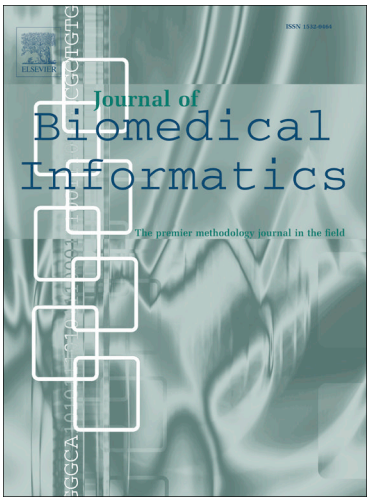
Flinders UNIVERSITY



Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges

Seong L. Cheah, Vanessa L. Scarf, Chris Rossiter, Charlene Thornton, Caroline S.E. Homer

PII: S1532-0464(19)30070-X
DOI: https://doi.org/10.1016/j.jbi.2019.103152
Article Number: 103152
Reference: YJBIN 103152

To appear in: *Journal of Biomedical Informatics*

Received Date: 26 December 2018
Revised Date: 14 March 2019
Accepted Date: 15 March 2019

Please cite this article as: Cheah, S.L., Scarf, V.L., Rossiter, C., Thornton, C., Homer, C.S.E., Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges, *Journal of Biomedical Informatics* (2019), doi: https://doi.org/10.1016/j.jbi.2019.103152

TITLE PAGE

Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges

Short title: Creating the first national linked dataset on perinatal and maternal outcomes in Australia

Seong L Cheah[a], Vanessa L Scarf*[a], Chris Rossiter[a], Charlene Thornton[b], Caroline S E Homer[a,c]

[a] Centre for Midwifery, Child and Family Health, Faculty of Health, University of Technology Sydney, NSW, Australia

[b] College of Nursing and Health Sciences, Flinders University, Adelaide, South Australia, Australia

[c] Burnet Institute, Melbourne, Victoria, Australia

* Corresponding author ORCID iD: 0000-0003-3503-9610

Email: vanessa.scarf@uts.edu.au (VS)

Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges

ABSTRACT

Background

Data linkage offers a powerful mechanism for examining healthcare outcomes across populations and can generate substantial robust datasets using routinely collected electronic data. However, it presents methodological challenges, especially in Australia where eight separate states and territories maintain health datasets. This study used linked data to investigate perinatal and maternal outcomes in relation to place of birth. It examined data from all eight jurisdictions regarding births planned in hospitals, birth centres and at home. Data linkage enabled the first Australia-wide dataset on birth outcomes. However, jurisdictional differences in data collection created challenges in obtaining comparable cohorts of women with similar low-risk pregnancies in all birth settings. The objective of this paper is to describe the techniques for managing previously linked data, and specifically for ensuring the resulting dataset contained only low-risk pregnancies.

Methods

This paper indicates the procedures for preparing and merging linked perinatal, inpatient and mortality data from different sources, providing technical guidance to address challenges arising in linked data study designs.

Results

We combined data from eight jurisdictions linking four collections of administrative healthcare and civil registration data. The merging process ensured that variables were consistent, compatible and relevant to study aims. To generate comparable cohorts for all three birth settings, we developed increasingly complex strategies to ensure that the dataset eliminated women with pregnancies at risk of complications during labour and birth. It was then possible to compare birth outcomes for comparable samples, enabling specific examination of the impact of birth setting on maternal and infant safety across Australia.

Conclusions

Data linkage is a valuable resource to enhance knowledge about birth outcomes from different settings, notwithstanding methodological challenges. Researchers can develop and share practical techniques to address these challenges. Study findings suggest that jurisdictions develop more consistent data collections to facilitate future data linkage.

KEYWORDS

Medical record linkage, pregnancy outcome, retrospective studies, pregnancy complications

ABBREVIATIONS

ABS	Australian Bureau of Statistics
ACT	Australian Capital Territory
APDC	Admitted Patient Data Collection
BIA	Birthplace in Australia (study)
DLU	Data linkage unit
ICD-10-AM	International Classification of Diseases – Australian modification. 10th edition
NICU	Neonatal intensive care unit
NSW	New South Wales
NT	Northern Territory
PDC	Perinatal Data Collection
PPN	Person project number
RBDM	Registry of Births, Deaths and Marriages
SA	South Australia
SPSS	Statistical Package for the Social Sciences
WA	Western Australia

Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges

INTRODUCTION

Different countries provide maternity services in a variety of ways. In Australia, most births (97.5%) take place in public or private hospitals. Other settings include birth centres (either attached to a hospital or stand-alone) (1.8% of births) or at home (0.2%) (1). In high-resource countries like Australia, healthy women giving birth generally have very good outcomes. Given the small number of adverse events overall, it is complicated to determine which place of birth is safest. Combining data from multiple states and territories or across several years is more likely to provide evidence about safety. Generating such combined data, however, involves complex methodological and technical challenges to optimise the quality of the evidence from which to guide policy decisions.

This paper presents methodological experiences in the *Birthplace in Australia: a population-based cohort study* (BIA), examining the perinatal and maternal outcomes from births planned in hospital obstetric units, birth centres and at home [de-identified reference]. This research was a nation-wide retrospective cohort study, combining linked data from Australia's eight jurisdictions (six states and two territories) for the period 2000-2012.

Similar large-scale studies examining outcomes by place of birth have been conducted in other high-income countries, including England (2), the Netherlands (3-6), Nordic countries (7, 8), Canada (9-12), the United States (13-15) and New Zealand (16, 17). Although previous Australian research has investigated outcomes related to place of birth in single states (18-20), none has attempted to examine outcomes for women nation-wide. In Australia, the

data collections and variables are not uniform across the country, presenting challenges to creating a standard national dataset.

Linking administrative data gathered over several years can generate the statistical power necessary to detect and compare rare outcomes such as perinatal mortality (21-23) or to examine health amongst vulnerable population groups (24). Data linkage combines electronic data from separate collections to amalgamate information about the same individual, facilitating research while maintaining privacy. However, its limitations include the time, technical intricacy and clerical burden involved, as well as concerns over the accuracy, consistency and comparability of data collected primarily for administrative purposes (25-32). Several studies of maternal or perinatal health and wellbeing have utilised data linkage, in Australia (33-36) and many other countries (37), particularly in Britain, the United States, and the Nordic countries.

Linking health-related data requires care, patience and expertise (24), especially linking maternal or neonatal datasets (22, 38, 39), and identifying and adjusting for errors and disparities (32, 33, 40-45). Some researchers have described data linkage techniques for perinatal health research within one Australian state (24, 46, 47), although compiling and merging maternity data across jurisdictions is less common (41, 48).

In Australia, state and territory governments are responsible for much healthcare delivery (via public hospitals and community health), monitoring and administration. The BIA study used data from up to four comprehensive data collections linked within each of Australia's eight jurisdictions, merging them into a robust dataset, containing information on multiple maternal and perinatal outcomes. However, the data screening and merging process spawned many methodological issues, related to the diversity of underlying data collections.

The BIA study required not only merging perinatal datasets, but also ensuring that the identified cohorts for the three different birth places (home, birth centre and hospital) were as similar as possible. This was imperative to eliminate potential variation from confounding factors that may affect safety other than the birth setting, especially differences in underlying risk factors that women may have which will influence their birth outcomes. To be eligible for home birth or birth centres, women must have a healthy pregnancy at low risk of complications. Therefore, women in the hospital cohort should have a comparable risk profile. It was important to exclude women with complex pregnancies from the hospital cohort, as they are likely to have higher rates of intervention and poorer outcomes than healthy women without any known risk factors. Thus, outcomes could be more reliably related to birth setting rather than the characteristics of the women in each cohort. This stipulation added further complexity to the already intricate process of combining multiple collections of linked data.

The objective of this paper is to describe data preparation techniques used to produce a national linked dataset on Australian women with low-risk pregnancies. It presents more detail on procedures than was feasible in the original BIA study report [de-identified]. Further, this paper aims to assist other researchers by highlighting and addressing problems encountered in managing and merging data from multiple sources in a complex healthcare system.

MATERIALS AND METHODS

The study retrospectively analysed routinely collected health data on women with low-risk pregnancies who gave birth between 2000 and 2012 (inclusively) in Australia. The BIA study

was approved by the [University] Human Research Ethics Committee (reference 2012000167).

