children

MDPI

Article

Improving Cohort-Hospital Matching Accuracy through Standardization and Validation of Participant Identifiable Information

Yanhong Jessika Hu [1,2], Anna Fedyukova [1], Jing Wang [1,2], Joanne M. Said [1,3,4], Niranjan Thomas [3,4], Elizabeth Noble [4], Jeanie L. Y. Cheong [1,3,5], Bill Karanatsios [6], Sharon Goldfeld [1,7] and Melissa Wake [1,2,*]

1 Murdoch Children's Research Institute, The Royal Children's Hospital, Parkville, VIC 3052, Australia
2 Department of Pediatrics, The University of Melbourne, Parkville, VIC 3052, Australia
3 Department of Obstetrics and Gynaecology, The University of Melbourne, Parkville, VIC 3010, Australia
4 Maternal Fetal Medicine, Joan Kirner Women's & Children's at Sunshine Hospital, St Albans, VIC 3021, Australia
5 Newborn Research, The Royal Women's Hospital, Parkville, VIC 3052, Australia
6 Western Health Chronic Disease Alliance, Western Health, St Albans, VIC 3021, Australia
7 Centre for Community Child Health, The Royal Children's Hospital, Parkville, VIC 3052, Australia
* Correspondence: melissa.wake@mcri.edu.au

Abstract: Linking very large, consented birth cohorts to birthing hospitals clinical data could elucidate the lifecourse outcomes of health care and exposures during the pregnancy, birth and newborn periods. Unfortunately, cohort personally identifiable information (PII) often does not include unique identifier numbers, presenting matching challenges. To develop optimized cohort matching to birthing hospital clinical records, this pilot drew on a one-year (December 2020–December 2021) cohort for a single Australian birthing hospital participating in the whole-of-state Generation Victoria (GenV) study. For 1819 consented mother-baby pairs and 58 additional babies (whose mothers were not themselves participating), we tested the accuracy and effort of various approaches to matching. We selected demographic variables drawn from names, DOB, sex, telephone, address (and birth order for multiple births). After variable standardization and validation, accuracy rose from 10% to 99% using a deterministic-rule-based approach in 10 steps. Using cohort-specific modifications of the Australian Statistical Linkage Key (SLK-581), it took only 3 steps to reach 97% (SLK-5881) and 98% (SLK-5881.1) accuracy. We conclude that our SLK-5881 process could safely and efficiently achieve high accuracy at the population level for future birth cohort-birth hospital matching in the absence of unique identifier numbers.

Keywords: birth cohort; hospital; data linkage; pregnant women; newborn; hospital records; information retrieval; personally identifiable information; data accuracy; demographics

Citation: Hu, Y.J.; Fedyukova, A.; Wang, J.; Said, J.M.; Thomas, N.; Noble, E.; Cheong, J.L.Y.; Karanatsios, B.; Goldfeld, S.; Wake, M. Improving Cohort-Hospital Matching Accuracy through Standardization and Validation of Participant Identifiable Information. *Children* **2022**, *9*, 1916. https://doi.org/10.3390/children9121916

Academic Editor: Henry C. Lee

Received: 15 October 2022
Accepted: 3 December 2022
Published: 7 December 2022

1. Introduction

Health services research using de-identified methods to link administrative datasets, typically undertaken with ethical approval for a waiver of consent, can generate powerful population insights to help shape care and policy [1,2]. Such linkage often uses a unique national administrative identifier, such as Scandinavian countries' unique personal identity number [3], US social security number [1], UK National Health Service (NHS) and National Insurance numbers [4], and Australia's national health insurance number (Medicare) [5]. Because linkage via these unique identifiers is highly accurate, it tends to dominate the published data linkage literature [5–7].

We propose that more, and more impactful, health services research could be undertaken if very large, consented birth cohorts were able to readily link to clinical data, including from hospitals. Firstly, such cohorts typically collect molecular, phenotypic, and participant-reported data that administrative datasets lack. Secondly, hospitals clinical data

may be especially important for the pregnancy and newborn periods, as (1) these are the only life points at which hospital care is essentially universal in many countries [3,6,8–10]; (2) there is substantial variation in care [11]; and (3) these periods shape subsequent outcomes for the entire maternal and child life course [12]. By accessing deep clinical data, consented cohorts could fill many gaps in both predicting health care and understanding its impacts on health and wellbeing [13]. However, other than minimum administrative datasets, most elements of Australian hospital care are neither available for research nor in a standardized format; for example, hospitals vary widely in how they record prescriptions and pathology results [5] and whether/how they collate them. Thus, data extraction may need to be on a hospital-by-hospital basis with subsequent standardization via common data models.

High matching rates between cohorts and hospital records have been achieved when unique identifiers are available, for example true matching rates above 97% for a UK study [14] and nearly 100% for a Hong Kong study [10]. Unfortunately, many consented cohorts do not hold hospital unique identifiers; participants may not know them, and it may not be easily possible to obtain them from services due to privacy and legal considerations [4,5]. This poses challenges for accurate matching, which then relies on participants' details such as name, postcode and date of birth. Such non-unique identifiers can facilitate linkage but also lead to linkage error and uncertainty [15]. While large-scale anonymized analyses may absorb some uncertain matches or mismatches, higher accuracy may be needed when consented clinical data permanently enter a major dataset designed to address future research questions, especially if return of results is part of the cohort's framework. Consequently, even a small improvement in matching accuracy could improve the health services research utility of population-based cohorts.

However, it is not clear how best to achieve high matching rates for cohort studies in the absence of a hospital unique identifier. Demographic identifiers vary in format and details both among cohorts and among hospitals. In Australia, the GRHANITE^{TM} Linkage Tool reached high sensitivity (95–100%) for clinical and pathology service data for linkage including Medicare number, but sensitivity dropped to only 66% in the absence of Medicare number [16]. The Australian Institute of Health and Welfare has created Statistical Linkage Key (SLK-581) to link hospital and death records through a probabilistic strategy, achieving matching rates of 97.5% [17]. However, this strategy may not be suited to deterministic matching, matching at an individual level, or where names differ significantly between two datasets (which is more likely at the start than the end of life) or large missing values [18].

We address this gap with the Generation Victoria (GenV) cohort, a statewide population-based birth cohort now open to all newborn babies and their mothers over two years in all 58 birthing hospitals in the state of Victoria (population 6.5 million), Australia [19]. Consent includes permission to bring into GenV information and samples that services already collect in clinical practice. In a one-year GenV sub-cohort from a single birthing hospital, this pilot study aimed to (1) compare different approaches to improve patients-participants matching accuracy after standardization and validation, and (2) recommend an approach for statewide scale up.

