ARDC
Australian Research Data Commons

Secondary use of clinical trials data in health research

A PRACTICAL GUIDE

Kylie Hunter, Jonathan Williams, Talia Palacios, Aidan Tan, Kristy Robledo, Angelina Tjokrowidjaja, Amany Gouda-Vossos, Kristan Kang, Anna Lene Seidler

17/07/2024

Cover Image - Siarhei - 337007459 / adobestock.com

DOI: http://doi.org/10.5281/zenodo.12768050

Suggested citation:
Hunter, K.E., Williams, J., Palacios, T., Tan, A.C., Gouda-Vossos, A., Kang, K., Robledo, K, Tjokrowidjaja, A., Seidler, A.L. (2024). *Secondary use of clinical trials data in health research: A practical guide v.1*

Viewed online at:
http://doi.org/10.5281/zenodo.12768050

Published:
<17/07/2024>

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

The ARDC is enabled by NCRIS

Acknowledgement of Country

We acknowledge the traditional custodians throughout Australia and their continuing connection to, and deep knowledge of, the land and waters. We pay our respects to Elders both past and present.

Version	*Start Date*	*Authors*	*Notes*
1.0	17/07/2024	ARDC	Original version

ARDC is enabled by NCRIS

CONTENTS

Background **3**

Scenario 1: Evidence synthesis **5**

What is evidence synthesis? 5

What data sources do I need to perform an evidence synthesis analysis? 5

What are the steps required to perform an evidence synthesis analysis? 6

What are the advantages and considerations of evidence synthesis analysis? 7

What are some examples of evidence synthesis? 9

Case Study: Evidence Synthesis Scenario (Individual participant data meta-analysis) 10

What are some key resources for evidence synthesis? 15

Scenario 2: Secondary analyses **16**

What is a secondary analysis? 16

What data sources do I need to perform a secondary analysis? 17

What are the steps required to perform a secondary analysis? 18

What are the advantages and considerations of secondary analysis? 19

What are some examples of secondary analysis? 20

Case Study: Secondary Analysis Scenario (Answer New Research Questions) 21

What are some key resources for secondary analysis? 23

Scenario 3: Reproducibility, replication and validation **25**

What is reproducibility, replication and validation with secondary data? 25

What data sources do I need to perform reproducibility, replication and validation? 26

What are the steps required to perform reproducibility, replication and validation studies? 26

What are the advantages and considerations of reproducibility, replication and validation? 28

What are some real-world examples of reproducibility, replication and validation with secondary data? 29

What are some key resources for reproducibility, replication and validation with secondary data? 30

Scenario 4: Education and Methods Development **31**

What is secondary data use in education and methods development? 31

What data sources do I need to use for secondary data for education and methods development? 32

What are the steps required to use secondary data for education and methods development? 32

What are the advantages and considerations when using secondary data for education and methods development? 32

What are some real-world examples of secondary data used in education and methods development? 34

Case study: Education and Methods Development (Developing and Demonstrating New Statistical Methods) 35

What are some key resources for secondary data used in education methods development? 38

Background

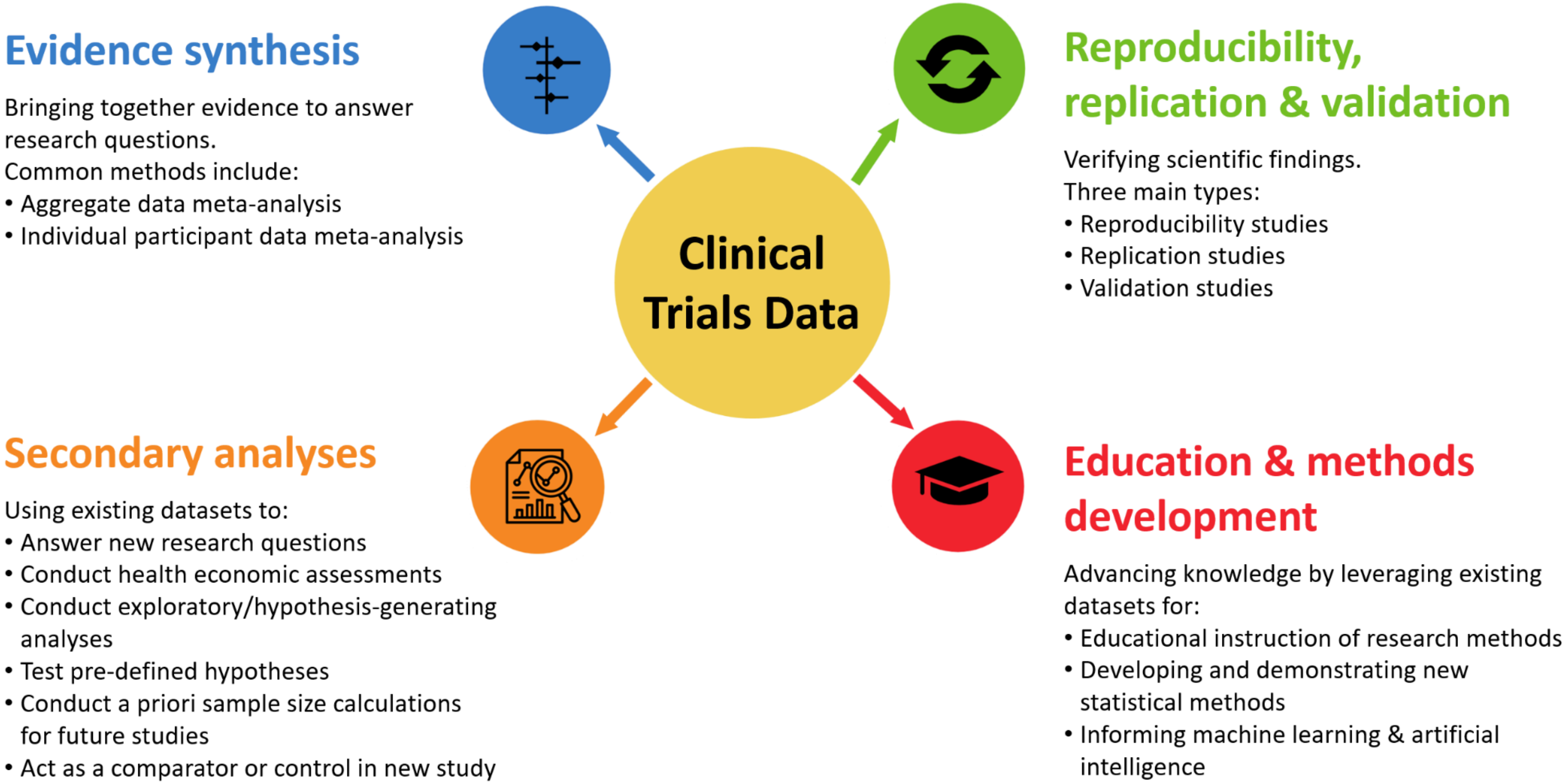
Data sharing is a core element of the Open Science[1] movement and adds tremendous value to existing data. The Open Science movement aims to make research, data, and their dissemination more accessible to all, and to increase research transparency and data reuse through the FAIR principles. Yet, structured guidance around different use-cases of secondary data, and their advantages, risks and limitations was lacking.

This document presents a theoretical framework for the use of clinical trials and other health data for secondary research purposes, which was derived from research papers, consultation with stakeholders and the research community. Click here to see a webinar about the theoretical framework. ARDC built the Health Data Australia platform to support the sharing of health research data for these secondary research scenarios. Four overall scenarios for data reuse were identified:

- **Scenario 1**: Evidence synthesis covers research projects bringing together evidence from different sources to answer a specific research question, e.g. effect of a health intervention, accuracy of a diagnostic test, prognostic effect of factors, or performance of risk prediction models.
- **Scenario 2:** Secondary analyses summarises research projects using existing data from one or more studies to answer a research question that is different from the original study(ies). This may include descriptive analyses, health economic assessments (e.g. cost-effectiveness of an intervention), exploratory analysis of trends or patterns in the data, or to test pre-defined hypotheses (e.g. about the relationships between variables or differences among subgroups of participants). This may also include using existing data to inform machine learning and artificial intelligence.
- **Scenario 3:** Reproducibility, replication and validation includes studies that aim to verify the accuracy, validity, and trustworthiness of the scientific findings of the original study.
- **Scenario 4**: Education and methods development covers the use of existing data as a valuable training resource to facilitate learning about data cleaning and analysis methods among researchers, students, and educators. This also includes the use of existing datasets to develop and demonstrate new statistical methods, and to inform machine learning and artificial intelligence.