Research questions

This paper addresses three research questions:

- How to merge already linked administrative data from eight states and territories?
- How to deal with inconsistencies in data collections?
- How to ensure a dataset that best represents women with low risk of obstetric complications, given the constraints of the underlying data collections?

Data sources

In Australia, data on healthcare service delivery and civil registrations are gathered at state/territory level. To examine maternal and perinatal outcomes comprehensively, we requested that each jurisdiction supply linked data from four sources:

- Perinatal Data Collections (PDC) collected by maternity providers
- Admitted Patient Data Collections (APDC) compiled by hospitals
- Registries of Births, Deaths and Marriages (RBDM) death registrations and
- Australian Bureau of Statistics (ABS) mortality data.

In Australia, data linkage uses probabilistic algorithms to match identifying data such as name, sex, date of birth and address across several databases. It allocates a match weight to similar pairs, representing the likelihood of a match (22, 32). These identifiers are separated from the health record to preserve privacy, and the health records are ultimately coupled with a randomly-assigned anonymous person project number (PPN) (22, 49). This method is

necessary because Australians do not possess a unique identifier that can be matched across datasets, which would facilitate a deterministic data linkage process (47).

The DLUs also provided a data linkage report on the quality of the linked data output, indicating the rate of false positive and false negative matching which all DLUs estimated to be around 0.5%.

While there is some consistency in data collection and linkage between jurisdictions, limited coordination in data management creates complications for handling and merging the state-based linked data into a readily accessible and efficient national database. Even though each jurisdiction recorded data on similar outcomes, these data are stored in discrete systems, use filenames of different formats, and comprise diverse variables with inconsistent value labels and contents. States also vary in quality assurance and linkage key protocols (41).

Access to linked datasets

Using a two-step process, we first applied to the Data Linkage Units (DLUs) in eight jurisdictions for linked data from the four data collections on specified variables. The Research Protocol assisted in completing disparate forms and ensuring that the data specifications remained consistent across the many data requests and ethics applications. We then applied for ethical approval from eight state-based ethics committees (except in the Australian Capital Territory [ACT], where ethical approval needed to precede application to the data custodian). We used a national ethics application process while complying with state-specific requirements such as the Confidentiality Agreement (in Western Australia - WA) and the Victorian Specific Module.

The aim was to create a standardised national dataset, with consistent and comparable variables. We had previously used linked data from four sources to explore perinatal and maternal outcomes within one state, New South Wales (NSW) [authors]. Using the NSW dataset as a template for the nation-wide master data collection, data were sequentially merged as they arrived from each state or territory, aligning them to the NSW variables. The filtering and screening processes further ensured that sample contained only women with low-risk pregnancies. This process took over two years and is described below.

Process undertaken

The following sections outline the approaches adopted to address the three research questions. Although presented here as three distinct and linear stages, in reality the process was more complicated and iterative. Linked data arrived from DLUs at various times and in widely differing formats. We loaded the state-based datasets into the Statistical Package for the Social Sciences (SPSS) v. 24 software, and cleaned, screened and merged data as they arrived while awaiting datasets from other jurisdictions.

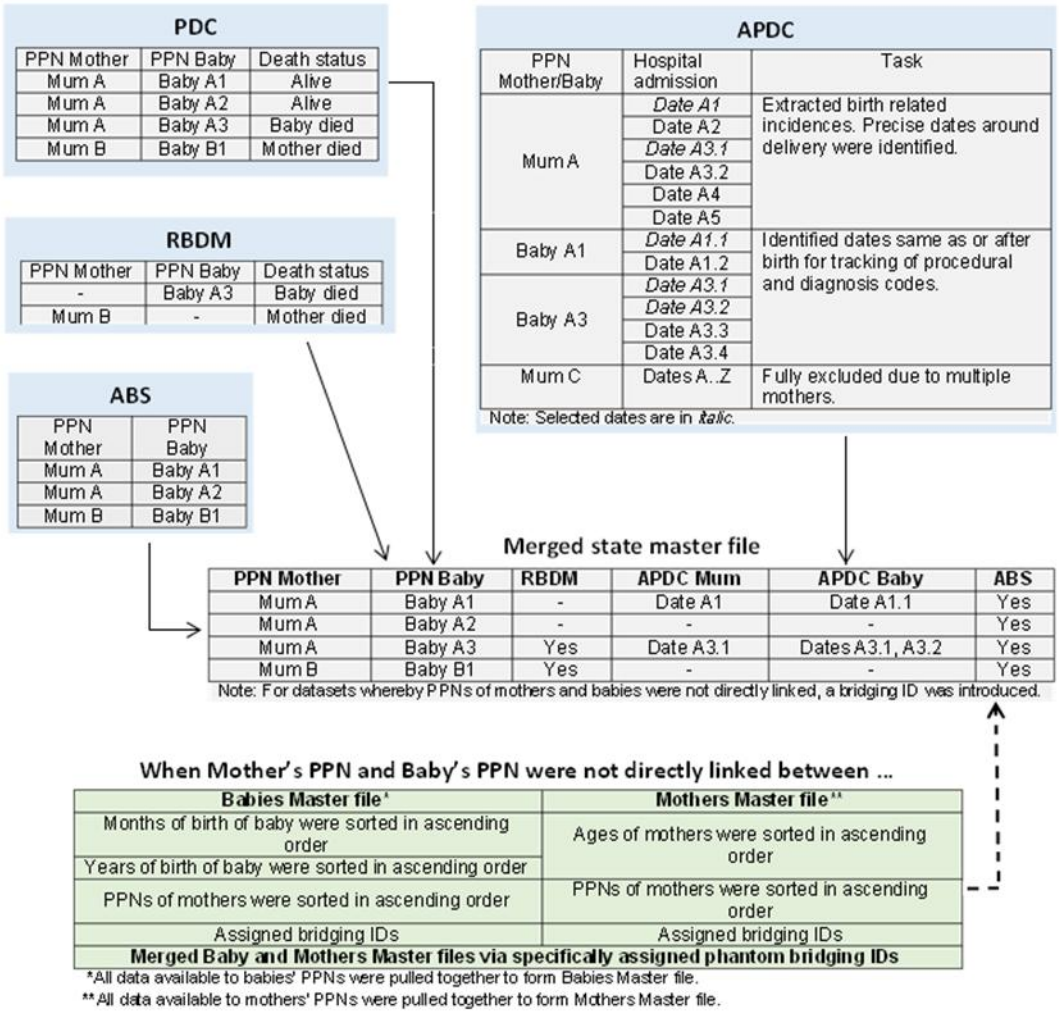
Stage 1: Cleaning and validating linked data

In order to develop an independent master file for each state, we extracted and merged the specially linked data from the PDC, APDC, ABS and RBDM collections (or state-based equivalents). PDC data for 2000-2012 (or available years) were linked with APDC data on hospital admissions for the nine months prior to the birth (mother) and twelve months after the birth (mother or baby), and with ABS and RBDM data on deaths up to twelve months after the birth (mother or baby). One state provided mortality data for the first six months only.

The DLUs generated two master files for each jurisdiction covering mothers and babies, although they generally assigned PPNs for mothers and babies to enable matching. Except for two jurisdictions, the Mothers' and Babies' master files could be directly merged to form a state master file, using the process illustrated in Figure 1. For situations where a mother's PPN and her baby's PPN were unmatched, we assigned a unique phantom ID to each of the records in the Mothers Master file and Babies Master file to match them within the state master file.

Basic cleaning during this stage involved validating and verifying demographic details across datasets, including age, gender, date of birth and date of death (some datasets only provided year and month). We compared consistency across all possible sources to minimise errors. For example, a particular PPN might contain gender or date of birth data which were the same in all datasets except one; in this case, we decided on the data with most occurrences. We applied these data cleaning and validation processes throughout the linked datasets for all variables of interest.