2. Materials and Methods

We conducted this data linkage matching report in accordance with reporting guidelines for studies involving data linkage [20] and in line with REporting of studies Conducted using Observational Routinely-collected Data (RECORD) guideline [21] (see Supplementary Tables S4 and S5). The first step was for the hospital to undertake initial selection of patients who appeared to also be GenV participants; for this, we tested three scenarios regarding how the hospital could best identify the possible probands. Once the hospital returned the initial potentially matched datasets, the second step was for GenV to then undertake further standardization/validation followed by optimization (highest accuracy, lowest effort of matching) comparing three different matching approaches. GenV data

scientist and hospital analyst both had the authorization of data access, separation principle has implemented for PII and clinical data to protect patients privacy.

2.1. Sampling Frame and Recruitment

GenV cohort participants: Participants for this study were recruited to GenV [22] from a single birthing hospital from commencement of recruitment on 5 December 2020 to 31 December 2021. The study includes all mother-baby pairs in this period, plus consented babies whose birth mothers were not themselves participating in GenV (see Table 1 for inclusion criteria).

Table 1. Criteria for inclusion or exclusion in GenV and the linkage matching cohort.

GenV Birthing Hospital	
Inclusion criteria	All children born in Victoria during the recruitment period whose parents/guardians have decisional capacity, and their parents Participants who leave Victoria may continue to take part via linked and contributed data Families who move to Victoria later and have children born within the recruitment period may join GenV GenV recruitment and data collection materials are offered in multiple languages to enhance accessibility
Exclusion criteria	Infants who die before recruitment to GenV (stillbirth or neonatal death) Families unable to consent in any available language
Linkage Matching Cohort	
Inclusion criteria	Baby is born between December 2020 and December 2021 Consented babies and parents who agreed to participate in GenV There is a record for admission between November 2020 and January 2022 at the selected Victorian birthing hospital.
Exclusion criteria	No additional exclusion criteria

GenV's sampling frame is the daily census of all births at each birthing hospital, plus notification of any in-transferred newborns not recruited at the birthing hospital. Hospital-based GenV-employed recruiters attempt to approach the parent/s of every baby face to face before discharge. They seek parent consent to follow participants indefinitely until study end or withdrawal, with a primary parent/guardian asked to provide consent for themselves and their child (index participant) and any additional parents/guardians asked to consent for themselves only. The broad consent includes to GenV accessing the information that services already collect for them and their child, from before the baby was born and in the future. The recruiter collects parent and child demographic details from the parent and records them in GenV's Participant Relationship Management System (PRMS). At the time of recruitment for this pilot, this did not include recording the unique hospital identifier the Unit Record (UR) Number, a permanent identifier that is assigned to the patient and comprises a digitized number and/or letter combination unique to individuals in an Australian health service [23].

The birthing hospital: The participating birthing hospital has around 5800 births annually and covers much of the culturally diverse western suburbs of Melbourne, Australia. The sampling frame for the hospital dataset spanned an additional month each side of the GenV sample (i.e., November 2020 to January 2022) to ensure data completeness.

2.2. Data Sources and Handling

This study required that we match participants on demographic details (personally identifiable information, PII) in the single GenV PRMS dataset with those that hospital held across two data systems. It further required that the hospital identified which individual babies belonged to which individual mothers. We explored matching using variables available in GenV (Table 2): mother and baby first name (FN), middle names, last name

(LN), date of birth (DOB), birth order (BO—baby only, for multiple births), baby sex (SEX), mother/the other parent's phone number (TN) and home address (ADD), and mother's and baby's computer-generated ID.

Table 2. PII data variables explored for matching from GenV mothers and babies.

Mother or Baby	PII Data Variables
Both	ID, generated by computer
Baby	First Name (B-FN) Middle Name Last Name (B-LN) Birthdate (B-DOB) Birth Order (BO) for multiple births (e.g., twins, triplets) Gender (Sex, female, male and unknown)
Biological Mother	First Name (M-FN) Middle Name Last Name (M-LN) Birthdate (M-DOB) Mother's Phone number (TN) Mother's street Address (ADD)
The other parent	Other parent's Phone number (TN) Other parent's Street Address (ADD)

Note: Middle names were not used in the initial and further matching analysis.

Figure 1 shows the process flows. The GenV data scientist securely extracted the participant variables from GenV's PRMS, encrypted the data in a single CSV file saved in a secured Owncloud account, and provided the login information to hospital analyst. The hospital data analyst undertook initial matching resulting in a separate hospital dataset (Supplemental Table S3) and then returned both the original GenV personally identifiable information (PII) and the new hospital datasets to GenV Owncloud account. The GenV data scientist then undertook all subsequent steps, in communication with the GenV authorized recruitment team and the hospital analyst, until the optimized matching rate was reached.

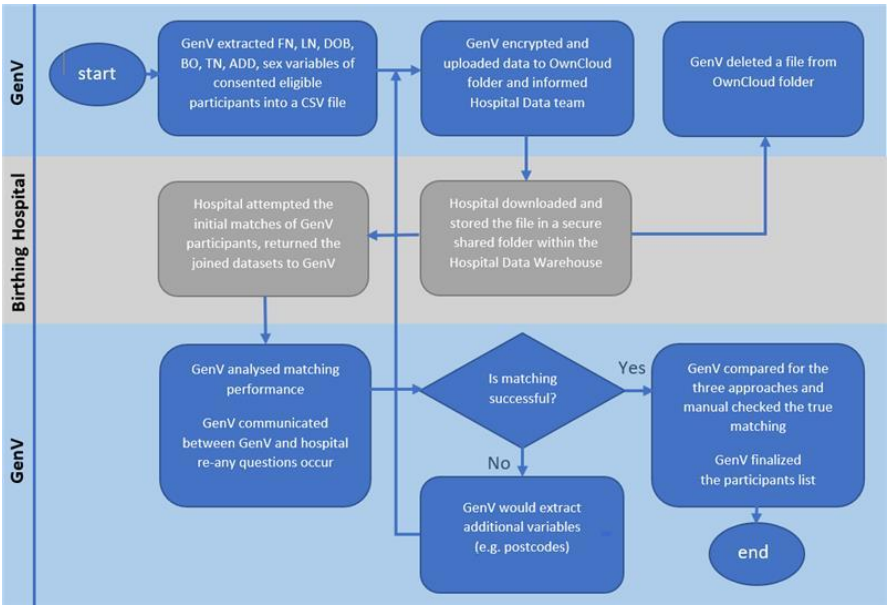

Figure 1. Linkage matching process flow. Note: All matching and data retrieval were done by authorized data scientist from GenV, data initial matching was done by authorized hospital data analyst. FN = first name; LN = last name; DOB = date of birth; BO = birth order; TN = telephone; ADD = addresses.