ARDC has created the Health Data Australia platform to help researchers get better access to data for their research. We'd like to hear more about your research and hear how we can improve the platform to make it more useful for you. Click here to access the survey and provide your feedback.

[1] https://www.unesco.org/en/open-science



Clinical Trials Data

Evidence synthesis

Bringing together evidence to answer research questions.
Common methods include:

- Aggregate data meta-analysis
- Individual participant data meta-analysis

Reproducibility, replication & validation

Verifying scientific findings.
Three main types:

- Reproducibility studies
- Replication studies
- Validation studies

Secondary analyses

Using existing datasets to:

- Answer new research questions
- Conduct health economic assessments
- Conduct exploratory/hypothesis-generating analyses
- Test pre-defined hypotheses
- Conduct a priori sample size calculations for future studies
- Act as a comparator or control in new study

Education & methods development

Advancing knowledge by leveraging existing datasets for:

- Educational instruction of research methods
- Developing and demonstrating new statistical methods
- Informing machine learning & artificial intelligence

Figure 1: Types of secondary research using clinical trials data

Scenario 1: Evidence synthesis

What is evidence synthesis?

Evidence synthesis involves the comprehensive compilation of data from a variety of sources to answer a specific research question. These sources may be in the form of aggregate data (i.e. data that have been summarised across participants), or raw line-by-line data, known as individual participant data (IPD). IPD meta-analyses are considered the gold standard approach for evidence synthesis.

Definition	Brings together evidence to answer a specific research question
Key steps	Develop protocol, systematic search, study selection, data collection, appraisal, analysis, dissemination, update
Data types	Aggregate data (i.e. data that have been summarised across participants), or raw line-by-line data, known as individual participant data
Data sources	Journal publications, the Health Data Australia (HDA) platform, study investigators, clinical trials registers, data repositories
Advantages	Comprehensive, systematic, minimise bias, rigorous
Challenges	Obtaining data, time-consuming
Types	Aggregate data meta-analysis, Individual participant data meta-analysis
Time	Depends on research topic, number of included studies, methods used (type of meta-analysis), researcher experience
Expertise	Information specialist, statistical support, administrative support, data management

What data sources do I need to perform an evidence synthesis analysis?

For evidence syntheses, ideally data for all studies fulfilling certain eligibility criteria are obtained. These studies should be identified through systematic literature searches. The Health Data Australia (HDA) platform may be a useful source of data for Australian trials (registered in ANZCTR, and trials registered with CT.gov with an Australian site). However, this is not a complete data source for evidence syntheses, since data for international studies should also be sought, as well as data for Australian studies not registered on HDA. Therefore, data listed in HDA would need to be used in conjunction with other data sources. Data for studies not on HDA may be accessed by directly contacting study authors, or from data repositories, clinical trial registries, or journal websites. Investigators can request either individual participant data or summary / aggregate data or a combination of both for this purpose. Importantly, facilitating access to outcome data that are not publicly available mitigates selective reporting and publication bias. Detailed intervention and population descriptions are required, at times with more detail than typically available from publications. These can be requested from trial investigators. For traditional aggregate data meta-analyses, national data sharing infrastructure may be used to access unpublished trials or unpublished outcomes of trials to mitigate publication bias. Guidance for how to identify such unpublished evidence and include it in systematic reviews and meta-analyses is available here and here. In most cases, we anticipate HDA to be accessed for individual participant data meta-analyses, and thus, this guide mainly focuses on these analyses from here onwards.

What are the steps required to perform an evidence synthesis analysis?

1. Develop the review question and define eligibility criteria.
2. Plan methods and develop protocol in accordance with the Preferred Reporting Items for Systematic Review and Meta-Analysis Protocols (PRISMA-P). Use core outcomes sets where available.
3. Obtain ethics approval (optional) and publish protocol in a peer-reviewed journal or open access registry such as PROSPERO or OSF.
4. Search for relevant studies on databases (e.g. Medline (Ovid), Embase (Ovid), CINAHL (EBSCO)) and clinical trial registers.
5. Screen and select eligible studies.
6. For individual participant data -meta-analysis: You may invite trial representatives to be part of the project and contribute to design, analysis, and interpretation. In this case, a governance plan should be prepared to outline how the collaboration will be managed.
7. Extract aggregate data from publications or registries and/or retrieve de-identified individual participant data and accompanying codebooks from repositories or in direct communication with trial representatives. This may involve negotiating data sharing agreements and contractual obligations with data providers.
8. Prepare a statistical analysis plan (optional for aggregate data meta-analysis, best practice for individual participant data meta-analysis). This should be agreed upon by all collaborators prior to analysis and time-stamped or made publicly available to prevent data dredging and potentially biased post-hoc analyses.

9. Assess risk of bias, quality, and integrity of studies and their data using appropriate tools.
10. Synthesise & analyse eligible studies according to pre-specified statistical analysis plan.
11. Present and interpret results. Report according to appropriate guidelines (e.g. PRISMA) and endeavour to publish / disseminate results regardless of outcome.
12. Improve and update review if necessary, e.g. to address outstanding research questions, where new evidence has become available.

What are the advantages and considerations of evidence synthesis analysis?

Advantages of individual participant data meta-analyses:

- Greater data availability than from publications alone (due to harmonisation and inclusion of unreported outcome data), enabling more powerful analyses
- More rigorous checking of data quality, and data integrity, leading to improved data quality and analysis
- Greater ability to harmonise outcomes and variables across studies, and therefore overcome problems of different definitions or measures
- Capacity to undertake more complex/appropriate statistical models, e.g. adjust for confounders, imputation for missing values
- Enables examination of potential effect modification for key subgroups, which is important to inform precision medicine

Resources and funding

Individual participant data-meta-analyses require significantly more resources than aggregate data meta-analyses and upwards of two years to complete, depending on the number of included studies and data availability. This is because data collection, processing, cleaning, and communication with collaborations are time-consuming activities. Items that need to be carefully budgeted for include: negotiating data sharing agreements, facilitating data transfer, re-coding, checking and cleaning data in duplicate and resolving any queries with study representatives, conducting integrity and bias assessments, and collating all data into a single merged dataset.

Potential challenges and strategies for mitigation

- Insufficient recruitment of trials to contribute – negotiating participation of trials can be difficult, timely and takes careful diplomacy skills. Although data sharing is generally supported in theory, in-practice participation is frequently far lower, with only around 25% of individual participant data meta-analyses successfully retrieving all eligible individual participant data for analysis. Early engagement of potential contributors to ensure they are willing to share data, and proactively assisting with necessary data sharing agreements and ethical approvals can help obtain higher levels of data.

- Trials unable to share data – depending on the consent obtained from participants at the time of the trial, some ethics committees may not grant permission for data sharing. This can often be overcome by explaining to ethics committees that the individual participant data meta-analysis project aligns with the objectives of the initial study, and is therefore eligible for a waiver of consent. If this approach is unsuccessful, there are methods that can be used to re-construct individual participant data using summary data, in collaboration with study representatives if they are willing.
- Diplomatic balancing of opinions – when gathering all the experts in a particular field, it is unlikely they will agree on all aspects. Diplomacy in this situation can be challenging, and how situations like this will be handled should be considered by the investigators. Poor handling of such a situation can potentially result in withdrawal of collaborators from a project. On the other hand, bringing together experts in the field for robust discussion, can greatly enhance the quality and impact of results. To maintain a harmonious and fruitful collaboration, it is important for the project steering group to maintain some level of independence from the data providers, so that they may objectively mediate any issues that arise. Strategies to deal with potential conflicts should also be clearly outlined in a governance plan.

What are some examples of evidence synthesis?

Research question Effectiveness of health care interventions

Required data All/ most relevant studies on an international scale, identified through systematic literature searches. These can be assessed either with IPD, aggregate data, or a combination.

Research question Examining the role of biomarkers (e.g. patient characteristics, molecular / genetic factors) in effectiveness of interventions (individual-level subgroup analyses)

Required data All/ most relevant studies on an international scale that have measured biomarker of interest, identified through systematic literature searches. These would be ideally analysed with IPD, or with aggregate data that has been stratified by biomarker. Exploratory biomarker analyses may be undertaken with a smaller/ more selective sample of studies (e.g. only those available in a data catalogue – see Scenario 2: secondary analyses

Research question Examining the role of settings, patient populations or intervention characteristics on intervention effectiveness (trial-level subgroup analyses)

Required data All/ most relevant studies on an international scale, identified through systematic literature searches. These can be assessed either with IPD, aggregate data, or a combination. Detailed intervention and population descriptions are required, at times with more detail than typically available from publications. These can either be extracted from intervention materials, or requested from trial investigators.

