



# Unlocking Potential within Health Systems Using Privacy-Preserving Record Linkage: Exploring Chronic Kidney Disease Outcomes through Linked Data Modelling

David Lim<sup>1</sup> Sean Randall<sup>1</sup> Suzanne Robinson<sup>1,2</sup> Elizabeth Thomas<sup>1,3</sup> James Williamson<sup>4</sup>  
Aron Chakera<sup>3,5</sup> Kathryn Napier<sup>6</sup> Carola Schwan<sup>7</sup> Justin Manuel<sup>7</sup> Kim Betts<sup>1</sup> Chris Kane<sup>8</sup>  
James Boyd<sup>1,9</sup>

<sup>1</sup> Curtin School of Population Health, Curtin University, Perth, Western Australia, Australia

<sup>2</sup> Deakin Health Economics, Deakin University, Burwood, Victoria, Australia

<sup>3</sup> Medical School, The University of Western Australia, Perth, Western Australia, Australia

<sup>4</sup> WA Department of Health, Perth, Western Australia, Australia

<sup>5</sup> Renal Unit, Sir Charles Gairdner Hospital, Perth, Western Australia, Australia

<sup>6</sup> Curtin Institute for Computation, Curtin University, Perth, Western Australia, Australia

<sup>7</sup> WA Country Health Service, Perth, Western Australia, Australia

<sup>8</sup> WA Primary Health Alliance, Perth, Western Australia, Australia

<sup>9</sup> La Trobe University, Melbourne, Bundoora, Victoria, Australia

**Address for correspondence** David Lim, MHA, BPharm, Curtin University, GPO Box U1987, Perth, Western Australia, Australia (e-mail: David.K.Lim@curtin.edu.au).

Appl Clin Inform 2022;13:901–909.

## Abstract

**Background** Chronic kidney disease (CKD) is a major global health problem that affects approximately one in 10 adults. Up to 90% of individuals with CKD go undetected until its progression to advanced stages, invariably leading to death in the absence of treatment. The project aims to fill information gaps around the burden of CKD in the Western Australian (WA) population, including incidence, prevalence, rate of progression, and economic cost to the health system.

**Methods** Given the sensitivity of the information involved, the project employed a privacy preserving record linkage methodology to link data from four major pathology providers in WA to hospital records, to establish a CKD registry with continuous medical record for individuals with biochemical specification for CKD. This method uses encrypted personal identifying information in a probability-based linkage framework (Bloom filters) to help mitigate risk while maximizing linkage quality.

**Results** The project developed interoperable technology to create a transparent CKD data catalogue which is linkable to other datasets. This technology has been designed to support the aspirations of the research program to provide linked de-identified pathology, morbidity, and mortality data that can be used to derive insights to enable better CKD patient outcomes. The cohort includes over 1 million individuals with creatinine results over the period 2002 to 2021.

## Keywords

- ▶ privacy preserving record
- ▶ linkage accuracy
- ▶ chronic kidney disease

received

April 29, 2022

accepted after revision

July 28, 2022

**DOI** <https://doi.org/10.1055/s-0042-1757174>.

**ISSN** 1869-0327.

© 2022. The Author(s).

This is an open access article published by Thieme under the terms of the Creative Commons Attribution License, permitting unrestricted use, distribution, and reproduction so long as the original work is properly cited. (<https://creativecommons.org/licenses/by/4.0/>)

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

**Conclusion** Using linked data from across the care continuum, researchers are able to evaluate the effectiveness of service delivery and provide evidence for policy and program development. The CKD registry will enable an innovative review of the epidemiology of CKD in WA. Linking pathology records can identify cases of CKD that are missed in the early stages due to disaggregation of results, enabling identification of at-risk populations that represent targets for early intervention and management.