2.3. Step 1 (Hospital): Initial Matching

Data linkage used two systems: Patient Administration System (iPM) [24] and the Birthing Outcomes System (BOS) [25]. iPM accumulates patients' details (e.g., name, address, DOB, telephone) and their admission details, such as admission and separation dates and admission type. BOS is a standalone product, which stores birthing and pregnancy data (including birth order, baby sex) independently from iPM. The two systems are internally linked within the hospital by the UR number.

The hospital data analyst linked GenV participants with minimum sets of variables utilizing three scenarios (see Table 3): (1) primarily using the mother's details to link the mother and baby, (2) primarily using the baby's details to link the mother and baby, and (3) using all available variables for babies with no information about their mothers but the other parents' information.

Table 3. Summary of the three hospital linkage scenarios.

Scenario	Details	Outcome
Scenario 1: Primarily using mother's details to link the mother and baby	1st letter of mother first name and first 2 letters of mother surname and mother's date of birth (dd/mm/ or yyyy). Any inpatient management (iPM, inpatient admission data) data would be joined where the Unit Record (UR) number from above criterion joins to IP admission UR number, and the Child's birthdate is between the IP admission date -1 day and separation date, and it is a maternity episode. Additional Birth Outcomes System (BOS)/iPM data will join where the UR number from the 1st criterion joined BOS UR number.	The hospital identified 1919 potential pairs
Scenario 2: Primarily using baby's details to link the mother and baby	1st two letters of baby surname OR Mother LN + baby (DOB). Join any inpatient episode through baby's UR number In addition, BOS/iPM data were joined by their common UR number.	The hospital identified 2289 potential pairs
Scenario 3: Using all available variables for babies without information about their mothers	Using babies' FN and LN, DOB Using baby's PII and obtained baby's other parent's TN and ADD.	The hospital identified 53 out of 58 babies

Note: iPM = Patient Administration System; BOS = Outcomes System; FN = first name; LN = last name; DOB = date of birth; BO = birth order; TN = telephone; ADD = addresses.

2.4. Step 2 (GenV): Standardization/Validation and Optimization of Matching Approaches

To improve matching rate, we implemented standardization and validation for the PII data variables of names, telephone number and address during further analysis of both the original GenV dataset and the three datasets returned from the birthing hospital after potential matches were identified.

LN and FN: As both datasets may have spelling errors, where they differed, we could not know which were the true correct names. We standardized names by removing spaces, hyphens and special symbols (') between two words; all letters were upper cased for both FN and LN. Although neither dataset had missing values for names of mothers or babies, this often included the non-specific 'Baby' in the FN field.

TN: We standardized telephone numbers in adherence to Australian Telecommunications standards (Figure 1) [26]. The hospital's dataset has two TN variables ('mobile number' and 'other phone') while the GenV cohort has one TN variable (with 1.1% of values missing). We combined the two hospital TN variables into one TN variable and reduced the missing values to 3.5%. This required converting raw telephone numbers into the standard format and the removal of special characters such as apostrophes or hyphens. In addition, country code was removed and the leading '0' was removed.

ADD: We applied Australian Postal Service certified address standardization rules for thoroughfare abbreviations (Figure 2) [27]. Hyphens, empty spaces, and special symbols were removed. For example, address variables in both datasets had the following format after pre-processing: '214SMITHST' (Supplemental Table S1). In the GenV dataset 1.4% of parents' address values were missing, but there were no missing values in the hospital dataset.

Figure 2 illustrates the example of data standardization and validation between the GenV and hospital datasets on the variables of phone number and address.

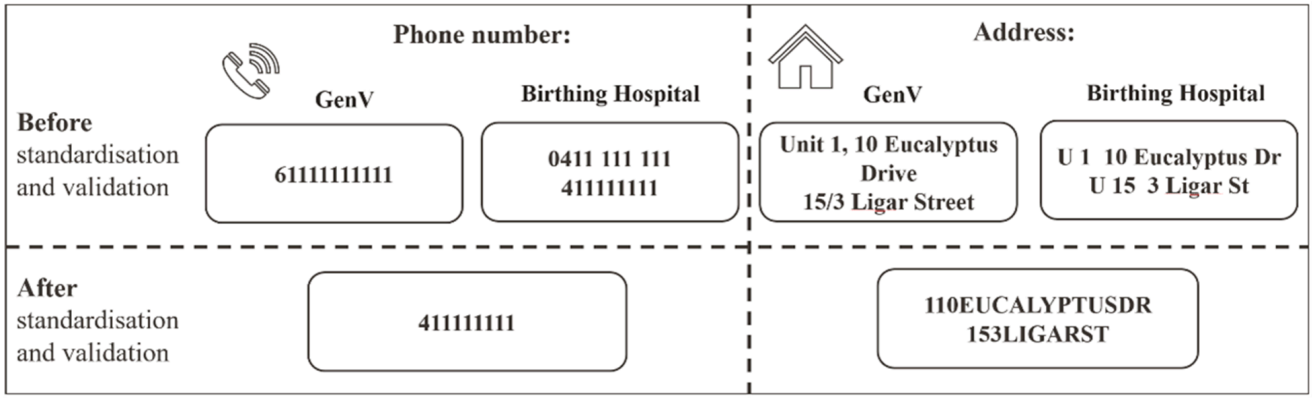

Figure 2. Data standardisation and validation examples for telephone and address. Note: Phone number and address are mock-up examples.

We tested three different approaches after data were standardized and validated: a deterministic rule-based approach and two modified SLK-581 approaches.