Case Study: Evidence Synthesis Scenario (Individual participant data meta-analysis)

With Dr Anna Lene Seidler and Dr Kylie Hunter

Sharing secondary data to give babies born too early a better chance of survival

Study name: Individual Participant data on Cord management at preterm birth (iCOMP)

Start date: 2018 **End date:** 2023

Website: https://www.icompstudy.org/

What was this study about?

Delayed cord clamping is now a recommended routine practice for babies born at full term. However, while previous research showed potential benefit for premature babies, best practice for this vulnerable group remained uncertain. This led to different recommendations in national and international guidelines, and uncertainty amongst clinicians. This question was too complex to answer in a single trial or simple meta-analysis based on publications alone. Instead, combining secondary data in an individual participant data meta-analysis enabled the research team to answer this important question.

What type of secondary data scenario is this case study?

This is a case study of using secondary data for evidence synthesis purposes, i.e. bringing together evidence to answer a specific research question. In particular, this study collated raw line-by-line data, known as individual participant data. Individual participant data meta-analysis is considered the gold standard approach for evidence synthesis, and further guidance on this methodology can be found here.

How was secondary data used in this study?

For evidence synthesis, it is important to include data from all eligible studies to avoid potential bias. For this reason, iCOMP was based on a systematic review of the literature, and all identified eligible studies were invited to join the collaboration and share their data.

This resulted in a massive global effort (the iCOMP collaboration) among more than 100 international researchers, who shared their original trial data for analysis. This created one of the largest databases in this research field, with over 60 international studies including more than 9,000 babies from all over the world (see Figure 2 below).

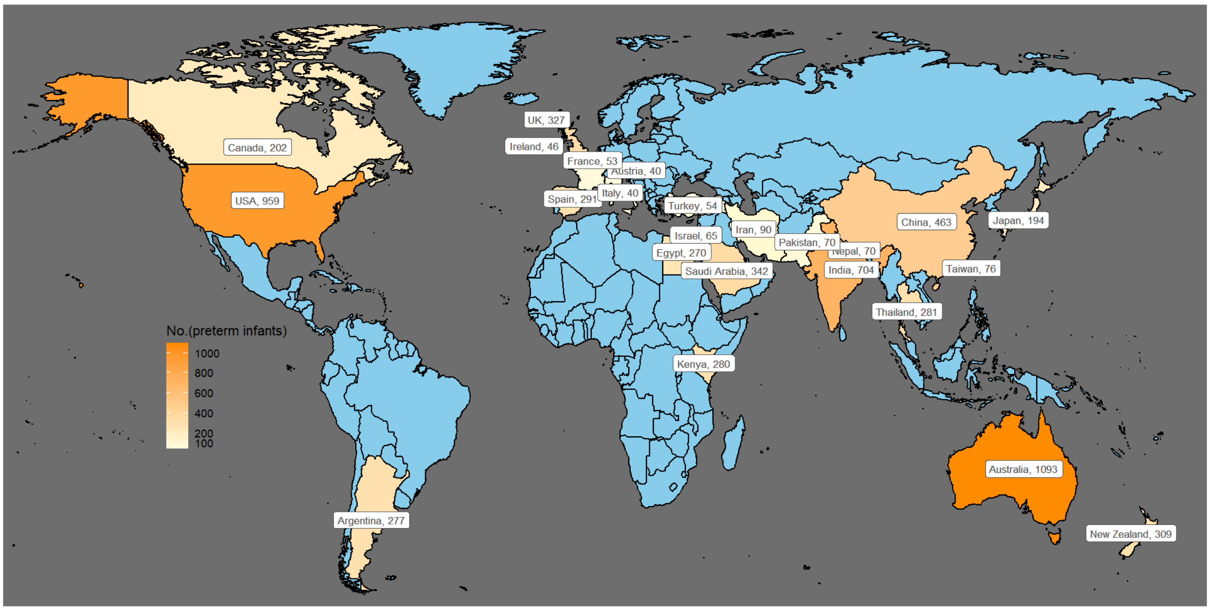


Figure 2: iCOMP collaboration trial data map.

What did the iCOMP studies investigate? And what did they find?

This large database was used to conduct two major studies: The first iCOMP study examined whether doctors should wait to clamp the cord, 'milk' the cord or clamp immediately, using data from 6367 infants across 48 studies. The second iCOMP study examined how long doctors should wait to clamp the cord, using data from 6,094 babies across 47 studies.

The first study found delaying umbilical cord clamping for 30 seconds or more after birth likely reduced mortality risk in premature babies compared to immediate clamping. The second study found waiting at least two minutes before clamping the cord may reduce mortality risk in premature babies compared with waiting less time.

What was the impact of the studies?

iCOMP showed with high certainty that waiting to clamp the umbilical cord reduces the risk of death for premature babies. These findings were published as a two-part fast-track series in the Lancet here and here, led to international media attention (including reports in the New York Times) and have already been implemented in international treatment recommendations by ILCOR, providing premature babies with a better chance of survival globally.

How did secondary data make this study and impact possible?

Previous standard reviews and meta-analyses based on publications alone where inconclusive, access to individual participant data resolved these problems for a number of reasons:

Improved data availability and quality: Previous reviews were limited by the fact that most available data could not be included in these reviews. Not all eligible studies were published and even those that were published often did not report all outcomes they collected. In addition, the ones that did report outcomes of interest often reported these outcomes using different data or categories, so it was difficult and sometimes even impossible to perform a meta-analysis of these outcomes. Access to individual participant data meant access to more and higher-quality outcome data. The iCOMP collaboration included a number of unpublished studies and many unpublished outcomes. The raw data enabled the study team to harmonise outcome variables, which greatly improved ability for analysis. In addition, individual participant data allowed in-depth checks, leading to a high-quality, more complete database underlying the analyses. This also enabled the study team to conduct careful analysis of potential adverse safety outcomes that are only possible to detect with such a large database.

Ability to conduct subgroup analyses: In addition, aggregate data meta-analyses are limited by their ability to assess differential treatment effects for different populations. Access to individual participant data allowed the study team to examine whether different groups (for example very early preterm infants, or twins) require a different treatment approach. The finding of the iCOMP study that the treatment effect was consistent for different groups of infants gave guideline developers and clinicians much greater confidence to widely apply this technique in practice.

How could Health Data Australia be used for this type of study?

The iCOMP study was conducted prior to the launch of Health Data Australia. This means that for the iCOMP study, there were few avenues to systematically find, understand and access secondary data underlying studies identified in the systematic review. The study team had to jump many legislative hurdles and decipher old datasets in a range of languages, often without a data dictionary. This process took many years and resources, and involved a lot of back and forth with the original trial investigators.

Health Data Australia will be an important resource to streamline access to data from Australian trials, making this process much easier for the secondary data study team but also the trial investigator. Yet, because evidence synthesis requires access to all eligible datasets, it is very important that also studies not available on Health Data Australia are included in this type of secondary analysis, even if finding and accessing their data may be more challenging.

Interview with the study team: Lessons from iCOMP for secondary data use for evidence synthesis

We interviewed some of the iCOMP study team to ask them about sharing the main lessons for secondary data use for the iCOMP study with other researchers embarking on similar projects.

How did you manage to bring together such a large collaboration?

“Maybe ignorance was bliss in our case. We did not expect to find that many eligible datasets when we first started, it was quite surprising to us and involved a lot of hard work. But in retrospect, having such a big dataset and collaboration really helped us get to the core of this research question, and hopefully make a real difference for babies and their families.”

“One thing we found really important was to form a real collaboration, and not just treat trial investigators as mere data providers. We found it was important to engage trial investigators early in the project to allow them to provide input into the protocol and analysis plan, and to improve willingness to share their data. This meant data sharers were an integral part of the study team and able to influence the direction of the study. We really appreciated all their insights as experts in the field. But not every investigator will have time to be this involved, so this model may not work for everyone.”

How do you plan this type of study?

“Learning from our experience, it is really important to do some preliminary searches to estimate the number of eligible studies for your research question. This information can then be used to inform resourcing requirements, in particular for collaboration management and data processing, checking, and cleaning, which can be very time-consuming. And one big piece of advice is to not underestimate the amount of time, resources and energy it takes to manage such a big collaboration and so many diverse datasets. This type of project requires sufficient funding, even if the funds needed are still only a fraction of what trying to collect new data on this research question would cost.”

What are some of the pitfalls or risks of this type of study for future researchers to be aware of?

“It is important to obtain as much relevant data as possible, to ensure that the meta-analysis has sufficient power to answer your research question and to mitigate data availability bias. This can be really difficult to achieve and is one of the main reasons individual participant data meta-analyses fail.”

So how did you then manage to retrieve such large proportions of available data?