## Background and Significance

Chronic kidney disease (CKD) is projected to be the fifth leading cause of death worldwide by the year 2040,<sup>1</sup> with an annual estimate of up to 10 million-associated deaths.<sup>2</sup> CKD is a complex disease associated with a spectrum of pathological, hereditary, and sociodemographic factors.<sup>3</sup> There are multiple risk factors known to contribute to a decline in kidney function, including age ( $\geq 60$  years), existing comorbidities, smoking, low socioeconomic status, cardiovascular disease, a body mass index ( $\geq 30$  kg/m<sup>2</sup>), family history, and the use of certain medications.<sup>4-7</sup>

Access to “joined up” service data from across the health system is required to improve our understanding of CKD.<sup>8-10</sup> CKD research would benefit from linked longitudinal de-identified pathology, morbidity, and mortality data that can be used by nephrology researchers to derive insights to enable better CKD patient plans and outcomes.<sup>8-10</sup> By accessing data from across the health systems, we can collect information about an individual when they present to a broad range of medical (both public and private sector) and community health services.<sup>11</sup> These large and complex data reservoirs provide the foundation to unlock potential within administrative-based “big data” assets.<sup>12,13</sup>

The availability of new types of previously unlinked administrative data, provides opportunities to look at early-stage CKD patient pathways. Big data<sup>12,13</sup> linked assets, based on data collections used to manage, monitor, assess, and review a wide range of health conditions and services, provide unique opportunities to exploit the advanced data science and big data analytics techniques to understand factors influencing CKD disease progression. These linked data assets can provide practical insights and improve clinical understanding and decision making for CKD patients.<sup>12-14</sup>

Worldwide however, a relatively small number of health systems have established and integrated a central population health database for research purposes, including countries such as Singapore,<sup>15</sup> Taiwan,<sup>16</sup> Denmark,<sup>11</sup> Sweden,<sup>17</sup> and Japan.<sup>18</sup> In most other health systems, missing or inconsistent data and privacy concerns with data sharing and linkage across organisations,<sup>11</sup> present major challenges.

While having one of the best health care systems in the world, Australia is no exception to the challenges associated with data integration within and across jurisdictions.<sup>19</sup> However, these challenges are magnified in Australia by a

complex health care landscape which receives its funding through both public (across State, Territory and Australian government arrangements) and private pathways.<sup>19</sup> These distinctive reimbursement pathways often encourage siloed data storage without the option for interoperability and effective data sharing.<sup>19</sup> For CKD research, data challenges emerge when CKD patients undergo pathology tests at different pathology providers (typically private companies). This presents challenges when we attempt to understand the characteristics of CKD in the population through fragmented pathology pathways that are not integrated together.<sup>20</sup>

While some government administrative health collections in Australia are routinely linked and analyzed, important health information such as general practice records and pathology data collected in the private sector<sup>10</sup> has not typically been available for research.<sup>21</sup> The underlying barriers limiting access to these datasets can be attributed to several factors, including privacy regulations, siloed health systems, and a reluctance to release personal information within and across organizational boundaries.<sup>10</sup> Although barriers affecting data sharing are being addressed through open data policies, more work is needed for governments to connect with the private industry to maximize information available on community needs,<sup>22,23</sup> while ensuring privacy concerns are recognized and mitigated.

Privacy preserving record linkage (PPRL) is a data linkage technology that allows organizations to securely link data without releasing personally identifiable information (PII), while maintaining linkage accuracy.<sup>24</sup> The technology has been tested and proven in a several linkage projects to support targeted early interventions, involving non-health datasets such as education, housing, and justice datasets.<sup>22,23</sup> PPRL has been a burgeoning research area, within numerous methodologies proposed in the academic literature.<sup>25</sup> It has also seen adoption in several forms across the world for real-world projects, including the phonetic-hash-based EUPID used in European cancer research.<sup>26</sup> Within Australia, PPRL methods utilizing Bloom filters have received adoption, including the LUMOS project connecting primary and secondary care data.<sup>27</sup> Bloom filter encodings have the key advantage of providing tolerance for spelling mistakes and other typographical errors, important for ensuring high linkage quality and robust analytical findings.<sup>23,24,28</sup> PPRL is an important technique enabling the linkage of otherwise unavailable private pathology datasets assisting to improve our understanding of the CKD and contributing to the clinical management of CKD.<sup>10,28</sup>

## Objectives

Using privacy preserving linkage techniques, the first objective of this research study is to link data from the four largest pathology service providers in Western Australia (WA) to WA hospital, emergency, and mortality data to identify a CKD cohort for analysis. PPRL provides the main enabling methodology to create a linked repository across pathology and hospital data collections. This study aims to demonstrate the use of PPRL as a viable method allowing for collaborative analysis of big datasets across different organizations while maintaining patient privacy. In further studies, linked data will enable the mapping of the burden of CKD in WA, including the incidence, prevalence, rate of progression, and economic cost to the health system.