- Deterministic rule-based approach: Use all mothers' and all babies' PII (Detail in Table 2).
- Modified SLK-581 approach 1—SLK-5881: SLK-581 [28] is a 14-character code comprising the 2nd, 3rd and 5th characters of the family name, the 2nd and 3rd of the given/first name ('5'), the date of birth (DDMMYYYY, '8') and sex ('1'). This has previously provided successful linkage in some datasets but less so in studies using data with a high rate of missing names [18]. As the GenV cohort includes twin and triplet babies with imprecise names (e.g., 'Boy', 'Girl', 'Twin1', 'Twin2') and the sex of the babies may also be imprecise, we modified the SLK-581 by adding babies' B-DOB and BO as 'birth order' was the only unique identifier for twins. We removed the mother's sex variable as mother's sex was the same in both datasets. We named this new linkage method as SLK-5881 as the '5' (2nd, 3rd letter of FN, and 2nd, 3rd, 5th letter of LN), the first '8' remains the mother's DOB, with the additional '8' for baby's DOB and '1' for birth order.
- Modified SLK-581 approach 2—SLK5881.1: Additionally, we tested a modified linkage method using the first 2 letters of first name and the first 3 letters of last name. M-DOB, B-DOB, B-DOB and BO were used as per SLK-5881.1.

2.5. Statistical Analysis and Evaluation Metrics

We defined three categories for matching results: fully matched, non-matched and partially matched. True matches were defined when all selected data variables were identical between GenV and the hospital datasets. Non-matches referred to when a GenV participant recruited from that hospital was not found in the hospital dataset. Partial matches referred to when one or more selected variables were matched while the rest of the variables had discrepancies, which required a further manual check to determine whether matches belong to true matches or non-matches. The matching performance quality was evaluated using Accuracy rate = (true matches)/(true matches + non-matches + partial matches) $\times$ 100%. We used manual comparison as the gold standard to confirm all true matches for this pilot.

2.6. Ethics, Privacy, and Data Protection

Ethical approval is in place for the GenV cohort (Royal Children's Hospital Human Research Ethics Committee (HREC)-2019/11), including consent to access clinical data. Written informed consent was obtained from participants. For this participating hospital data linkage matching, we further obtained site-specific governance authorization, including site-specific assessment, material transfer agreement and privacy assessment before PII data extraction commencement.

3. Results

3.1. The Sample

The GenV cohort dataset included 1819 mother-baby pairs and 58 babies without mothers' PII recruited from the participating hospital between December 2020 and December 2021. The hospital identified a total of 22,236 adult patients (by using 1st letter of M-FN and 2nd letter of M-LN and year of M-DOB), 5565 babies with mothers' PII (1st two letters of B-LN or M-LN and year of B-DOB) and 76 babies without mothers' PII (1st 2 letters of baby LN and year of B-DOB) from November 2020 to January 2022 with which to undertake their initial matching.

3.2. Step 1: Hospital Initial Matching in the Three Linkage Scenarios Analysis

The hospital data analysist assessed the initial matching and identified the most promising scenarios for further analysis. This process determined whether we use the mothers' or babies' PII for matching (Figure 3) and without manual check.

In Scenario 1, using adult patients' PII details for matching, 1919 possible mother-baby pairs were identified in the hospital dataset. After removing 69 duplicates (twins), there were 160 fully matched pairs, 1622 partially matched and 37 non-matched pairs.

In Scenario 2, using babies' PII details for matching, 2289 possible mother-baby pairs were identified in the hospital dataset. There were 94 fully matched pairs, 1555 partially matched and 170 non-matched pairs.

In Scenario 3, for the 58 babies without mothers' details, 53 babies from the hospital dataset were initially identified, with 28 babies partially or fully matched. A combination of other parents' and babies' PII details was used for further matching process.

The hospital then returned the potential matches from all three scenarios to GenV, comprising the selected PII variables of the overlapping 1919 mother-baby pairs and 2289 mother-baby pairs identified above (with 1650 pairs common to both) and 53 babies without mother's information but with the other parents' ADD and TN information.

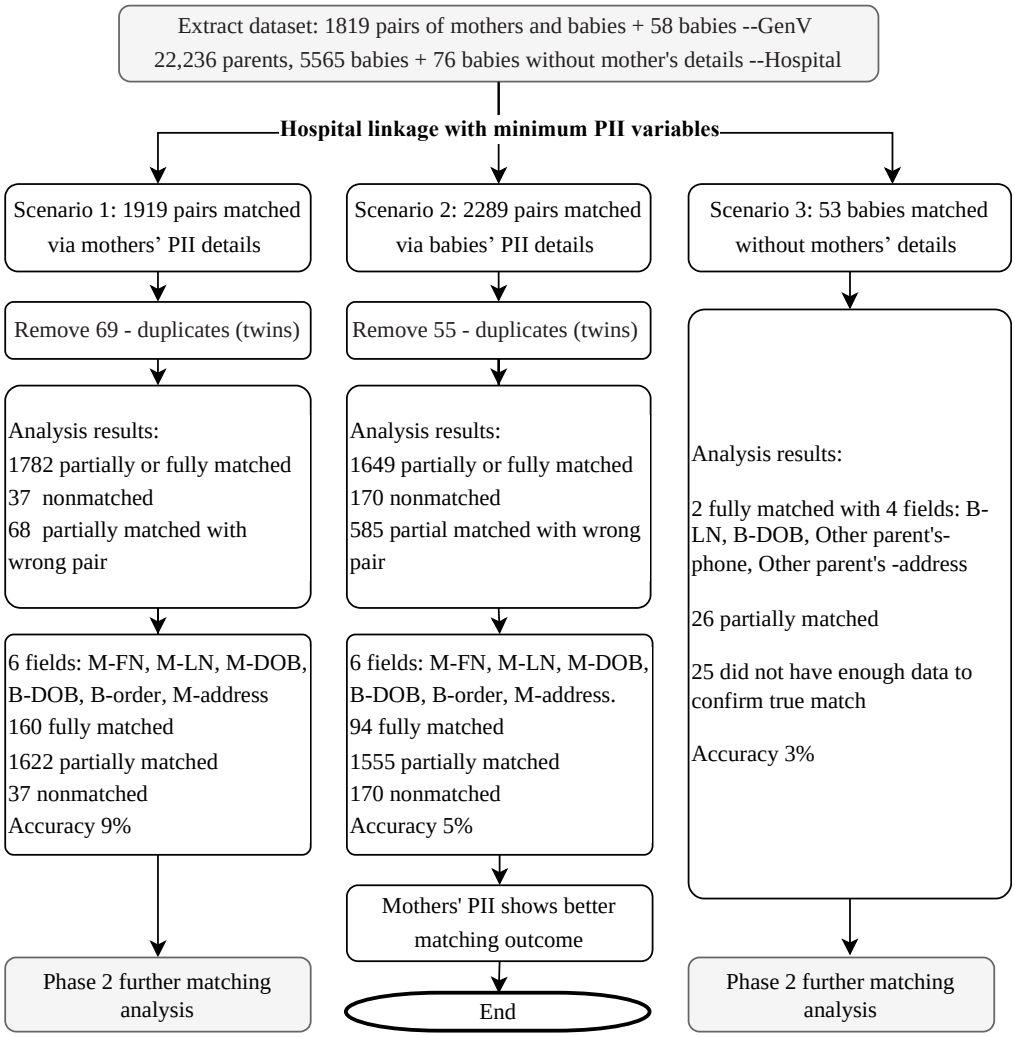

Figure 3. Three linkage scenarios of hospital initial matching. Note: PII = personally identifiable information.