“By employing several strategies and being very persistent. For example, we used several different methods to reach trial investigators, including email, phone, videoconferencing, face-to-face meetings at conferences and by contacting them through our networks. And as already mentioned above, we then sought to engage trial investigators from very early on in the project, by forming a collaboration, and seeking their input on the protocol (for which they were offered co-authorship) and analysis plan.”

How did you go about the data sharing process and requesting data?

“This was quite a work-intensive and manual process. We developed data sharing agreements, and then asked each investigator for their datasets and any additional information they can give you such as data dictionaries. Then, we had to try to figure out how to harmonise all these different datasets into one main dataset. A platform like Health Data Australia would have been so useful in streamlining some of these processes for Australian datasets.”

Were there any unexpected lessons from your study?

“The additional insights and opportunities that came from working in such a large collaboration with researchers from around the world. They knew the situation in their clinics and countries best, and helped make our research so much more relevant for guidelines and practice. There is a lot of power in international collaboration and it can lead to such great impact.”

Dr Anna Lene Seidler

Anna Lene Seidler (Lene) is a Senior Research Fellow and biostatistician at the NHMRC Clinical Trials Centre (CTC), University of Sydney where she leads the NextGen Evidence Synthesis team. She is also a Research Associate for the Australian New Zealand Clinical Trials Registry and Co-Convenor for the Cochrane Prospective Meta-Analysis Group. Lene specialises in systematic reviews, methods development, and individual participant data (IPD). She leads several large international research projects, such as the iCOMP collaboration and the TOPCHILD collaboration. Her clinical focus areas are obesity and neonatology.

Dr Kylie Hunter

Kylie Hunter is a Research Fellow for the NextGen Evidence Synthesis Team at the NHMRC Clinical Trials Centre, University of Sydney. She is Associate Convenor of the Cochrane Prospective Meta-Analysis Methods Group and board member of the Association for Interdisciplinary Meta-Research and Open Science. Kylie specialises in systematic reviews methodologies, such as individual participant data (IPD) and prospective meta-analysis (PMA), with a focus on obesity and neonatology.

What are some key resources for evidence synthesis?

- Aromataris E, Munn Z (Editors). JBI Manual for Evidence Synthesis. JBI, 2020. Available from https://synthesismanual.jbi.global. https://doi.org/10.46658/JBIMES-20-01

- Higgins JPT, Thomas J, Chandler J, et al., eds. Cochrane Handbook for Systematic Reviews of Interventions version 6.4: Cochrane, 2023. Available from www.training.cochrane.org/handbook.

- Hunter KE, Webster AC, Page MJ, et al. Searching clinical trials registers: guide for systematic reviewers. BMJ 2022; 377: e068791. https://doi.org/10.1136/bmj-2021-068791

- Hunter KE, Aberoumand M, Libesman S, et al. Development of the Individual Participant Data (IPD) Integrity Tool for assessing the integrity of randomised trials using individual participant data. medRxiv 2023.12.11.23299797; doi https://doi.org/10.1101/2023.12.11.23299797

- Hunter KE, Webster AC, Clarke M, Page MJ, Libesman S, Godolphin P, Aberoumand M, Rydzewska L, Wang R, Tan A, Li W, Mol BWJ, Willson M, Brown V, Palacios T, Seidler AL. Development of a checklist of standard items for processing individual participant data from randomised trials for meta-analyses: protocol for a modified e-Delphi study. PLOS One 2022;17(10):e0275893.https://doi.org/10.1371/journal.pone.0275893

- Riley RD, Tierney JF, Stewart LA, eds. Individual Participant Data Meta-Analysis: A Handbook for Healthcare Research, First Edition: John Wiley & Sons Ltd; 2021. https://www.ipdma.co.uk/textbook

- Seidler AL, Hunter KE, Cheyne S, Ghersi D, Berlin JA, Askie L et al. A guide to prospective meta-analysis BMJ 2019; 367 :l5342 https://doi.org/10.1136/bmj.l5342

- Sterne JAC , Savović J , Page MJ , et al. RoB 2: a revised tool for assessing risk of bias in randomised trials. BMJ 2019;366:l4898. https://www.bmj.com/content/366/bmj.l4898

- Stewart LA, Clarke M, Rovers M, et al. Preferred Reporting Items for a Systematic Review and Meta-analysis of Individual Participant Data: The PRISMA-IPD Statement. JAMA 2015; 313(16): 1657-65. https://jamanetwork.com/journals/jama/fullarticle/2279718

Scenario 2: Secondary analyses

What is a secondary analysis?

Secondary analyses involve using an existing dataset to answer a research question that is different from the research question of the original study. Types of secondary analyses may include:

- **Descriptive analyses,** which involve summarising the characteristics of a dataset using measures of frequency (e.g. counts, percentage), measures of central tendency (e.g. mean, median), and measures of variability (e.g. standard deviation, variance)
- **Health economic assessments,** which compare the costs and effectiveness of interventions
- **Prognostic and predictive analyses,** which aim to identify important prognostic or predictive factors of disease using modelling techniques
- **Power calculations for future trials,** which use effect sizes from existing datasets of similar studies to determine the sample size required to address a research question with sufficient confidence
- **Exploratory analyses,** which may explore trends or patterns in the data, associations or relationships between variables, biomarkers, mediatiors, or effectiveness among different subgroups of participants to generate new hypotheses
- **Hypothesis testing studies,** which involve testing pre-defined hypotheses generated from exploratory analyses above

Secondary analyses may also cover existing datasets being used as a comparator or control for a new study.

Definition	Using existing datasets to answer new research questions
Key steps	Develop protocol, obtain data, process and check data, conduct analysis, dissemination
Data types	Aggregate data (i.e. data that have been summarised across participants), or raw line-by-line data, known as individual participant data (IPD)
Data sources	The Health Data Australia (HDA) platform, trial investigators, clinical trials registers, journal websites, data repositories
Advantages	Maximise use of pre-existing data at little additional cost Inform sample size calculations to assist planning of new trials
Challenges	Obtaining data, generalisability/external validity

Types	Descriptive analyses, identification of important prognostic or predictive factors of disease, better understanding of disease history, informing sample size or power calculations for new study, hypothesis generating research questions about associations, biomarkers, mediations, effectiveness
Time	Depends on type and number of datasets
Expertise	Statistical expertise/support, data management

What data sources do I need to perform a secondary analysis?

The type of data required for secondary analyses depends on the type of research question that is being asked:

Research question: Descriptive analyses

Descriptive analyses require data from a sample that is assessed as likely representative of the population of interest. This can be the case in larger cluster-randomised trials, or whole-of-population trials (e.g. within a certain health district). Note that clinical trials are frequently not representative, so sample characteristics may differ from population characteristics.

Research question: Health economic assessments

Health economic assessments require data on costs of interventions and health outcomes, such as clinical effectiveness and health-related quality of life. Data could be sourced from one or multiple studies, noting that considerations of generalisability need to be undertaken based on the sample included in the original study/ies.

Research question: Identification of important prognostic or predictive factors of disease

Secondary analyses to identify prognostic or predictive factors require individual participant data from studies measuring factors and outcomes of interest. This could be one large study, or multiple studies, albeit considerations of generalisability need to be undertaken based on the sample included in the original studies.

Research question: Informing sample size through power calculations for new study

For secondary analyses to inform power calculations for a new study, ideally individual participant data would be obtained from studies measuring similar variables to the planned one, albeit specific aggregate data may be sufficient in some cases.

Research question: Hypothesis generating, e.g. about associations, biomarkers, mediators, effectiveness, etc.

For exploratory research questions, a subset of studies is usually sufficient. Note that findings then need to be validated in future studies.

What are the steps required to perform a secondary analysis?

Required steps depend on the type of secondary analyses planned, but typically would involve:

1. Determine the main objectives of the secondary analyses and scope availability of required data.
2. For descriptive analyses, health economic assessments, and prognostic/predictive studies, define the secondary analysis question(s) and detail planned methods in a protocol. The protocol should be reported in accordance with appropriate reporting guidelines (see EQUATOR network) and registered prospectively on open source platforms such as PROSPERO or OSF. For sample size calculations for a new study, report the methods used in the new study protocol that may either be uploaded to an open science platform such as OSF, uploaded as a pre-print or published in a peer-reviewed journal.
3. Check if ethical approval is required for the secondary analyses, or whether there are any ethical considerations.
4. Co-draft a data sharing agreement with the data provider, share with them your study protocol, ethics approval and data management plan, and adhere to any other specific legislative or regulatory requirements that their country or institution may have.
5. Where possible, it is recommended to give the data provider the opportunity to collaborate on your study, to acknowledge their contributions and benefit from their expertise and knowledge of the dataset. Offer authorship or acknowledgement if appropriate.
6. Request and obtain de-identified data and a data dictionary or codebook. Process, check and clean the data and clarify any uncertainties with the data provider to ensure accurate understanding.
7. Conduct analysis in accordance with pre-specified protocol. If possible, share your analytic code in a public repository, such as OSF.
8. Interpret, report and disseminate your findings according to appropriate guidelines (see EQUATOR network).