## Methods

### Data Collections

Retrospective pathology biochemistry data was extracted from participating pathology providers (PathWest, Australian Clinical Laboratories and Clinipath Pathology) for all Western Australians (WAs) aged 18 years or older, who had a serum creatinine test in the study period (2006–2021). Each pathology provider provided biochemistry results that utilized standardized Isotope Dilution Mass Spectroscopy calibration.<sup>29</sup> This enables comparable reporting of serum creatinine. Using the reported serum creatinine result, estimated glomerular filtration rate (eGFR) values were recalculated via the chronic kidney disease epidemiology collaboration equation shown in ▶Fig. 1. The stages of CKD are categorized according to the individual's eGFR, and take into consideration the individual's age, body size, and gender. (Western Diagnostic Pathology, also participating in this study, was not able to provide their data in time for inclusion in this manuscript).

The CKD cohort was identified from the pathology cohort, according to the following criteria, those who were ≥18 years old, with at least two eGFR results of <60 mL/min/1.73 m<sup>2</sup> performed at least 3 months apart. Additional data such as albuminuria was not available at this point for the study and has not been included for the confirmation of kidney damage. This definition will enable subsequent investigations of fluctuating trajectories of renal function.

Morbidity data for the CKD cohort was identified through the Hospital morbidity data system which captures administrative data from all public and private hospitals in WA. Mortality data was sourced from the WA Registry of Births, Deaths, and Marriages. Emergency data was extracted from the Emergency Department Data Collection which captures WA emergency department presentations. Data was extracted from these systems between 2002 and 2021.

$$eGFR = 141 \times \min\left(\frac{SCr}{\kappa}, 1\right)^\alpha \times \max\left(\frac{SCr}{\kappa}, 1\right)^{-1.209} \times 0.993^{Age} \times 1.018[\text{if female}]$$

**Fig. 1** CKD-EPI equation. CKD-EPI, chronic kidney disease epidemiology collaboration.

### Data Model

Data linkage models were developed based on the separation principle to help reduce privacy risks. The separation principle requires data to be split into PII, which are released for linkage, and clinical content data or payload data which are provided to researchers – a unique record number connects the PII record to its clinical content record. The data model and associated data flows have been outlined in ▶Fig. 2. Using these data flows, which separate content data from PII during the linkage process, the risk of revealing sensitive information about individuals is significantly reduced. While our model was designed to utilize record linkage techniques based on encoded information to ensure linkers did not see any PII, a separation data model was still implemented. A common key existed between each set of personal identifiers provided to the linkage unit and its associated clinical content data provided to researchers, to allow an overarching person identifier to be attached to the research data after linkage. The registry data (pathology results, hospital admission information, etc.) accessed by researchers were not encoded in any way, this ensures easy usability by the researchers. As such, for analysts utilizing the registry, there is no difference between a dataset linked using PPRL and one linked using another linkage method.

As highlighted in ▶Fig. 2, PII data was encoded at source (i.e., by the pathology providers and the WA Department of Health [DoH]) before being provided to the Centre for Data Linkage at Curtin University for PPRL. Specific software to enable encoding was provided to all parties. Linkage occurred within a specifically designed ISO27001 certified environment at Curtin University. Linkage between the pathology datasets and WA health records utilized encoded personal identifiers (names, date of birth, sex, and address). The clinical content/payload data for each dataset was securely transferred and stored in SeRP at Curtin University. Examples of structured data stored in the secured environment for the CKD registry include de-identified and anonymized pathology and hospital records, including biochemistry markers such as creatinine and electrolytes, demographic factors such as postcode and WA emergency, morbidity, and mortality data.