3.3. Step 2 (GenV): Standardization/Validation and Optimization of Matching Approaches

Without data pre-processing and standardization, the accuracy rate for scenario 1 and 2 with 6 variables (M-FN, M-LN, M-DOB, B-DOB, B-BO, M-address) was 9% (160/1819) and 5% (94/1819), respectively. In Scenario 3, with 4 variables B-LN, B-DOB, Other parent's address (ADD) and phone (TN) used for initial matching, the accuracy rate was 3% (2/53).

Supplemental Table S6 describes the three different approaches and their sequential steps with different demographic variables, all true matches were confirmed manually. Figure 4 shows the three different approaches and their accuracy rates at each step (noting that subsequent recommendations related not only to accuracy but also to effort required).

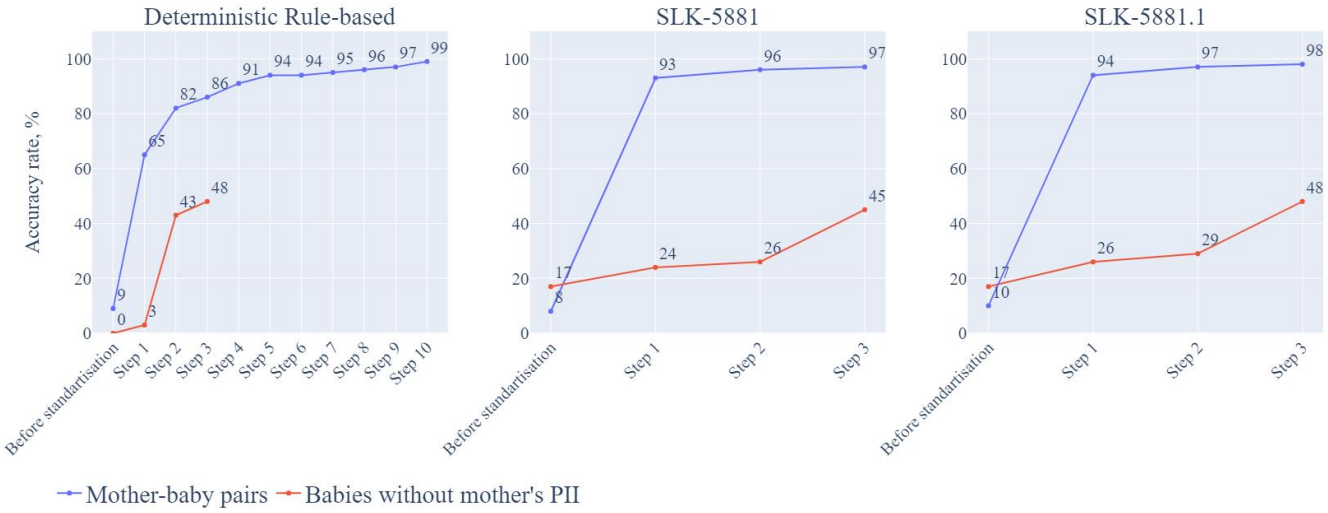

Figure 4. Accuracy rates for mother-baby pairs and babies without mothers' PII information and their required steps for the three approaches. Note: details of included variables for each step listed in Supplemental Table S6. PII = personally identifiable information; SLK = Statistical Linkage key.

3.3.1. Deterministic Rule-Based Approach

Mother-Baby Pairs

The matching accuracy rate was 9% (168/1819) before standardization. After data standardization and validation, we proceeded to sequential matching steps. Step 1 employed 7 variables from mothers and babies (M-FN, M-LN, M-DOB, M-ADD, M-TN, B-DOB, B-BO) as linkage keys for matching, and the accuracy rate was 65% (1189/1819). In Step 2, we used 6 variables, excluding address (M-FN, M-LN, M-DOB, M-TN, B-DOB, B-BO) and the accuracy rate increased to 82% (1486/1819). In Step 3, we used 6 variables without FN, and the accuracy rate increased to 86% (1571/1819). In Step 4, we used 6 variables without mother's TN and the accuracy rate reached 90% (1648/1819). In Step 5, we used 6 variables without M-DOB, and the accuracy rate increased to 93% (1694/1819). By step 10, the accuracy rate reached 99% (1800/1819). Supplemental Table S6 shows the details of all 10 steps and the variables used for each step. Supplemental Figure S1 provide examples of and reasons for partial matching requiring further manual checks.

After manually checking through both the hospital and cohort datasets, we then reached 100% matching accuracy rate for mother-baby pairs.

Babies without Mother's PII

In the GenV cohort dataset, 58 babies recruited from this sampling frame did not have mothers' PII. Before standardization and validation, the matching rate was 0% as there were no matches for any of the 58 babies in the hospital dataset. After standardization and validation, it took 3 steps to match those babies. In Step 1, with baby variables of name, DOB, SEX, BO and the other parent's TN and ADD, the accuracy rate was 3% (2/58). In Step 2, baby variables included LN, DOB, SEX, BO and other parent's ADD; the accuracy rate was 43% (25/58). In Step 3, only baby's FN, LN, DOB, SEX, BO were included; the accuracy rate was 48% (28/58) (see Table 3).

3.3.2. SLK-5881 Matching Approach

Mother-Baby Pairs

Similar to the deterministic rule-based approach, we applied the SLK-5881 approach with matching steps sequentially. The matching accuracy rate was 5% (99/1819) before standardization. In step 1, accuracy rate reached 93% (1687/1819) by using 5 variables including 2nd and 3rd letters of M-FN, 2nd and 3rd and 5th letters of M-LN, M-DOB, B-DOB, B-BO. In step 2, accuracy rate achieved 96% (1755/1819) by using 5 letters of

mother's name, M-DOB, B-DOB, B-BO and M-TN. In step 3, accuracy rate reached 97% (1764/1819) (Supplemental Table S6).