Figure 1: The process of merging PDC, APDC, ABS and RBDM datasets into a state master file



Missing data sometimes arose from changed versions in a particular data collection system and were reported accordingly. Linked data received from all sources were cleaned by eliminating duplicated data (e.g. similar records within one unique personal number) and cases that were clearly inaccurate or extreme (e.g. one child with several birth mothers). We validated data entries for discrepancies between sources (e.g. stillbirth with date of death), and determined value labels upon consensus among team members (e.g. regrouping of third degree and fourth degree perineal tears into a single variable) for subsequent merging into the master file. Owing to the large file size (often over 2 gigabytes), we generally commenced screening the linked data after applying some selection criteria beforehand

(e.g. spontaneous onset of labour, gestation between 37 and 41 weeks), to streamline cleaning and checking processes.

Having collated the state master file, we performed more robust data validation and verification to minimise discrepancies across contents with similar variable types. The DLUs assigned PPNs for mothers and babies to facilitate linkage. However, sometimes information on mothers and babies was independently stored in addition to the data with the PPN; this made merging the contents between mothers and babies more involved. Box 1 illustrates various potential errors we identified as arising from the linkage process, and indicating reasons for excluding some records. It indicates that when DLUs link data from several sources, some records may be repeated containing slightly different information.

Box 1: Potential scenarios in merging mother/baby data during validation

PPN mother*	Birth event	PPN baby*	DOB baby MMM/YYYY	Included in study?	Reason for exclusion
7654321	1	3456789	FEB/1999	No	Before 2000
7654321	2	5678934	JAN/2000	No	Multiple birth
7654321	2	6789543	JAN/2000	No	Multiple birth
7654321	3	4567893	DEC/2012	Yes	-
7654321	4	2345678	JAN/2013	No	After 2012
2345678	1	1234567	MAR/2004	No	Multiple mothers
3456712	1	1234567	MAR/2004	No	Multiple mothers

* Dummy PPNs used here as examples

We developed an approach for verifying records where one baby appeared to have several birth mothers using Algorithm 1. Initially, both variables PPN_Mum and PPN_Baby were sorted consecutively in ascending order, followed by an iterative one-step increment count

for Mum_Num when the condition of PPN_Baby equal to previous PPN_Baby *and* PPN_Mum not equal to previous PPN_Mum was met.

***Algorithm 1*: Identifying the number of mothers per baby**

```
*/Both PPN_Baby and PPN_Mum were consecutively sorted in ascending order.

SORT CASES BY PPN_Baby(A)PPN_mum(A).
COMPUTE MumNum=1.
IF (PPN_Baby = LAG(PPN_Baby)and PPN_mum<>LAG(PPN_mum))
Mum_Num=LAG(Mum_Num)+1.
EXECUTE.
FREQUENCIES VARIABLES=Mum_Num
 /ORDER=ANALYSIS.
```

Then Algorithm 2 provided a more comprehensive approach to detecting records with the same birth mothers for the same babies, same birth mothers for different babies, different birth mothers for the same baby and different birth mothers for different babies. When a baby appeared to have more than one birth mother, DiffMum>1, then that record was omitted.

***Algorithm 2*: Comprehensive combination between number of mothers and number of babies**

```
*/Both PPN_Mum and PPN_Baby were consecutively sorted in ascending order. Baby1 was
assigned to count the number of similar PPN_Baby.
SORT CASES BY PPN_baby(A)PPN_mum(A).
COMPUTE baby1=1.
IF (PPN_Baby = LAG(PPN_Baby))baby1=LAG(baby1)+1.
VARIABLE LABELS baby1 'baby1 order'.
EXECUTE.

*/Both PPN_Baby and PPN_Mum were consecutively sorted in ascending order and baby2
was assigned to count the number of similar PPN_Baby.
SORT CASES BY PPN_mum(A)PPN_Baby(A).
```

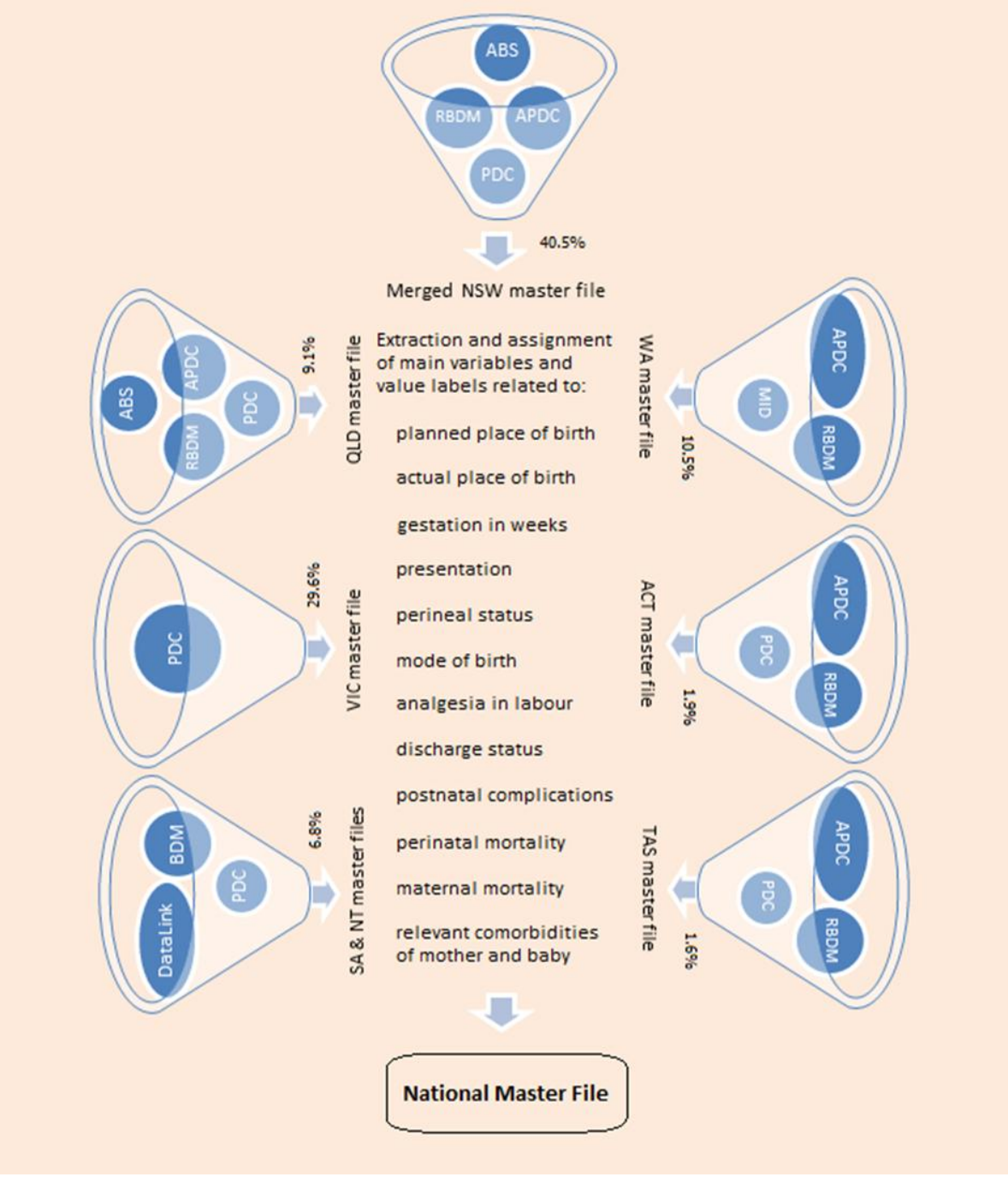
```
COMPUTE baby2=1.
IF (PPN_Baby = LAG(PPN_Baby))baby2=LAG(baby2)+1.
VARIABLE LABELS baby2 'baby2 order'.
EXECUTE.

*/Crosstabulating baby1 and baby2 to view the uniqueness in PPNs between baby and mother.
CROSSTABS
 /TABLES=baby1 BY baby2
 /FORMAT=AVALUE TABLES
 /CELLS=COUNT
 /COUNT ROUND CELL.

*/If DiffMum>1, a baby has more than one mother, and the associated record has to be excluded.
IF (baby1 = 2 & baby2 = 1)DiffMum=2.
EXECUTE.
FREQUENCIES VARIABLES=DiffMum
 /ORDER=ANALYSIS.
```

The study methodology was initially pilot-tested within one jurisdiction – NSW, the most populous state which accounts for approximately one-third (31.9%) of Australian births (1). This test used de-identified data from the NSW DLU, covering the four data collections (above) and the Register of Congenital Conditions. The original NSW master file sample comprised 258 161 women at low risk of complications who gave birth over an eight-year period (mid-2000 to mid-2008). Analysis of the linked data identified significant differences by planned place of birth in type of birth and interventions [de-identified]. The NSW dataset formed the master file for the national study, which adopted the NSW data variables and value labels.