What are the advantages and considerations of secondary analysis?

Advantages of secondary analyses

- Answering new research questions using existing datasets increases the utility of the original data, for little cost, thereby reducing research waste
- Collaborating with data providers can improve the quality and impact of results, and encourage future synergies and efficiencies
- Using existing data for sample size calculations can inform resourcing and planning of new studies, and ensure sufficient statistical power to obtain meaningful results
- Existing data can be used to generate new hypotheses and advance research

Resources and funding

This may vary greatly depending on the type of secondary analyses, number of data sources, researcher experience and expertise. Using pre-existing data to answer new research questions can be much more efficient and cheaper than collecting new data. Depending on the number of data providers, it may take some time to finalise data sharing agreements, so this process should be initiated as early as is feasible. While sample size calculations may be relatively straightforward, more complex analyses, for instance generating advanced statistical models will be more time-consuming.

Potential challenges and strategies for mitigation

- Difficulties in obtaining data. These may arise due to reluctance on the part of the data provider, ethical approval issues, or local legislative or regulatory requirements. Accessing data via data sharing catalogues such as Health Data Australia may address some of these issues. Researchers can also try to proactively address any concerns of data providers, e.g. by sharing a data management plan detailing how participant privacy will be protected, and by demonstrating they have requisite skills and expertise to conduct analyses.
- Researchers may find it difficult to understand the provided data, particularly if a detailed codebook or dictionary are not provided. To mitigate this, try to obtain any relevant meta-data from the data provider and reach out with any queries to prevent misinterpretation.
- Datasets may not be generalisable, depending on the sample included. Attempts should be made to obtain representative data. Where this is not possible, generalisability should be noted as a potential limitation in any publications or reports of results.

What are some examples of secondary analysis?

Research question Descriptive analyses

Required data Dataset from a sample that is assessed as likely representative of the population of interest. This can be the case in larger cluster-randomised trials, or whole-of-population trials (e.g. within a certain health district). Note that clinical trials are frequently not representative, so sample characteristics may differ from population characteristics.

Research question Identification of important prognostic or predictive factors of disease, better understanding of disease history.

Required data Ideally individual participant data (IPD) from studies measuring similar variables to the planned one, albeit specific aggregate data may be sufficient in some cases.

Research question Informing sample size or power calculations for new study.

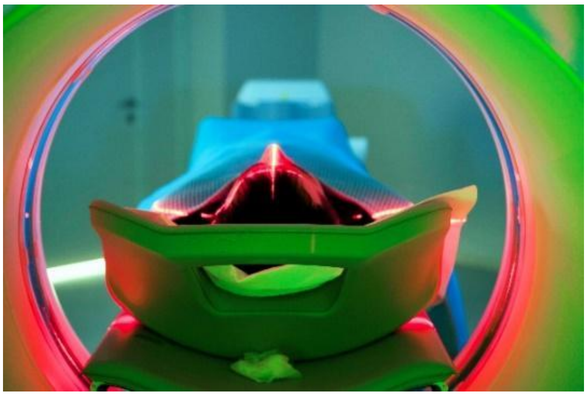
Required data Ideally IPD from studies measuring similar variables to the planned one, albeit specific aggregate data may be sufficient in some cases.

Research question Hypothesis generating research questions about associations, biomarkers, mediatiors, effectiveness, etc.

Required data Subset of studies sufficient, noting that findings then need to be validated in future studies.

Case Study: Secondary Analysis Scenario (Answer New Research Questions)

With Dr Angelina Tjokrowidjaja



Using existing trial datasets to determine the clinical accuracy of tumour marker blood test CA-125 versus CT imaging criteria to detect cancer progression in patients with ovarian cancer.

Study name: Poor concordance between Cancer Antigen-125 and RECIST Assessment for Progression in Patients with Platinum-Sensitive Relapsed Ovarian Cancer on Maintenance Therapy with a Poly(ADP-ribose) Polymerase Inhibitor

Start date: 2018 **End date:** 2023

Website: https://ascopubs.org/doi/full/10.1200/JCO.23.01182

What was this study about?

In women with ovarian cancer, CA-125 is a blood-based tumour marker widely used in low and high resource clinical settings to monitor for disease progression and this is reflected in current guidelines. However, CA-125 has only been validated as a biomarker of progression in the setting of chemotherapy and not for newer treatments, particularly poly(ADP-ribose) polymerase inhibitor (PARPi) therapy, which is a current standard of care for women with relapsed ovarian cancer.

In the primary trials of PARPi therapy, regular CT imaging was performed to diagnose disease progression and not CA-125. There are no evidence-based guidelines to inform clinicians on the optimal surveillance for women on PARPi therapy. Clinical practice is variable and some clinicians perform regular CA-125 while reserving CT imaging for rising CA-125 markers or symptoms concerning progression. Combining secondary data by pooling existing trial datasets enabled the research team to determine whether CA-125 can accurately detect disease progression in women with ovarian cancer on PARPi therapy.

What type of secondary data scenario is this case study?

This is a case study of using existing datasets for the purpose of secondary analyses to answer new research questions. The primary randomised controlled trials established maintenance PARPi therapy as a standard of care for women with relapsed ovarian cancer. By performing secondary analyses, this case study was able to answer the new research question: "W*hat is the concordance between CA-125 and CT criteria for disease progression in patients with relapsed ovarian cancer on maintenance PARPi therapy?*"

How was secondary data used in this study?

This work involved international collaboration between academia and industry, under the auspice of the Gynaecologic Cancer Intergroup, a collaborative research organisation involving more than 30 gynaecological cancer clinical trial groups worldwide.

What did this case study investigate? And what did they find?

This case study was a pooled analysis that included over 1,200 patients with relapsed ovarian cancer treated in 4 randomised controlled trials using contemporary targeted treatment with maintenance PARPi.

We found poor concordance between CA-125 and CT criteria for disease progression in these women. In particular, approximately one in two patients with radiologic progression did not have CA-125 progression and the majority of these patients had CA-125 levels that remained within the normal range. We also found that types of recurrence had different profiles. Cancer cells around the intestines gave rise to elevated CA-125 more than those from solitary organ (e.g. liver or lung) metastases. These findings demonstrate that CA-125 testing alone may not necessarily be a reliable marker to detect cancer growth and we should consider more periodic imaging to diagnose cancer progression.

What was the impact of the studies?

This case study's research in surveillance for women with relapsed ovarian cancer has generated practice-changing results with immediate implications on health practice. These findings were published in the Journal of Clinical Oncology here and received commentaries here and here.

These results inform national and international clinical guidelines to include periodic imaging as part of surveillance rather than relying on CA-125 tumour marker alone to detect disease progression, as the presence of normal CA-125 levels can give false reassurance to the oncologist and patient. Diagnosing disease progression at an earlier stage would allow patients to avoid continuing potentially futile treatment, toxicities and unnecessary costs associated with maintenance therapy. Given the importance of our findings, our case study translates to immediate uptake by oncologists in implementing periodic imaging for patients with ovarian cancer on contemporary PARPi therapy.

How did secondary data make this study and impact possible?

Previous primary studies established the efficacy of PARPi therapy in women with relapsed ovarian cancer. Using secondary data from these primary studies allowed us to address the new research question of whether CA-125 is a reliable surrogate for disease progression in this setting and its concordance with CT defined cancer progression.

Close collaboration with industry allowed us to access unpublished high-quality trial data from well conducted randomised trials. By pooling these data together, we were able to achieve more robust findings than would have arisen from a single trial. In addition, pooling previously unpublished subgroup data gave this case study greater power to detect differences in the concordance between CA-125 and CT criteria for cancer growth in particular patient subgroups, e.g. with different profiles according to the different sites of cancer growth.

How could Health Data Australia be used for this type of study?

This case study, led by Dr Angelina Tjokrowidjaja, showcases the value of data sharing to address new research questions and improve current health guidelines. However, there were several challenges encountered while undertaking this case study. Data sharing of randomised controlled trials requires close collaboration between academic groups, industry sponsors and trialists. Procedures to access data varied for each study and different platforms were required for data access and analyses.

By creating a more efficient way to share and manage data, Health Data Australia can foster new opportunities for research and mitigate the challenges faced by future researchers undertaking secondary analyses.