Babies without Mother's PII

All steps for SLK-5881 approach were sequential with 53 babies identified in hospital dataset without manual check. Before standardization and validation, the accuracy rate was 17% (10/58). After standardization and validation, we used 2nd, 3rd and 5th letters of B-LN for all steps in this approach. In step 1, we included additional 2nd and 3rd letters of B-FN, B-DOB, B-Sex and BO, the accuracy rate was 24% (13/53). In step 2, additional B-DOB, B-Sex and BO and the other parent's M-TN were included, the accuracy rate was 26% (14/53). In step 3, additional B-DOB, B-Sex and BO and other parent's address were used for matching, accuracy rate was 45% (26/53).

3.3.3. SLK-5881.1 Matching Approach

Mother-Baby Pairs

Similar to SLK-5881 approach, we employed SLK-5881.1 approach for matching linkage. The matching accuracy rate was 9% (168/1819) before standardization. After standardization and validation, we achieved 94% accuracy rate (1709/1819) in step 1, 97% (1764/1819) in step 2, and 98% (1782/1819) in step 3. See supplemental Table S6 for the data variables used.

Babies without Mother's PII

We adjusted SLK-5881.1 approach by using baby's names and the other parent's TN and ten symbols (letters and numbers) of the other parent's ADD variable to match those babies without mothers' PII. Like mother-baby pair matching, all steps were sequential. The accuracy rate for SLK-5881.1 was 17% (10/58) before standardization and validation. After standardization and validation, the accuracy was 26% (15/58), 29% (17/58) and 48% (28/58), respectively in step 1, step 2 and step 3 (Supplemental Table S6).

Supplemental Figure S2 shows reasons for partially matched pairs requiring further manual check for both SLK-5881 and 5881.1.

4. Discussion

4.1. Principal Results

High matching to birthing hospital records of mother-baby pairs is possible for a consented cohort without unique identifiers. Before standardization and validation, the accuracy rate was less than 10%. After standardization and validation, all 3 approaches showed very high success rates for the mother-baby pairs (100% vs. 97% vs. 98%). Modified SLK-581 approaches were faster and involved much less effort (3 steps) than the deterministic rule-based approach (10 steps) and should enable efficient scaling when linking the whole GenV cohort to multiple birthing hospitals across our state. For babies without mothers' PII, none of our three standardized approaches achieved a matching accuracy rate above around 50%; direct manual perusal of hospital and cohort records did eventually achieve 100% matching, but this may not be scalable for a large number.

While some demographic variables were already largely standardized (e.g., DOB) across both data sources, standardization and validation of other demographic variables greatly improved matching performance between the GenV and hospital datasets. Babies' first names could not be standardized as for many this was simply listed as 'Baby', especially in the hospital dataset.

4.2. Comparisons with Published Studies

For our consented cohort using only demographic summary variables, we achieved high matching rates to birthing hospitals records using efficient, secure methods. These matching rates were comparable to studies using combinations of unique identifiers (lacking in our dataset) and the child's and mother's demographic information, e.g., the studies

from UK [14] and Hong Kong [10]. We have found no other reported studies that have achieved this.

Several studies have, however, shown that standardized variables can improve matching over and above that with unique identifier numbers alone. A study from the US (matching health information exchange (HIE) records, public health registry, Social Security Death Master File records and newborn screening records) showed standardizing individual variables (telephone and date of birth, as well as social security number) increased matching sensitivity from 81.3% to 91.6%; however, standardizing address and last name showed no improvement [29]. A Canadian study achieved 95% matching of records across community health service agencies via a weighted approach; matches based on Health Card Number and Last name were weighted 1.0, while matches based on the Last name, First name, DOB, and Gender were weighted 0.7 [30].

4.3. Implications

Taking Australian experience as a case study, multiple barriers face researchers attempting to bring together birth cohort and hospitals data. Because hospitals use diverse records of varying sophistication, ranging from fully electronic to handwritten medical records within and across hospitals, clinical data are not brought together in any unifying or collated way beyond the minimum Victorian Admitted Episodes Dataset (VAED), and each hospital's unique identifier number applies only to that hospital. In this situation, cohorts must work across inconsistent health record systems on a hospital-by-hospital basis, without the benefit of a unique identifier—but, until now, there have been no reports as to how to achieve this to the high level of accuracy cohorts may require. Our study shows that it is possible to navigate the necessary privacy and legal requirements and work in a cohort-hospital partnership to this end. Further, by modifying an existing Australian linkage key (SLK-581) we provide a short set of variables (phone, addresses with additional mothers and babies' date of birth, baby's birth order) and steps that can be efficiently applied to achieve high rates of matching for mother-baby pairs in birthing hospitals. On the other hand, for many Australian babies are born under mothers' name, but not all, Australian babies after birth (e.g., babies born to sole mothers, gay parents, or whose parents choose to give them the mother's surname). Our approach has proved sufficiently flexibility to match regardless of whether babies' surname is the same or different from their mothers' surnames. These are likely to be generalizable to other hospitals, regardless of which health record system they are using and what the formats of those demographic data are.

Performance of the SLK-5881 and SLK-5881.1 approaches was comparable. The former might be seen as offering superior privacy protection by virtue of its use of variables in selected positions, rather than consecutive letters, which perhaps could be more identifiable in the event of data breaches.

While undertaken securely, our processes were not fully de-identified, in keeping with the consent of the cohort participants. Based on our experience, highly accurate fully de-identified linkage is unlikely without unique numbers such as unit record (UR) number, Medicare number [31] or some other individual healthcare identifier [32]. However, there is increasing concern about sharing such numbers both by individuals and from a legal perspective [33]. With many techniques proposed for privacy-preserving clinical linkage [34], we hope that future safe access to such datasets for beneficial research will be widely supported and enabled [13,35] in ways that meet the needs both of citizens and good governance.

4.4. Limitations

There are several limitations to this study. First, our results are specific to the hospital we selected. However, there is no reason to think that the demographic variables we used would differ greatly from those in other Australian hospitals [36]. Therefore, our findings are likely to apply for birthing hospital-to-cohort exchanges throughout Victoria and potentially nationally and to other countries with computerized administrative

databases; however, we acknowledge that these may not always be available in low-middle income countries, which may raise different security concerns than managed here. Second, this approach may not apply to matching of babies without mothers' information as the matching accuracy rate was low with all three approaches. Third, this standardization might not cover all the potential partial matches which might occur in future data sources with, for example, name changes with altered family compositions over time. However, this first experience is very promising for GenV's ability to access hospitals data at a critical universal lifecourse juncture (the pregnancy, birth and newborn periods) and will help optimize our process as we scale up to include more hospitals.