Stage 2: Merging data from different states and territories

Using NSW linked data as the base for the national master file, data were sequentially merged from other state master files as they were received and verified (Figure 2).

Fig 2: Development of national master file, indicating data sources and proportion of sample



Note:

Maternal mortality – demise of mother between start of labour and 42 days postpartum, relating to the birth of a child.

Perinatal mortality – intrapartum fetal death resulting in a stillbirth OR death of a live born infant up to 7 days (early neonatal death) or between 8 and 28 days (late neonatal death).

Linked data collections varied between states and territories, often involving different variables or similar variables in different formats, requiring adjustment of the syntax to ensure that the values within data from every other state were compatible with NSW values. For example, what is called 'mode of birth' in one state, is termed 'birth method' or 'delivery' in others, with varying constituent values (see Box 2 for examples of the syntax required). Even within NSW, the components of the dataset changed during the study period. For example, the variable 'model of care' was introduced to the PDC in 2006.

Box 2: Syntax required for merging data on mode of birth

In **NSW**, there were two versions of the mode of birth ('delivery') variable (*deliv98* and *deliv2011)* with differing value labels. This called for a new set of value labels that would accommodate all options. We therefore developed the new variable *deliv* and proceeded to match similar values in data from other states or territories. For example, matching data to the value "6" for "Caesarean Section" required the following syntax to capture the data from other jurisdictions.

Northern Territory: '*IF CHAR.INDEX(delivery_BB,"CS")>0 Y_deliv2015new=6.'*

Queensland: '*IF CHAR.INDEX(delivery_BB,"CS")>0 deliv=6.'*

Tasmania: '*IF (INDEX(BIRTH_MODE_P,"Caes")>0 or INDEX(ModeBirth_O,"Caes")>0) deliv=6.'*

Victoria: '*RECODE BirthMethod_Ds ('6'=6) INTO deliv.*' and '*RECODE BirthMethod_old ('3'=6) ('4'=6) INTO deliv.*'

WA: '*IF (INDEX(mid_method,"03")=1) deliv=6.*'

For effective data management, we established various strategies to avoid unnecessary reshuffling of data or other cumbersome processes. In practice, analysts devise their own syntax strategy. Box 3 outlines some examples of the many techniques that we developed to simplify data and procedures.

Box 3: SPSS syntax for simplifying data handling

Tip	Syntax	Function
A	COMPUTE id=$CASENUM. FORMAT id (F8.0). EXECUTE.	Keeping all records intact for quick reference if necessary. Assigned before any work was done on the original dataset.
B	COMPUTE DiffY=(date_dth - dob) / (365.25 * time.days(1)). VARIABLE LABELS DiffY. VARIABLE LEVEL DiffY (SCALE). FORMATS DiffY (F8.2). VARIABLE WIDTH DiffY(8). EXECUTE.	Computing and validating the age provided in original dataset.
C	STRING DiagP (A3). COMPUTE DiagP=CHAR.SUBSTR(Diag_codeP,1,3).	Assigning substring of a principal diagnosis code up to 3 characters.

D	COMPUTE pdc_mumAgeOK1=TRUNC(pdc_mumAgeOK). VARIABLE LABELS pdc_mumAgeOK1 'Truncated to Year'. EXECUTE.	Truncating mother's age into year to be consistent across all states.
E	RECODE VAR1 (1=1) (2 thru 4=2) (5 thru 7=3) (8 thru Highest=4) INTO VAR2. EXECUTE.	Recoding a variable into another variable by regrouping the value contents.
F	IF CHAR.INDEX(delivery_BB,"Forc")>0 deliv=2. EXECUTE.	Recoding values within a string variable into numeric values of a new variable.

Tip A illustrates a technique for backtracking the original order of the dataset for quick reference on data entries; Tip B computes the mother's age if age is not provided or validates the age if given age is in doubt; Tip C extracts the first three characters of a principal diagnosis code to help categorise risk status; Tip D truncates mother's age into year of birth for consistency across all datasets (as maternal age data was presented inconsistently); Tip E recodes a series of value labels in one variable into another series of values in a new variable; Tip F captures contents that contain a certain string and assigns it into a new variable.

Stage 3: Identifying low-risk cohorts

The BIA study needed to minimise confounding variables in order to examine the specific impact of birth setting on perinatal safety and well-being. To ensure that the women in different birth place cohorts (hospital, birth centre, home birth) shared similar risk status, a

three-phase process filtered out most pregnancies likely to have moderate or high risk of complication for labour and birth.

Phase 1: Applying common inclusion criteria

In this phase, the initial inclusion criteria specified:

- Birth between 2000 and 2012 (inclusive)
- Singleton pregnancy
- Gestation between 37 weeks and 41 weeks + 6 days
- Spontaneous labour onset

The linked data received from DLUs varied considerably in the extent to which they addressed the risk factors requested. In one state 33.3% of data supplied met the study criteria for low-risk pregnancy, whereas in another 88.3% of records met the criteria. This variation does not indicate that states and territories varied markedly in the proportion of pregnancies which are relatively low-risk; rather it demonstrates the extent to which DLUs could filter and extract relevant data.

The first phase of filtering the linked data commenced before we merged each state master files into the national master file, not least to reduce the size of the unwieldy datasets and to facilitate computer execution speeds. This initial screening, either by DLUs or by the research team, resulted in eight separate state-based datasets of women who met the initial inclusion criteria, which we cumulatively combined.

Phase 2: Clarifying indicators of pregnancy risk

This phase eliminated cases with other factors that also contribute to complexity in pregnancy, by closely examining variables provided and the related value labels. This process excluded women who:

- Received no antenatal care
- Attempted a vaginal birth after one or more previous caesarean sections
- Had a breech presentation

Phase 3: Excluding complicating medical conditions

The third phase of screening for risk factors involved scrutinising APDC and PDC data (or equivalents) for International Classification of Diseases (ICD-10-AM) codes, to identify diagnoses and procedures related to pregnancy complications. We excluded women with high-risk O codes (Box 4) and/or infants with Q codes indicating congenital abnormalities in the current pregnancy. We generally used broader diagnosis categories (e.g. O14 Pre-eclampsia, rather than O14.0/1/2) to ensure consistency with earlier versions of the ICD, avoiding the potential inconsistencies in more refined categories in later versions.

Box 4: ICD10[1] codes indicating pregnancy complications

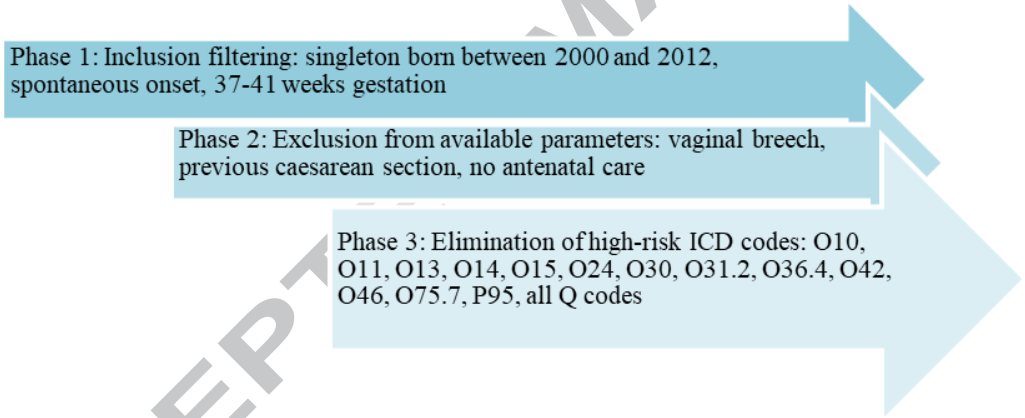
ICD-10-AM	**Diagnosis**
O10	Pre-existing hypertension complicating pregnancy, childbirth and the puerperium
O11	Pre-eclampsia superimposed on chronic hypertension
O13	Gestational [pregnancy-induced] hypertension
O14	Pre-eclampsia
O15	Eclampsia
O24	Diabetes mellitus in pregnancy
O30	Multiple gestation
O31.2	Continuing pregnancy after intrauterine death of one fetus or more

O36.4	Maternal care for intrauterine death
O42	Premature rupture of membranes
O46	Antepartum haemorrhage
O75.5	Delayed delivery after artificial rupture of membranes
O75.7	Vaginal delivery following previous caesarean section
P95	Fetal death of unspecified cause
Q codes	Reportable congenital abnormalities

[1] Australian Consortium for Classification Development *The International Statistical Classification of Diseases and Related Health Problems,* Tenth Revision, Australian Modification (ICD-10-AM). Darlinghurst NSW, Independent Hospital Pricing Authority, 2014.