What are some key resources for secondary analysis?

- Chan A-W, Tetzlaff JM, Gøtzsche PC, et al. SPIRIT 2013 Explanation and Elaboration: Guidance for protocols of clinical trials. BMJ 2013;346:e7586. https://doi.org/10.1136/bmj.e7586
- Cooksey RW. Descriptive Statistics for Summarising Data. Illustrating Statistical Procedures: Finding Meaning in Quantitative Data. 2020 May 15:61–139. https://doi.org/10.1007%2F978-981-15-2537-7_5
- Hunter KE, Tan AC, Webster AC, et al. Responsibilities for receiving and using individual participant data. Cochrane Ev Synth. 2023; 1:e12028. doi:10.1002/cesm.12028

- Kent P, Cancelliere C, Boyle E, et al. A conceptual framework for prognostic research. BMC Med Res Methodol. 2020;20:172. https://doi.org/10.1186/s12874-020-01050-7
- Turner HC, Archer RA, Downey LE, et al. An Introduction to the Main Types of Economic Evaluations Used for Informing Priority Setting and Resource Allocation in Healthcare: Key Features, Uses, and Limitations. Front Public Health. 2021:9 https://doi.org/10.3389/fpubh.2021.722927

Scenario 3: Reproducibility, replication and validation

What is reproducibility, replication and validation with secondary data?

The main aim of reproducibility, replication and validation studies is to verify the accuracy, validity, and trustworthiness of scientific findings.

- In **reproducibility studies**, researchers re-analyse existing data from a previous study using the same methods in an attempt to verify the study's findings or uncover potential concerns.
- In **replication studies**, researchers attempt to replicate a previous study by applying the same methods to a different dataset (usually in a new study but at times accessed through a data repository) to see if results are comparable.
- In **validation studies**, researchers attempt to replicate a previously found effect estimate/ prediction model in a new dataset, e.g. to externally validate a model's predictive performance across settings, or determine if a treatment effect is generalisable to a different setting

Definition	Reproducibility: re-analysing data from an original study to verify its findings Replication: recreating an existing study to assess reliability of results Validation: attempts to recreate an effect estimate/ model in a new dataset
Key steps	Define study objectives, develop protocol, obtain data and research materials, re-create study (primary replication), re-analyse data according to original methods, assess consistency across studies, report and disseminate findings
Data types	Ideally individual participant data
Data sources	The Health Data Australia (HDA) platform, study investigators, clinical trials registers, journal websites, data repositories
Advantages	Fosters rigorous and transparent research practices, enhances confidence and credibility where research findings are found to be reproducible/replicable, enhances generalisability and external validity of findings.
Challenges	Obtaining data and detailed methodology, including analysis plan and coding Having sufficient resources and expertise to conduct study
Types	Reproducing statistical analyses, comparison of two or more studies (secondary replication), comparison of original study to a new replicated study (primary replication)
Time	Depends on complexity of original study, and whether replication is primary (more time consuming) or secondary
Expertise	Statistical expertise, topic experts, laboratory, or field staff for primary replication

What data sources do I need to perform reproducibility, replication and validation?

Data sharing catalogues such as HDA are an excellent resource to access data for reproducibility, replication, and validation studies. Data might also be obtained from publications, online repositories or via direct communication with study investigators.

For reproducibility studies, researchers require access to individual participant data from the original study, as well as detailed information about the methods of analysis used in the original study, including analysis code if possible.

For replication studies, ideally researchers would have access to individual participant data from the existing study that they wish to replicate, although aggregate data may be sufficient in some cases.

For validation studies, large individual participant datasets are ideal, e.g. from health catalogues, e-health records, or individual participant data meta-analysis.

What are the steps required to perform reproducibility, replication and validation studies?

At times, the specific question being addressed may use different steps or methods than below. The information provided is to be used as a guide only.

Reproducibility studies

1. Define the research question and the specific aspect of an original study that you plan to reproduce.
2. Select the original study and obtain individual participant data, data dictionary, statistical analysis plan, software code, and any other materials to enable you to reproduce the original study methods as closely as possible. These may need to be acquired directly from the original study investigators, and may require ethical approval, data sharing agreements and adherence to other local regulations.
3. Reproduce the methods and analyses applied in the original study as closely as possible using the same dataset, tools, and analysis techniques.
4. Compare the results obtained from the reproducibility study with results from the original study and assess the level of agreement. If there are inconsistencies, explore potential reasons.
5. Document, report and disseminate results.

Replication studies

1. Select the original study that you wish to replicate. Consider feasibility, topic of interest, and required expertise.
2. Define your research question and objectives.
3. Obtain individual participant data, data dictionary, statistical analysis plan, software code, and any other materials to enable you to replicate the original study as closely as possible. These may need to be acquired directly from the original study investigators, and may require ethical approval, data sharing agreements and adherence to other local regulations,
4. Develop a replication protocol that follows the original research methodology as closely as possible, and share with original study investigators to provide feedback if they wish. Be sure to describe any deviations from the original study, e.g. different measurement tools, different settings.
5. Pre-register the replication protocol on an open source platform such as OSF.
6. Conduct the replication study, including participant recruitment, experiment, data collection and analyses.
7. Compare the results obtained from the replication study with results from the original study and assess the level of agreement. If there are inconsistencies, explore potential reasons.
8. Document, report and disseminate results.

Alternatively, replication studies may involve attempts to replicate two pre-existing studies (secondary replications).

Validation studies

1. Select the effect estimate, model, or result that you wish to validate from an original study.
2. Develop a protocol or a statistical analysis plan, documenting your methods of validation. You may need access to the data or model from the original study, depending on your question and approach.
3. Obtain individual participant data and data dictionaries from separate study/ies that address a similar research question to the original study. Data may be accessed from health catalogues or repositories, where available. Alternatively, data may need to be acquired directly from the original study investigators, and may require ethical approval, data sharing agreements and adherence to other local regulations.
4. Analyse your data following your protocol or statistical analysis plan, comparing the results obtained from this validation with results from the original study.
5. Document, report and disseminate results.

What are the advantages and considerations of reproducibility, replication and validation?

Advantages of reproducibility, replication and validation studies

- Verification of previous research findings enhances confidence in these results, thereby contributing to the reliability and credibility of scientific research.
- Advances knowledge by identifying and correcting any errors, biases or limitations in previous research studies.
- Enhances generalisation and external validity of findings if they can be replicated across different settings and populations.
- Improves methodological conduct in research by encouraging transparency, rigour and adoption of best practices to facilitate reproducibility and replication (and therefore credibility of findings).

Resources and funding

Generally, resources and funding required will depend on the complexity, size and duration of the original study that is to be reproduced or replicated.

For both study types, resourcing is required to obtain original research data and materials (including any necessary agreements or approvals), and sufficient expertise, software, and skills are required to reproduce analyses, and then disseminate findings.

Primary replication studies are typically more resource intensive because they involve repeating an entire study experiment, including participant recruitment, data collection and analyses.

Potential challenges and strategies for mitigation

- It may be difficult to obtain data and research materials that are sufficiently detailed to enable reproducibility or replication of an original study as closely as possible. Deviations in data collection methods or analytical techniques may affect results, though robust research findings should hold up to slightly different methods. To mitigate this risk, researchers should attempt to obtain as much information as possible from original study investigators. Conversely, study investigators should be encouraged and supported to routinely make detailed protocols, analysis plans, data and coding openly available on a suitable repository or framework.
- Researchers may find it difficult to understand and implement original study methods, particularly if they are highly complex. To address this, they need to ensure appropriate expertise and skills are available among their team.

What are some real-world examples of reproducibility, replication and validation with secondary data?

Research question	If an existing dataset is re-analysed using the same methods and code as the original study, are the results obtained consistent?
Required data	IPD and detailed information about the methods of analysis used in the original study.
Research question	Do two or more studies addressing the same research question obtain consistent results?
Required data	Ideally IPD from an existing study that the researcher wishes to replicate, though aggregate data may be sufficient in some cases. The data for the new replication study can either be accessed through secondary data sources if a suitable dataset is available (i.e. data catalogue), or it can be 'de novo' conducted by the researcher.

What are some key resources for reproducibility, replication and validation with secondary data?

- Calin-Jageman R. Getting Started – A step-by-step guide to developing a replication project. Open Science Framework. 2016. https://osf.io/jx2td/wiki/Geting%20Started%20-%20A%20step-by-step%20guide%20to%20developing%20a%20replication%20project/

- Errington TM, Iorns E, Gunn W, et al. An open investigation of the reproducibility of cancer biology research. eLife. 2014;3:e04333. https://doi.org/10.7554%2FeLife.04333

- National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303.