5. Conclusions

Standardizing participants' names, phone and address demographic data led to very high cohort-to-hospital matching rates both for the mothers and the babies in our large, consented birth cohort. This was achieved despite the cohort not holding hospital unique identifier numbers. We recommend our modified SLK-5881 approach as achieving the best balance of accuracy, efficiency, safety and scalability to the population level.

Because health care and health status during pregnancy, birth and the newborn periods can influence the whole lifecourse, linking to hospital health records covering these periods could provide immense additional value to cohorts in their quest to improve maternal and child health.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/children9121916/s1. Table S1: Mapping dictionary for standardization of address suffixes. Table S2: Examples of address mismatching reasons requiring manual check. Table S3: Hospital returned the variable list for GenV participants for further matching analysis. Table S4: Reporting guidelines for studies using data linkage. Table S5: The RECORD statement—checklist of items, extended from the STROBE statement, that should be reported in observational studies using routinely collected health data. Table S6: Summary of the three matching approaches and their interactions steps with different variables. Figure S1: The reasons for partial matching for mother-baby pairs through deterministic rule-based approach. Figure S2: The reasons for partial matching for mother-baby pairs through SLK-5881 and SLK-5881.1 approaches.

Author Contributions: Y.J.H. conceived this concept and made the first draft. J.W. prepared the required datasets and undertook the phase I data analysis. A.F. undertook data extraction and analysis. J.M.S., N.T., E.N., J.L.Y.C., B.K., S.G., M.W. provided the critical consultation and comments for this study. M.W. supervises Y.J.H. and is the Scientific Director of GenV. All authors have read and agreed to the published version of the manuscript.

Funding: This pilot work was conducted by Generation Victoria (GenV), which is supported by grants from the Paul Ramsay Foundation, the Victorian Government and the Royal Children's Hospital Foundation. Research at the Murdoch Children's Research Institute is supported by the Victorian Government's Operational Infrastructure Support Program. J.W. was supported by a Melbourne Children's LifeCourse postdoctoral fellowship, funded by Royal Children's Hospital Foundation grant (reference number 2018–984). S.G. was supported by the Australian National Health and Medical Research Council (NHMRC) Practitioner Fellowship (reference number 155290). M.W. was supported by NHMRC Principal Research Fellowship (reference number 1160906). J.C. was supported by Career Development Fellowship from the MRFF 1141354.

Institutional Review Board Statement: Ethical approval was received from Royal children's Hospital Human Research Ethics Committee (HREC), reference number: HREC/51302/RCHM-2019. Privacy assessment was done by Corrs Chambers Westgarth, a leading independent Australian law firm. Written informed consent was obtained from participants.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: We are not able to share the participant data for this article due to privacy and legislation requirements. For matching algorithms, please contact Y.J.H.

Acknowledgments: We would like to thank the Performance Unit from Joan Kirner Hospital for the initial data matching; the GenV Legal, Cohort and Platform teams; and the Joan Kirner GenV participants.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Cowie, M.R.; Blomster, J.I.; Curtis, L.H.; Duclaux, S.; Ford, I.; Fritz, F.; Goldman, S.; Janmohamed, S.; Kreuzer, J.; Leenay, M.; et al. Electronic health records to facilitate clinical research. *Clin. Res. Cardiol.* **2017**, *106*, 1–9. [CrossRef]
2. Farmer, R.; Mathur, R.; Bhaskaran, K.; Eastwood, S.V.; Chaturvedi, N.; Smeeth, L. Promises and pitfalls of electronic health record analysis. *Diabetologia* **2018**, *61*, 1241–1248. [CrossRef]
3. Colombo, F.; Oderkirk, J.; Slawomirski, L. Health information systems, electronic medical records, and big data in global healthcare: Progress and challenges in oecd countries. In *Handbook of Global Health*; Springer: Berlin/Heidelberg, Germany, 2020; Chapter 71-1, pp. 1–31.
4. Harron, K.; Dibben, C.; Boyd, J.; Hjern, A.; Azimaee, M.; Barreto, M.L.; Goldstein, H. Challenges in administrative data linkage for research. *Big Data Soc.* **2017**, *4*, 2053951717745678. [CrossRef] [PubMed]
5. Smith, M.; Flack, F. Data linkage in australia: The first 50 years. *Int. J. Environ. Res. Public Health* **2021**, *18*, 11339. [CrossRef]
6. Tew, M.; Dalziel, K.M.; Petrie, D.J.; Clarke, P.M. Growth of linked hospital data use in australia: A systematic review. *Aust. Health Rev.* **2017**, *41*, 394–400. [CrossRef]
7. Wood, A.; Denholm, R.; Hollings, S.; Cooper, J.; Ip, S.; Walker, V.; Denaxas, S.; Akbari, A.; Banerjee, A.; Whiteley, W.; et al. Linked electronic health records for research on a nationwide cohort of more than 54 million people in england: Data resource. *BMJ* **2021**, *373*, n826. [CrossRef]
8. Casey, J.A.; Schwartz, B.S.; Stewart, W.F.; Adler, N.E. Using electronic health records for population health research: A review of methods and applications. *Annu. Rev. Public Health* **2016**, *37*, 61–81. [CrossRef]
9. Darke, P.; Cassidy, S.; Catt, M.; Taylor, R.; Missier, P.; Bacardit, J. Curating a longitudinal research resource using linked primary care ehr data-a uk biobank case study. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 546–552. [CrossRef]
10. Gao, L.; Leung, M.T.Y.; Li, X.; Chui, C.S.L.; Wong, R.S.M.; Yeung, S.L.A.; Chan, E.W.W.; Chan, A.Y.L.; Chan, E.W.; Wong, W.H.S.; et al. Linking cohort-based data with electronic health records: A proof-of-concept methodological study in Hong Kong. *BMJ Open* **2021**, *11*, e045868. [CrossRef]
11. Mykletun, A.; Widding-Havneraas, T.; Chaulagain, A.; Lyhmann, I.; Bjelland, I.; Halmøy, A.; Elwert, F.; Butterworth, P.; Markussen, S.; Zachrisson, H.D.; et al. Causal modelling of variation in clinical practice and long-term outcomes of adhd using norwegian registry data: The adhd controversy project. *BMJ Open* **2021**, *11*, e041698. [CrossRef]
12. Reed, B.D.; Schibler, K.R.; Deshmukh, H.; Ambalavanan, N.; Morrow, A.L. The impact of maternal antibiotics on neonatal disease. *J. Pediatr.* **2018**, *197*, 97–103.e3. [CrossRef] [PubMed]
13. Young, A.; Flack, F. Recent trends in the use of linked data in australia. *Aust. Health Rev.* **2018**, *42*, 584–590. [CrossRef] [PubMed]
14. Tingay, K.S.; Bandyopadhyay, A.; Griffiths, L.; Akbari, A.; Brophy, S.; Bedford, H.; Cortina-Borja, M.; Setakis, E.; Walton, S.; Fitzsimons, E.; et al. Record linkage to enhance consented cohort and routinely collected health data from a uk birth cohort. *Int. J. Popul. Data Sci.* **2019**, *4*, 579. [CrossRef]
15. Cox, S.; Martin, R.; Somaia, P.; Smith, K. The development of a data-matching algorithm to define the 'case patient'. *Aust. Health Rev.* **2012**, *37*, 54–59. [CrossRef] [PubMed]
16. Nguyen, L.; Stoové, M.; Boyle, D.; Callander, D.; McManus, H.; Asselin, J.; Guy, R.; Donovan, B.; Hellard, M.; El-Hayek, C. Privacy-preserving record linkage of deidentified records within a public health surveillance system: Evaluation study. *J. Med. Internet Res.* **2020**, *22*, e16757. [CrossRef] [PubMed]
17. Coulson, T.G.; Bailey, M.; Reid, C.; Shardey, G.; Williams-Spence, J.; Huckson, S.; Chavan, S.; Pilcher, D. Linkage of australian national registry data using a statistical linkage key. *BMC Med. Inform. Decis. Mak.* **2021**, *21*, 37. [CrossRef]
18. Taylor, L.K.; Irvine, K.; Iannotti, R.; Harchak, T.; Lim, K. Optimal strategy for linkage of datasets containing a statistical linkage key and datasets with full personal identifiers. *BMC Med. Inform. Decis. Mak.* **2014**, *14*, 85. [CrossRef] [PubMed]
19. Australian Bureau of Statistics. Popuation: Census. 2021. Available online: https://www.abs.gov.au/statistics/people/population/2021-census-overcount-and-undercount/2021 (accessed on 1 October 2022).
20. Bohensky, M.A.; Jolley, D.; Sundararajan, V.; Evans, S.; Ibrahim, J.; Brand, C. Development and validation of reporting guidelines for studies involving data linkage. *Aust. N. Z. J. Public Health* **2011**, *35*, 486–489. [CrossRef]
21. Benchimol, E.I.; Smeeth, L.; Guttmann, A.; Harron, K.; Moher, D.; Petersen, I.; Sørensen, H.T.; von Elm, E.; Langan, S.M. The reporting of studies conducted using observational routinely-collected health data (record) statement. *PLoS Med.* **2015**, *12*, e1001885. [CrossRef]
22. Generation Victoria (GenV). What's Genv. 2020. Available online: https://genv.org.au/about-genv/what-is-genv/ (accessed on 11 June 2022).
23. Department of Health Standard. Assignment of Unique Unit Record Number. Standard QH-IMP-280-3:2014. Queensland Department of Health. 2015. Available online: https://www.health.qld.gov.au/__data/assets/pdf_file/0030/397254/qh-imp-280-3.pdf (accessed on 11 June 2022).