The iterative three-phase screening process is illustrated in Figure 3.

Figure 3: Phased procedures to identify low-risk cohorts



Box 5 indicates the SPSS syntax used in the three phases of filtering, identifying different indicators of potential obstetric risk, in order to generate a sample of women whose pregnancies were low-risk regardless of their planned place of birth.

Box 5: SPSS syntax for identifying low-risk pregnancies

Filtering	**Syntax**	**Explanation**

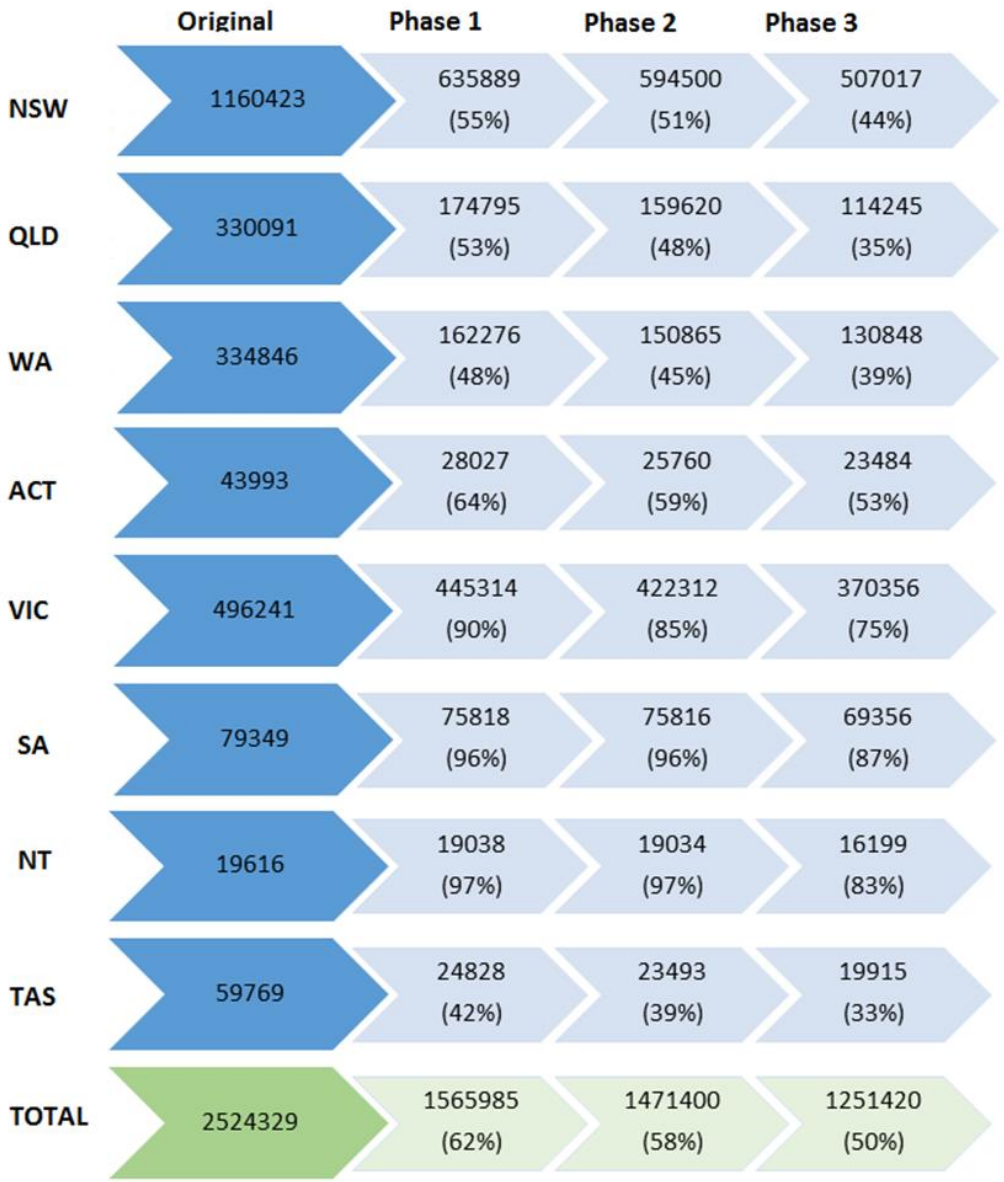
Phase 1	COMPUTE filter_$=((pdc_gestage >= 37 & pdc_gestage<42) & (Bdob_Yr >= 2000 & Bdob_Yr<=2012) & pdc_labons = 1 & Non_Singleton_OK=0). VARIABLE LABELS filter_$ '(pdc_gestage >= 37 & pdc_gestage<42) & (Bdob_Yr >= 2000 & Bdob_Yr <= 2012) & pdc_labons = 1 & Non_Singleton_OK=0 (FILTER)'. VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'. FORMATS filter_$ (f1.0). FILTER BY filter_$. EXECUTE.	Selected singleton born between 2000 and 2012 (inclusive), spontaneous onset of labour, gestation 37 to 41 weeks inclusive
Phase 2	IF Presentation = “Breech” or Delivery = “Breech” or prev_csbirth=”Yes” Exclude1=1.	Excluded vaginal breech and previous caesarean section
Phase 3	IF INDEX(Diag_code,"O10")>0 or INDEX(DIAG_code,"O11")>0 or INDEX(DIAG_code,"O13")>0 or INDEX(DIAG_code,"O14")>0 or INDEX(DIAG_code,"O15")>0 or INDEX(DIAG_code,"O24")>0 or INDEX(DIAG_code,"O30")>0 or INDEX(DIAG_code,"O31.2")>0 or INDEX(DIAG_code,"O36.4")>0 or INDEX(DIAG_code,"O42")>0 or INDEX(DIAG_code,"O46")>0 or INDEX(DIAG_code,"O75.7")>0 or INDEX(DIAG_code,"P95")>0 or INDEX(Diag_code,"Q")>0 Exclude2=1.	Eliminated high-risk ICD codes: O10, O11, O13, O14, O15, O24, O30, O31.2, O36.4, O42, O46, O75.7, P95, all Q codes

In order to scrutinise outcomes on perinatal mortality, we conducted a line-by-line investigation of individual de-identified records of all deaths from births planned in birth centres and at home, and a random one-in-ten sample of planned hospital births that resulted in mortality. This highlighted a number of factors that indicated additional risk; some of these had not been detected by previous screening processes and were then added to the codes applied to the whole dataset. Given the complexity of the data and the recognised gaps in administrative data for the purposes of health research (especially for home births), it was essential to carry out manual validation to complement aggregate data preparation.

RESULTS

Over the period investigated by the BIA study, there were 3 171 800 births across Australia (50). This figure represents all births in Australia 2007-2012, plus births 2000-2006 in NSW, Victoria, WA, SA, ACT and NT, and from 2005-2006 in Tasmania (Queensland and Tasmania provided data for limited time periods).

Figure 4 here: Refinement of sample size following three phases of data screening by state and territory



Phase 1: Inclusion filtering: singleton born between 2000 and 2012, spontaneous onset, 37-41 weeks gestation
Phase 2: Exclusion from available parameters: vaginal breech, previous caesarean section, no antenatal care
Phase 3: Elimination of high-risk ICD codes: O10, O11, O13, O14, O15, O24, O30, O31.2, O36.4, O42, O46, O75.7, P95, all Q codes

Overall, the BIA study received linked data about 2 524 329 births from the DLUs, which met the initial selection criteria to varying degrees. Figure 4 illustrates the effect of the various phases of screening and filtering, demonstrating how the sample size decreased with the increasing specificity of requirements for a low-risk cohort.