- Riley RD, Ensor J, Snell KIE, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges BMJ 2016; 353 :i3140. https://doi.org/10.1136/bmj.i3140

Scenario 4: Education and Methods Development

What is secondary data use in education and methods development?

Pre-existing datasets are a valuable training resource to facilitate learning about data cleaning and analysis methods among students, researchers, and other interested parties. Upskilling such parties can enable implementation of scenarios 1, 2, and 3. Existing datasets may also be used to develop and demonstrate new statistical methods, and to inform machine learning and artificial intelligence algorithms.

Definition	Using existing datasets to: teach/learn about data cleaning and analysis methods, or develop and demonstrate new statistical methods
Key steps	Determine primary education,or methodological objectives, obtain appropriate dataset in adherence with regulations, use dataset for intended purpose
Data types	Dependant on learning outcomes, but usually individual participant data are required, particularly to allow greater depth of learning
Data sources	The Health Data Australia (HDA) platform, study investigators, clinical trials registers, journal websites, data repositories
Advantages	Improves data cleaning and analysis capacity among students and researchers, which supports confidence in findings, and facilitates data re-use for other purposes. Enables methodological developments
Challenges	Obtaining suitable data and navigating regulatory requirements to enable re-use for education and methods development purposes
Types	Education about data processing, cleaning, coding, analysis, including learning new software and tools for these purposes. Develop and demonstrate new statistical methods
Time	Can be adjusted to cater for student/teacher/research capacity, expertise, specific requirements, and key objectives
Expertise	The teacher/trainer/methodologist should be skilled in the methods they are disseminating and/or developing, and possess good communication skills. The learner may have any level of expertise, from novice to expert

What data sources do I need to use for secondary data for education and methods development?

Data sharing catalogues such as HDA are an excellent resource to access data for education and methods development purposes. Data might also be obtained from publications, online repositories or via direct communication with study investigators.

Most datasets may be suitable for teaching purposes, but it is important to consider potential for re-identification and other ethical considerations prior to sharing. Generally, any variables that may allow re-identification should be removed from the dataset.

If existing data are being used to inform new study design, then datasets of well-designed studies assessing similar research questions are required.

What are the steps required to use secondary data for education and methods development?

The ways in which data can be used in this space is vast. The steps provided here are to be used as a guide only.

1. Determine primary objective and target audience, e.g. teaching university students how to clean, and analyse data, training researchers to harmonise their studies with existing research to facilitate evidence synthesis.
2. Scope and obtain appropriate datasets from available sources, including data dictionaries, study protocol, analysis code and other helpful materials, where possible. Choose a topic area of relevance to your target audience or appropriate for your methodological objectives.
3. Ensure appropriate approvals or permissions are obtained to enable re-use of data for education or methods development purposes.
4. Ensure all data are de-identified to protect participant privacy and confidentiality.
5. Prepare and disseminate education or training activity (e.g. teach a tutorial at a university) or develop/demonstrate new statistical methods.
6. Evaluate teaching and identify areas for improvement, or test new methods on a broader range of datasets.

What are the advantages and considerations when using secondary data for education and methods development?

Advantages of data re-use for education and methods development

- Learn and improve expertise and competency in data processing, cleaning, and statistical analyses, leading to improved rigour of research studies
- Upskilling best practice research and analyses across various parties enables re-use of pre-existing data to generate new knowledge
- Improved confidence in findings of research studies that employ best practice methods

Resources and funding

Generally, few resources and funding are required to obtain data for education and methods development processes. However, the time required will vary depending on legislative, regulatory, and ethical, and other requirements of the specific data source. Further, data may need to be manipulated to ensure it is appropriate for the intended purpose, i.e. no re-identification is possible, and it is not too complex for the required training purposes.

Potential challenges and strategies for mitigation

- Difficulties in finding a dataset appropriate for the intended purpose. To overcome this, try not to be overly specific with the type of dataset you are looking for. Adjustments can always be made if necessary.
- Navigating legislative, regulatory, and ethical requirements may be tricky. Try to find an appropriate data source with the least ‘red tape’ to access, while ensuring key data sharing principles are adhered to (e.g. participant confidentiality). Datasets may also be re-used across several different classes, semesters, or training activities.

What are some real-world examples of secondary data used in education and methods development?

Teaching Teach students and researchers how to clean, and analyse data.

Required data: Most datasets may be suitable for teaching purposes, but it is important to consider potential re-identification and other ethical considerations prior to sharing. Generally, any variables that may allow re-identification should be removed from the dataset.

Inform new study design: Review dataset, including variable names, categories, etc in an existing trial to harmonise conduct and data collection for a new trial.

Required data: Datasets of well-designed studies assessing similar research questions.

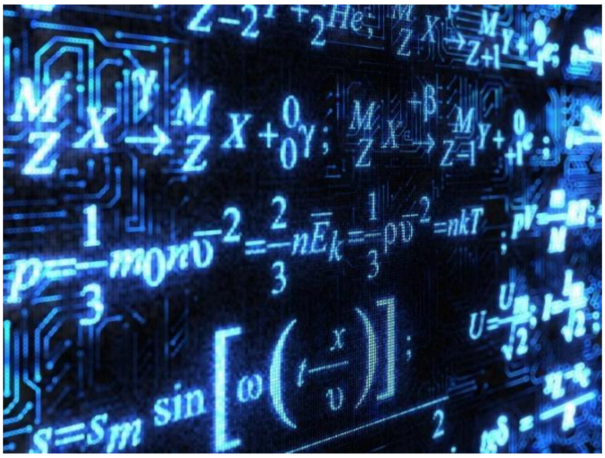
Methods development: Use existing data to develop, test, and demonstrate new statistical methods

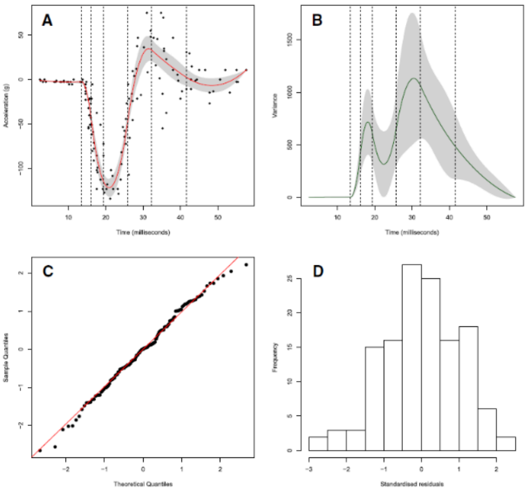
Required data: Ideally, a wide variety of datasets to demonstrate validity across disciplines

Case study: Education and Methods Development (Developing and Demonstrating New Statistical Methods)

With Dr Kristy Robledo

Sharing secondary data to develop, test, and demonstrate new statistical methods





Study name: A new algorithm for fitting semi-parametric variance regression models

Start date: 2016 **End date:** 2021

Website: https://doi.org/10.1007/s00180-021-01067-6

What was this study about?

In this study, secondary data were used to illustrate development of new statistical methods. In Stats 101, you may have learnt that constant variance (or spread) is a standard assumption in linear regression. Variance regression allows for changing (heterogeneous) variance, by allowing a researcher to fit a model to the variability of data, i.e. allowing the variability in a dataset to change depending on other factors. In this study a new statistical method was developed that is more stable and flexible than some other methods.

Let us consider an example. For a patient, having high blood pressure has been linked to cardiovascular disease, stroke, kidney damage… and the list goes on. However, the variability in a patient's blood pressure measurements has also been shown to be linked to increased risk of cardiovascular events. So, for a new treatment, it would be advantageous to investigate not just that it can reduce the average blood pressure of a patient, but also to reduce the variability of measurements. While there were existing ways to do this, this new method is (1) more reliable, (2) better able to handle complex data (like censored or truncated data), and (3) avoids some of the problems associated with older methods. The researchers tested this new method using simulations and real data, demonstrating it works well in many different situations.

What type of secondary data scenario is this case study?

This case study highlights the use of secondary data for education and methodological development. Specifically, existing datasets were utilised to illustrate to other researchers how to apply a new statistical method in 'real life', thereby advancing knowledge and learning in the field.

How was secondary data used in this study?

Secondary datasets were crucial for testing and applying the new algorithm. They allowed the researchers to assess how well their method performs across a range of different 'real-world' scenarios.

What did the study investigate? What were the key findings?

The study developed a new statistical algorithm and provided two different applications of how the method could be used. The method applies a model not just to the average (or mean), but also models the variability. The first dataset, the 'motorcycle crash dataset,' includes 133 measurements of the head acceleration of a crash dummy during a motorcycle crash, recorded in g units over time in milliseconds. The second dataset, the LIDAR dataset, consists of 221 measurements from an experiment using light to measure distances.