24. Sarkies, M.N.; Bowles, K.-A.; Skinner, E.H.; Mitchell, D.; Haas, R.; Ho, M.; Salter, K.; May, K.; Markham, D.; O'Brien, L.; et al. Data collection methods in health services research: Hospital length of stay and discharge destination. *Appl. Clin. Inform.* **2015**, *6*, 96–109. [CrossRef]
25. Knight-Agarwal, C.R.; Williams, L.T.; Davis, D.; Davey, R.; Cochrane, T.; Zhang, H.; Rickwood, P. Association of bmi and interpregnancy bmi change with birth outcomes in an australian obstetric population: A retrospective cohort study. *BMJ Open* **2016**, *6*, e010667. [CrossRef] [PubMed]
26. Horsley, A.; Gerrand, P. Major policy gaps in Australian telecommunications. *Telecommun. J. Aust.* **2011**, *61*, 1–6. [CrossRef]
27. Alam, Q.; Grose, R. Australia post. In *Regional Businesses in a Changing Global Economy: The Australian Experience*; Routledge: London, UK, 2022.
28. Australian Institute of Health and Welfare. Slk-581 Guide for Use. 2016. Available online: https://www.aihw.gov.au/getmedia/e1d4d462-8efa-4efa-8831-fa84d6f5d8d9/aodts-nmds-2016-17-SLK-581-guide.pdf.aspx (accessed on 1 October 2022).
29. Grannis, S.J.; Xu, H.; Vest, J.R.; Kasthurirathne, S.; Bo, N.; Moscovitch, B.; Torkzadeh, R.; Rising, J. Evaluating the effect of data standardization and validation on patient matching accuracy. *J. Am. Med Inform. Assoc.* **2019**, *26*, 447–456. [CrossRef] [PubMed]
30. Eze, B.; Kuziemsky, C.; Peyton, L. A patient identity matching service for cloud-based performance management of community healthcare. *Procedia Comput. Sci.* **2017**, *113*, 287–294. [CrossRef]
31. Duckett, S. Expanding the breadth of medicare: Learning from australia. *Health Econ. Policy Law* **2018**, *13*, 344–368. [CrossRef]
32. Bidargaddi, N.; van Kasteren, Y.; Musiat, P.; Kidd, M. Developing a third-party analytics application using australia's national personal health records system: Case study. *JMIR Med Inform.* **2018**, *6*, e28. [CrossRef]
33. Xafis, V. The acceptability of conducting data linkage research without obtaining consent: Lay people's views and justifications. *BMC Med. Ethics* **2015**, *16*, 79. [CrossRef]
34. Boyd, J.H.; Ferrante, A.M.; O'Keefe, C.M.; Bass, A.J.; Randall, S.M.; Semmens, J.B. Data linkage infrastructure for cross-jurisdictional health-related research in australia. *BMC Health Serv. Res.* **2012**, *12*, 480. [CrossRef]
35. Costa, J.D.O.; Bruno, C.; Schaffer, A.L.; Raichand, S.; A Karanges, E.; Pearson, S.-A. Pearson. The changing face of australian data reforms: Impact on pharmacoepidemiology research. *Int. J. Popul. Data Sci.* **2021**, *6*, 1418. [CrossRef]
36. Dixit, S.K.; Sambasivan, M. A review of the australian healthcare system: A policy perspective. *SAGE Open Med.* **2018**, *6*, 2050312118769211. [CrossRef]