The various stages of cleaning, validating and screening records left 1,039,478 women remaining whose pregnancies were identified as low-risk by eliminating complicating conditions. The sample included data on outcomes from 1,251,420 births to these women between 2000 and 2012 (or 2005-2012 in Tasmania and 2007-2012 in Queensland). The final low-risk cohort contained half the number of records originally received, following the multiple screening processes.

Table 1 presents the sources of data received from each jurisdiction and its contribution to the final master dataset.

Table 1: Data sources, timeframe and contribution to national dataset, by state and territory

State or Territory	Data source	Years data available	Proportion of Australian births [51]	Proportion of final dataset
Australia Capital Territory (ACT)	ACT Death and Cause of Death Unit Record File ACT Perinatal Data Collection ACT Admitted Patients Care ACT Registrar of Births, Deaths and Marriages	13 years (2000-2012)	1.9%	1.9%
New South Wales (NSW)	NSW Perinatal Data Collection NSW Admitted Patient Data Collection NSW Register of Births, Deaths and Marriages Australian Bureau of Statistics - Mortality	13 years	31.9%	40.5%
Northern	SA NT DataLink	13 years	1.3%	1.3%

Territory (NT)	Perinatal Trends Inpatient Client Master Index Death Registry			
Queens-land	Queensland Perinatal Data Collection Queensland Health Admitted Patient Data Collection Queensland Registrar General Australian Bureau of Statistics - Deaths	6 years (2007-2012)	20.4%	9.1%
South Australia (SA)	SA NT DataLink Perinatal Maternal Family Link Birth, Death and Marriage	13 years	6.6%	5.5%
Tasmania	Tasmanian Data Linkage Unit Cause of Death Unit Record File	8 years (2005-2012)	1.9%	1.6%
Victoria	Victorian Perinatal Data Collection	13 years	25.1%	29.6%
Western Australia (WA)	Hospital Morbidity Data System Midwives Notification System - Mortality	13 years	10.9%	10.5%

Table 1 indicates that the two states that provided data covering a shorter timeframe were under-represented in the final national dataset. Two other states were accordingly over-represented. Figure 2 illustrates the complex process of developing the national master file from linked data from all states or territories. It indicates the sources for which data were available nationally.

DISCUSSION

The BIA study uniquely and ambitiously aimed to merge linked data from six states and two territories that used diverse methods for recording, storing and naming variables. The study entailed a complex and time-consuming process to verify, match, screen and clean data, and to ensure comparable cohorts. This paper has described several intricate procedures used to enhance the compatibility and integrity of data used to compare the outcomes for women with low-risk pregnancies who planned to give birth in three birth settings.

Data linkage is valuable for increasing the size and utility of datasets, especially to examine uncommon outcomes such as mortality (22) which are critical in evaluating the evidence on birthplace safety. However, merging linked data is beset with challenges arising from datasets that are not always complete or compatible, and whose original purpose is not research (31).

Linked data offer enormous potential for exploring birth outcomes, albeit with an added degree of difficulty in a federation such as Australia where states and territories are responsible for relevant data collections (52). To ascertain the outcomes related to the different birth settings, the BIA study endeavoured to distinguish three comparable cohorts of women with similar low-risk pregnancies. Inconsistency in variables and data quality hampered this task. Eliminating obstetric complexity from the sample necessitated increasing the technical and analytical complexity to generate the most appropriate dataset.

Refining the Australia-wide master file over the stages of data preparation resulted in a final dataset of more than 1 million women with low-risk pregnancies, approximately two-fifths (39.5%) of all births during the study period (50). This proportion highlights how the various

phases of data screening greatly reduced the study denominator by limiting the sample to low-risk pregnancies necessary for comparing outcomes meaningfully.

As well as reducing the number of women in the sample, combining linked data from eight jurisdictions also reduced the number of variables available for analysis. Differences between variables necessitated the most basic approach for the combined data. For example, not all jurisdictions distinguished between neonatal intensive care units and special care nurseries for unwell newborns, so we combined admissions to both facilities to ensure data consistency. Some states did not record data on specific items, meaning that we either did not report them or analysed a smaller sample for those variables; for example, maternal mortality outcome data were only available for six jurisdictions. At times, the process of adopting the lowest common denominator also potentially eroded the quality of data from some variables. Intended place of birth was a variable critical to our analysis as it is vital to exclude, for example, unintended home births without skilled birth attendants. However, in most states, women's intended place of birth was recorded at an unspecified time in the pregnancy and not necessarily updated in administrative records if the location changed. This is less ideal than documenting intention at onset of active labour given the potential for complicating conditions to occur throughout pregnancy. However, the other screening processes would have largely excluded those women who developed risk factors prior to labour onset and who therefore required a hospital birth.

Limitations

This paper is limited in presenting only some of the techniques used to generate accurate and appropriate data for the BIA study. Other minor procedures were necessary to increase the integrity of the sample. However, this paper aimed to address several important

considerations in using linked perinatal data in Australia and to provide guidance for future researchers attempting similar ventures.

The procedures to eliminate obstetric complexity in the dataset may have failed to detect all high-risk pregnancies as demonstrated by our final manual scrutiny of mortality data. A review of studies validating perinatal datasets suggested that diagnoses were less effectively recorded than procedures; it identified under-reporting of hypertension and diabetes amongst pregnant women (30). Further, the selected ICD-10-AM codes used for screening may have overlooked other complicating conditions. Given these data are not collected primarily for research, they are dependent on the quality of data collection and entry by healthcare professionals and administrative staff.

CONCLUSION

This paper has illustrated the unique contribution that merging linked data can make to understand the impact of planned place of birth on maternal and perinatal outcomes, while expanding on the challenges involved. By offering practical guidance to help overcome common difficulties, this study aims to contribute to knowledge and research practice using this methodology.

The complexity and variability of data encountered in this study highlight the urgent need for more effective, transparent and uniform methods to collect and share healthcare data across Australian states and territories.

REFERENCES

1. Australian Institute of Health and Welfare. Australia's mothers and babies 2015 – in brief. Perinatal statistics series no. 33. Cat no. PER 91. Canberra: AIHW; 2017.
2. Birthplace in England Collaborative Group. Perinatal and maternal outcomes by planned place of birth for healthy women with low risk pregnancies: the Birthplace in England national prospective cohort study. BMJ. 2011.
3. de Jonge A, Geerts CC, van der Goes BY, et al. Perinatal mortality and morbidity up to 28 days after birth among 743 070 low-risk planned home and hospital births: a cohort study based on three merged national perinatal databases. BJOG: An International Journal of Obstetrics and Gynaecology. 2015;122(5):720-8.
4. de Jonge A, Mesman J, Mannien J, et al. Severe adverse maternal outcomes among low risk women with planned home versus hospital births in the Netherlands: nationwide cohort study. BMJ. 2013.
5. van der Kooy J, Poeran J, de Graaf JP, Birnie E, Denktass S, Steegers EA, et al. Planned home compared with planned hospital births in the Netherlands: intrapartum and early neonatal death in low-risk pregnancies. Obstetrics & Gynecology. 2011;118(5):1037-46 10p.
6. Wiegerinck MM, van der Goes BY, Ravelli AC, van der Post JA, Klinkert J, Brandenbarg J, et al. Intrapartum and neonatal mortality in primary midwife-led and secondary obstetrician-led care in the Amsterdam region of the Netherlands: A retrospective cohort study. Midwifery. 2015;31(12):1168-76.
7. Blix E, Huitfeldt AS, Oian P, Straume B, Kumle M. Outcomes of planned home births and planned hospital births in low-risk women in Norway between 1990 and 2007: A retrospective cohort study. Sexual and Reproductive Healthcare. 2012;3(4):147-53.
8. Halfdansdottir B, Smarason AK, Olafsdottir OA, et al. Outcome of Planned Home and Hospital Births among Low-Risk Women in Iceland in 2005-2009: A Retrospective Cohort Study. Birth. 2015;42(1):16-26.
9. Hutton EK, Cappelletti A, Reitsma AH, Simioni J, Horne J, McGregor C, et al. Outcomes associated with planned place of birth among women with low-risk pregnancies. CMAJ: Canadian Medical Association Journal. 2016;188(5):E80-E90 11p.
10. Hutton EK, Reitsma AH, Kaufman K. Outcomes associated with planned home and planned hospital births in low-risk women attended by midwives in Ontario, Canada, 2003-2006: a retrospective cohort study. Birth: Issues in Perinatal Care. 2009;36(3):180-9 10p.
11. Janssen PA, Lee SK, Ryan EM, et al. Outcomes of planned home births versus planned hospital births after regulation of midwifery in British Columbia. Canadian Medical Association Journal (CMAJ). 2002;166(3):315-23.
12. Janssen PA, Saxell L, Page LA, et al. Outcomes of planned home birth with registered midwife versus planned hospital birth with midwife or physician. Canadian Medical Association Journal (CMAJ). 2009;181(6-7):377-83.
13. Pang JW, Heffelfinger JD, Huang GJ, Benedetti TJ, Weiss NS. Outcomes of planned home births in Washington State: 1989-1996. Obstetrics & Gynecology. 2002;100(2):253-9.
14. Wax JR, Pinette MG, Cartin A, Blackstone J. Maternal and newborn morbidity by birth facility among selected United States 2006 low-risk births. American Journal of Obstetrics & Gynecology. 2010;202(2):152.e1-5 1p.
15. Cheng YW, Snowden JM, King TL, Caughey AB. Selected perinatal outcomes associated with planned home births in the United States. American Journal of Obstetrics & Gynecology. 2013;209(4):325.e1-8 1p.
16. Davis D, Baddock S, Pairman S, Hunter M, Benn C, Wilson D, et al. Planned Place of Birth in New Zealand: Does it Affect Mode of Birth and Intervention Rates Among Low-Risk Women? Birth: Issues in Perinatal Care. 2011;38(2):111-9 9p.