Using these datasets, the researchers were able to show that the algorithm could accurately identify a model across various sample sizes, and different data characteristics.

What was the impact of the study?

Methodological research is research on how we conduct research. It could be research on how we collect data (trial design), who to collect data from, how to analyse data (current example) or how to report it. Once a new method or idea has been developed, it is important to apply it to real world examples and provide other researchers with tutorials on how to use the new method. If this is done well, and across different clinical disciplines, then the new method will gain traction in these fields. In this example, using a better statistical method will not only allow researchers to model the variability of their data, but also modelling the variability allows for more accurate estimation of average effects. So even if the variability is not the main interest of researchers, modelling it appropriately enables more precise estimates of average effects.

This method is flexible, and allows researchers to investigate smoothed curves, and deal with complex data like biomarkers with upper or lower limits of detection (censoring). It also allows not just one, but many factors to be investigated. The new algorithm is available for public use through the VarReg package in R.

How did secondary data make this study and impact possible?

It is important to test new methods on a variety of real datasets, since 'made up' data cannot give researchers the same confidence in the validity of their new method. The availability of secondary data enabled the researchers to test their new algorithm on real-world data and evaluate its performance. The authors plan to continue testing their algorithm on other existing datasets and adapt it to better handle different types of complex data, such as non-normal distributions, or incomplete data. These refinements will also make running the algorithm more flexible and efficient. But it is often hard for researchers to find and access such a large variation of datasets needed to develop valid methods.

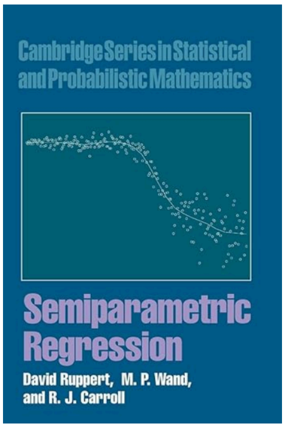
How could Health Data Australia be used for this type of study?

Health Data Australia is an excellent resource to find and access datasets for training, learning and methods development. It can help researchers identify and apply for access to a wide variety of datasets they may need to develop and test new statistical methods, ensuring they are robust, widely applicable and efficient. This case study highlights how secondary data can be instrumental in testing statistical methods across a broad range of data types.

Interview with Dr Kristy Robledo: Accessing data for methods purposes.

Why is it important to access a variety of real datasets for developing methods?

"For researchers to be able to use our new methods, they need to see that it works on real data, not just simulated data! So, having case studies is really important, and even better, case studies across different disciplines to show how the method can be used more broadly."



Cambridge Series in Statistical and Probabilistic Mathematics

Semiparametric Regression

David Ruppert, M. P. Wand, and R. J. Carroll

How did you go about finding datasets prior to the launch of Health Data Australia? What were the limitations of this approach?

"Textbooks. Other research papers in the area. Google isn't very useful for this. These two datasets are classic examples used for variance regression, as demonstrated by the LIDAR dataset being on the front page of a textbook called "Semiparametric Regression"!"

"This means that these datasets are very well known within this field – which is both positiveand negative. Sometimes we are after quite a specific example, and that's really hard to find."

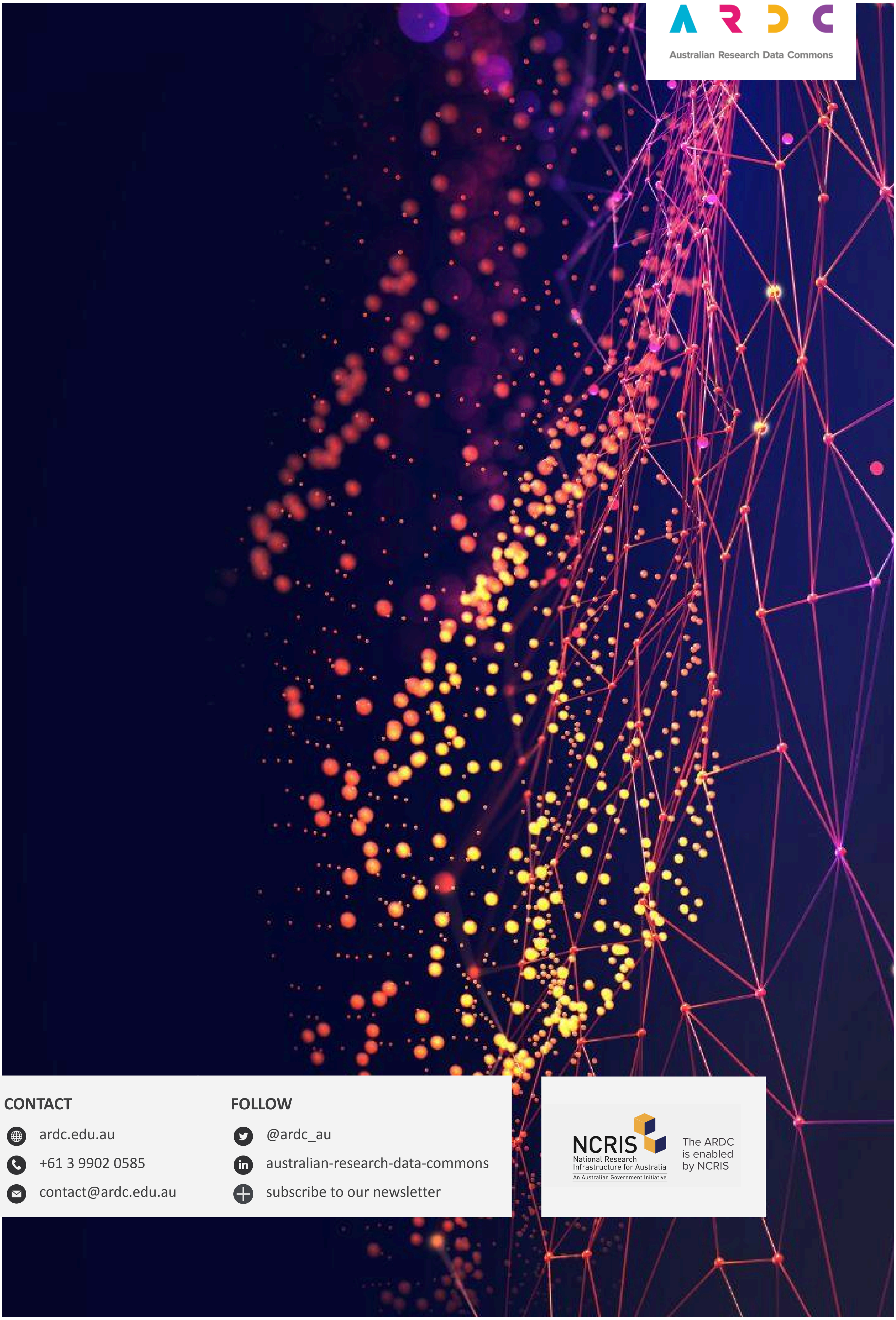
What do non-statisticians not understand about the importance of the development of new methods?

“It takes a reallllly long time! There are a lot of hours that go into developing the methodology, and then getting it coded up and working as it should. After that is the publication - that paper was under review for almost 12 months, at the second journal it was submitted to.”

What are some key resources for secondary data used in education methods development?

- Andrews DF, Herzberg AM. Data: A Collection of Problems From Many Fields for The Student and Research Worker. New York, NY: Springer; 1985.
- DASL – The Data And Story Library - https://dasl.datadescription.com/
- Lu Y, Web-Based Applets for Facilitating Simulations and Generating Randomized Datasets for Teaching Statistics, Journal of Statistics and Data Science Education, 2023;31:3, 264-272, https://doi.org/10.1080/26939169.2022.2146614
- Moreau D, Wiebels K. Ten simple rules for designing and conducting undergraduate replication projects. PLoS Comput Biol. 2023;19(3):e1010957. doi: 10.1371/journal.pcbi.1010957.
- R Built-in Data Sets - http://www.sthda.com/english/wiki/r-built-in-data-sets
- Robledo KP & Marschner IC. A new algorithm for fitting semi-parametric variance regression models. Computational Statistics, 2021; 36(4):2313-2335. https://doi.org/10.1007/

Australian Research Data Commons



CONTACT

ardc.edu.au

+61 3 9902 0585

contact@ardc.edu.au

FOLLOW

@ardc_au

australian-research-data-commons

subscribe to our newsletter

NCRIS
National Research
Infrastructure for Australia
An Australian Government Initiative

The ARDC
is enabled
by NCRIS