may be interesting to have an addition to Table 1 that includes a breakdown of publications by year.

The suggestions made above are mainly for clarification to help understand the parameters of the data set. I would like to again emphasise that this data note does represent a huge task undertaken and has resulted in a very worthwhile output.

Is the rationale for creating the dataset(s) clearly described?
Partly

Are the protocols appropriate and is the work technically sound?
Partly

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Referee Report 29 January 2019

https://doi.org/10.21956/wellcomeopenres.16347.r34616

Carly Strasser
Fred Hutchinson Cancer Research Center (Fred Hutch), Seattle, WA, USA

The manuscript reports on creating a complete bibliography of publications associated with the ALSPAC longitudinal study. Collecting such data is not trivial given the duplication via preprints, PMIDs, and variable metadata associated with articles. This work is important for understanding the impacts of the study, as well as potential future meta-analyses.

Is the rationale for creating the dataset(s) clearly described?
Yes

Are the protocols appropriate and is the work technically sound?
Yes

Are sufficient details of methods and materials provided to allow replication by others?
Yes

Are the datasets clearly presented in a useable and accessible format?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: open science, data management

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.