17. Dixon L, Prileszky G, Guillilan K, Miller S, Anderson J. Place of birth and outcomes for a cohort of low risk women in New Zealand: A comparison with Birthplace England. New Zealand College of Midwives Journal. 2014(50):11-8 8p.
18. Homer CS, Thornton C, Scarf VL, Ellwood DA, Oats JJ, Foureur MJ, et al. Birthplace in New South Wales, Australia: an analysis of perinatal outcomes using routinely collected data. BMC Pregnancy & Childbirth. 2014;14:206.
19. Laws PJ, Tracy SK, Sullivan EA. Perinatal outcomes of women intending to give birth in birth centers in Australia. Birth. 2010;37(1):28-36.
20. Kennare RM, Keirse MJ, Tucker GR, Chan AC. Planned home and hospital births in South Australia, 1991-2006: differences in outcomes. Medical Journal of Australia. 2010;192(2):76-80.
21. Olsen O, Clausen JA. Planned hospital birth versus planned home birth. Cochrane Database of Systematic Reviews. 2012(9).
22. Méray N, Reitsma JB, Ravelli AC, Bonsel GJ. Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. Journal of clinical epidemiology. 2007;60(9):883. e1-. e11.
23. Ravelli AC, Tromp M, van Huis M, Steegers EA, Tamminga P, Eskes M, et al. Decreasing perinatal mortality in The Netherlands, 2000-2006: a record linkage study. Journal of Epidemiology & Community Health. 2009;63(9):761-5.
24. Hilder L, Walker JR, Levy MH, Sullivan EA. Preparing linked population data for research: cohort study of prisoner perinatal health outcomes. BMC Medical Research Methodology. 2016; 16:72.
25. Stanley F, Glauert R, McKenzie A, O'Donnell M. Can joined-up data lead to joined-up thinking? The Western Australian Developmental Pathways Project. Healthcare Policy. 2011;6(Sp):63.
26. Mitchell RJ, Cameron CM, McClure RJ, Williamson AM. Data linkage capabilities in Australia: practical issues identified by a Population Health Research Network 'Proof of Concept project'. Australian and New Zealand journal of public health. 2015;39(4):319-25.
27. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health services research and data linkages: issues, methods, and directions for the future. Health services research. 2010;45(5p2):1468-88.
28. Brook EL, Rosman DL, Holman CAJ. Public good through data linkage: measuring research outputs from the Western Australian Data Linkage System. Australian and New Zealand journal of public health. 2008;32(1):19-23.
29. Andrew NE, Sundararajan V, Thrift AG, Kilkenny MF, Katzenellenbogen J, Flack F, et al. Addressing the challenges of cross-jurisdictional data linkage between a national clinical quality registry and government-held health data. Australian and New Zealand journal of public health. 2016;40(5):436-42.
30. Lain SJ, Hadfield RM, Raynes-Greenow CH, Ford JB, Mealing NM, Algert CS, et al. Quality of data in perinatal population health databases: a systematic review. Medical care. 2012;50(4):e7-e20.
31. van Walraven C, Austin P. Administrative database research has unique characteristics that can risk biased results. Journal of Clinical Epidemiology. 2012;65(2):126-31.
32. Harron K, Wade A, Gilbert R, Muller-Pebody B, Goldstein H. Evaluating bias due to data linkage error in electronic healthcare records. BMC medical research methodology. 2014;14(1):36.
33. Miller JE, Hammond GC, Strunk T, Moore HC, Leonard H, Carter KW, et al. Association of gestational age and growth measures at birth with infection-related admissions to hospital throughout childhood: a population-based, data-linkage study from Western Australia. The Lancet Infectious Diseases. 2016;16(8):952-61.
34. Khambalia AZ, Algert CS, Bowen JR, Collie RJ, Roberts CL. Long-term outcomes for large for gestational age infants born at term. Journal of Paediatrics and Child Health. 2017;53(9):876-81.
35. Xu F, Austin M-P, Reilly N, Hilder L, Sullivan EA. Major depressive disorder in the perinatal period: using data linkage to inform perinatal mental health policy. Archives Of Women's Mental Health. 2012;15(5):333-41.

36. Dahlen HG, Tracy S, Tracy M, Bisits A, Brown C, Thornton C. Rates of obstetric intervention and associated perinatal mortality and morbidity among low-risk women giving birth in private and public hospitals in NSW (2000-2008): a linked data population-based cohort study. BMJ Open. 2014;4(5):e004551-e.
37. Delnord M, Szamotulska K, Hindori-Mohangoo A, Blondel B, Macfarlane A, Dattani N, et al. Linking databases on perinatal health: a review of the literature and current practices in Europe. The European Journal of Public Health. 2016;26(3):422-30.
38. Holian J. Client and birth record linkage: a method, biases, and lessons. Evaluation Practice. 1996;17(3):227-35.
39. Hall ES, Goyal NK, Ammerman RT, Miller MM, Jones DE, Short JA, et al. Development of a linked perinatal data resource from state administrative and community-based program data. Maternal And Child Health Journal. 2014;18(1):316-25.
40. Leiss JK. A new method for measuring misclassification of maternal sets in maternally linked birth records: true and false linkage proportions. Maternal And Child Health Journal. 2007;11(3):293-300.
41. Tran DT, Havard A, Jorm LR. Data cleaning and management protocols for linked perinatal research data: a good practice example from the Smoking MUMS (Maternal Use of Medications and Safety) Study. BMC Medical Research Methodology. 2017;17(1):97.
42. Gialamas A, Pilkington R, Berry J, Scalzi D, Gibson O, Brown A, et al. Identification of Aboriginal children using linked administrative data: Consequences for measuring inequalities. Journal Of Paediatrics And Child Health. 2016;52(5):534-40.
43. Morgan AS, Marlow N, Costeloe K, Draper ES. Investigating increased admissions to neonatal intensive care in England between 1995 and 2006: data linkage study using Hospital Episode Statistics. BMC medical research methodology. 2016;16(1):57.
44. Nedkoff L, Knuiman M, Hung J, Sanfilippo FM, Katzenellenbogen JM, Briffa TG. Concordance between administrative health data and medical records for diabetes status in coronary heart disease patients: a retrospective linked data study. BMC medical research methodology. 2013;13(1):121.
45. Tromp M, Ravelli AC, Bonsel GJ, Hasman A, Reitsma JB. Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. Journal of clinical epidemiology. 2011;64(5):565-72.
46. Xu F, Hilder L, Austin MP, Sullivan EA. Data preparation techniques for a perinatal psychiatric study based on linked data. BMC Medical Research Methodology. 2012;12:71.
47. Bentley JP, Ford JB, Taylor LK, Irvine KA, Roberts CL. Investigating linkage rates among probabilistically linked birth and hospitalization records. BMC medical research methodology. 2012;12(1):149.
48. Moore HC, Guiver T, Woollacott A, de Klerk N, Gidding HF. Establishing a process for conducting cross-jurisdictional record linkage in Australia. Australian and New Zealand journal of public health. 2016;40(2):159-64.
49. Kelman CW, Bass AJ, Holman C. Research use of linked health data—a best practice protocol. Australian and New Zealand journal of public health. 2002;26(3):251-5.
50. Australian Bureau of Statistics. Births, Australia 3301.0. Various years.
51. Hilder L, Zhichao Z, Parker M, Jahan S, Chambers GM 2014. Australia's mothers and babies 2012. Perinatal statistics series no. 30. Cat. no. PER 69. Canberra: AIHW.
52. Mitchell RJ, Cameron CM, Bambach MR. Data linkage for injury surveillance and research in Australia: perils, pitfalls and potential. Australian And New Zealand Journal Of Public Health. 2014;38(3):275-80.

Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges

FINANCIAL DISCLOSURE STATEMENT

This study was funded by the National Health and Medical Research Council of Australia, Project Grant ID APP1022422. Principal Investigators: Professor Caroline Homer, Professor Hannah Dahlen, Professor David Ellwood, Professor Della Forster, Professor Maralyn Foureur, Professor Helen McLachlan, Professor Jeremy Oats and Professor David Sibbritt.

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

Competing interests

The authors declare that they have no competing interests

Ethics approval and consent to participate

The study was approved by the Human Research Ethics Committee at the University of Technology Sydney, Reference Number 2012000167.

Each state and territory also provided ethical approval as outlined below.

State or territory	Name of the ethics committee(s)	Number/ID of the approval
Australian Capital Territory	ACT Health - Human Research Ethics Committee	ETH 3.14.061
New South Wales	NSW Population and Health Services Human Research Ethics Committee	HREC/14/CIPHS/15
Northern Territory	Department of Health of the Northern Territory and the Menzies School of Health Research - Human Research Ethics Committee	HREC 2014-2247
Queensland	Qld Health: Office of the Human Research Ethics Committee	HREC/14/QGC/175
South Australia	South Australian Health - Human Research Ethics Committee	HREC/14/SAH/117
Tasmania	Human Research Ethics Committee – Tasmania Network	Ref No: H0015023
Victoria	Department of Health and Human Services – Consultative Council on Obstetric and Paediatric Mortality and	Ref: 14/12

	Morbidity (CCOPMM)	
Western Australia	Government of Western Australia, Department of Health, Human Research Ethics Committee	HREC 2014/57

Authors' contributions

Author contributions

Formal analysis: SC
Methodology: SC, VS, CR, CT, CH
Project administration: VS, SC, CH
Supervision: CH
Writing – original draft preparation: SC, VS, CR
Writing – review and editing: SC, VS, CR, CT, CH

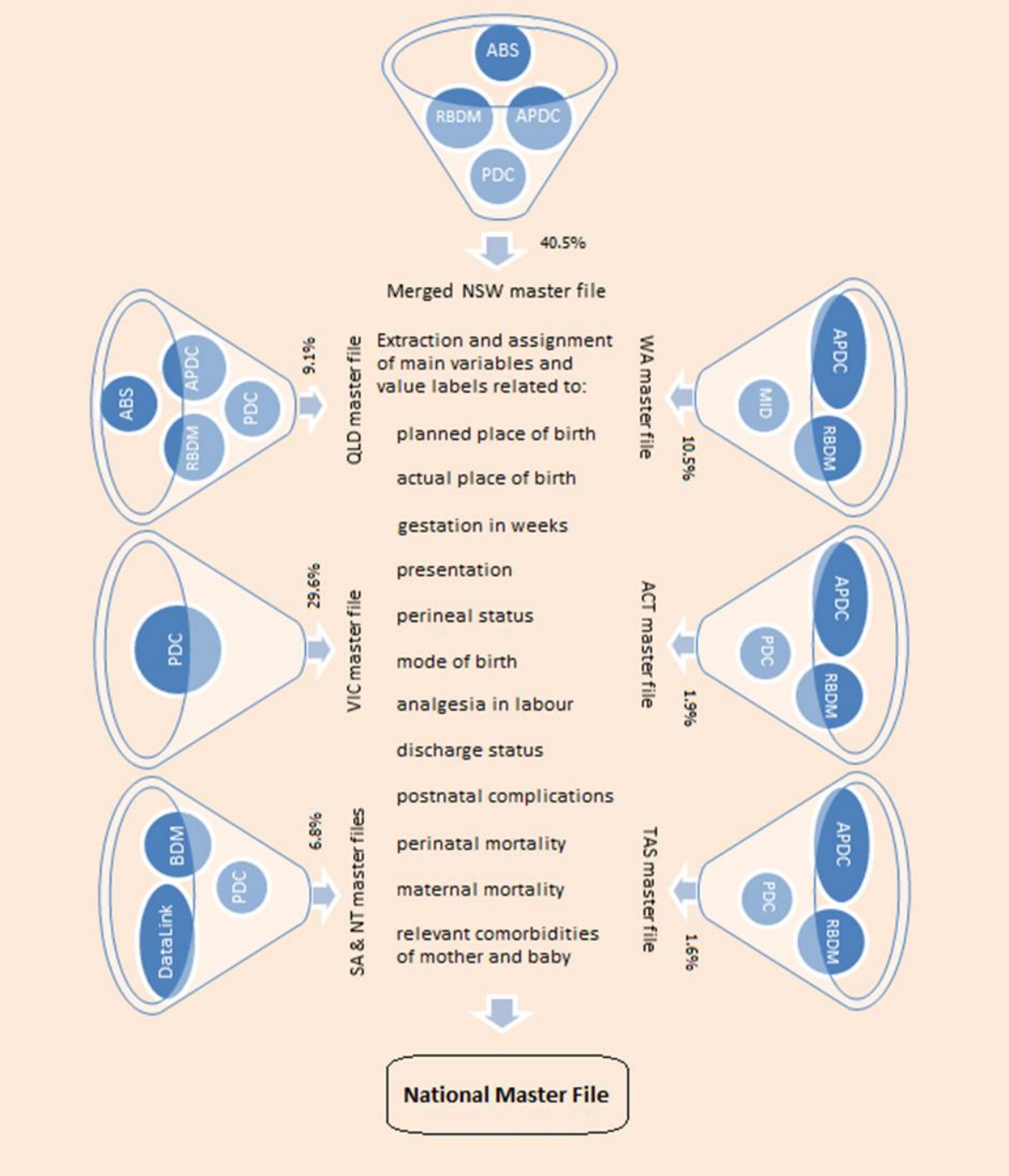
Acknowledgements

The authors are very grateful to the data custodians and helpful and resourceful staff at the data linkage units in all Australian states and territories for providing the de-identified data which formed the basis of this study. We thank the Centre for Health Research Linkage (CHeReL), NSW and ACT; Data Linkage Branch, WA; SA-NT DataLink; Tasmanian Data Linkage Unit; Data Linkage Queensland; Consultative Council on Obstetric and Paediatric Mortality and Morbidity (CCOPMM), Victoria for providing the de-identified data that formed the basis of this study. We would also like to acknowledge the assistance from the Centre for Data Linkage at Curtin University in planning the complex data application process. We thank the NHMRC for funding the study and our colleagues on the Birthplace in Australia

research team: Professor Hannah Dahlen, Professor David Ellwood, Professor Della Forster, Professor Maralyn Foureur, Professor Helen McLachlan, Professor Jeremy Oats and Professor David Sibbritt.

Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges

GRAPHICAL ABSTRACT



Creating the first national linked dataset on perinatal and maternal outcomes in Australia: Methods and challenges

HIGHLIGHTS

- This describes the first Australia-wide perinatal data linkage project
- We successfully compared outcomes from hospitals, birth centres and homebirths
- Comparisons require cohorts with equivalent risk profiles from linked datasets
- Variation in state-based data collection creates challenge to national data linkage