### Data Linkage

Data linkage techniques were used to identify records belonging to each individual within and between the various data collections. PPRL involved carrying out data linkage of extracted data using encoded versions of the PII, allowing linkage to be performed without disclosing PII. To ensure consistency during the encoding process, an extraction template and password to encode the data were agreed between the data custodians before securely transferring encoded data to the linkage unit.

While several methods for PPRL have been successfully piloted, PPRL using Bloom filters has shown to offer better all-round linkage quality.<sup>30</sup> Consequently this method was chosen for this project. As outlined in ▶Fig. 3, Bloom filters are binary vectors constructed based on the letter/number combinations contained within the PII using a series of

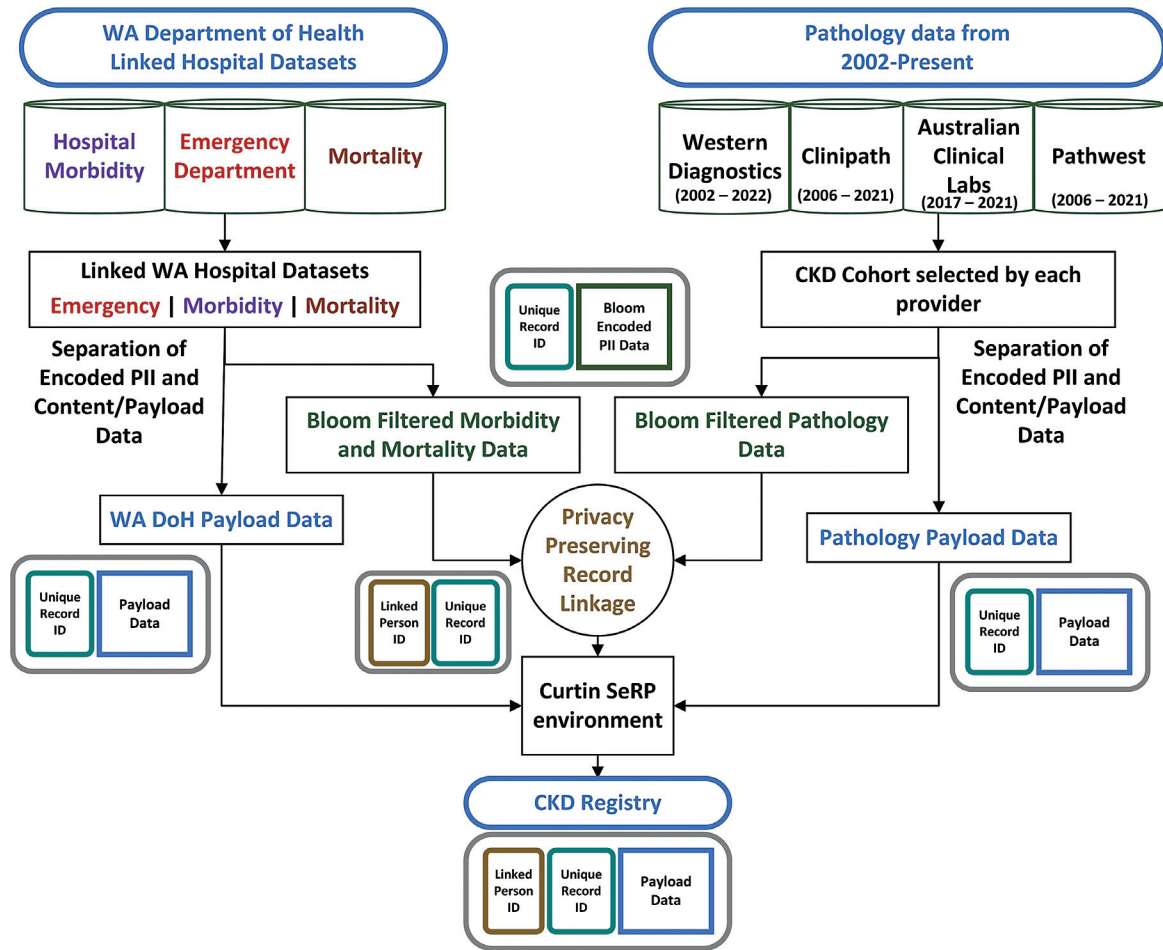


Fig. 2 Data model and data flows.

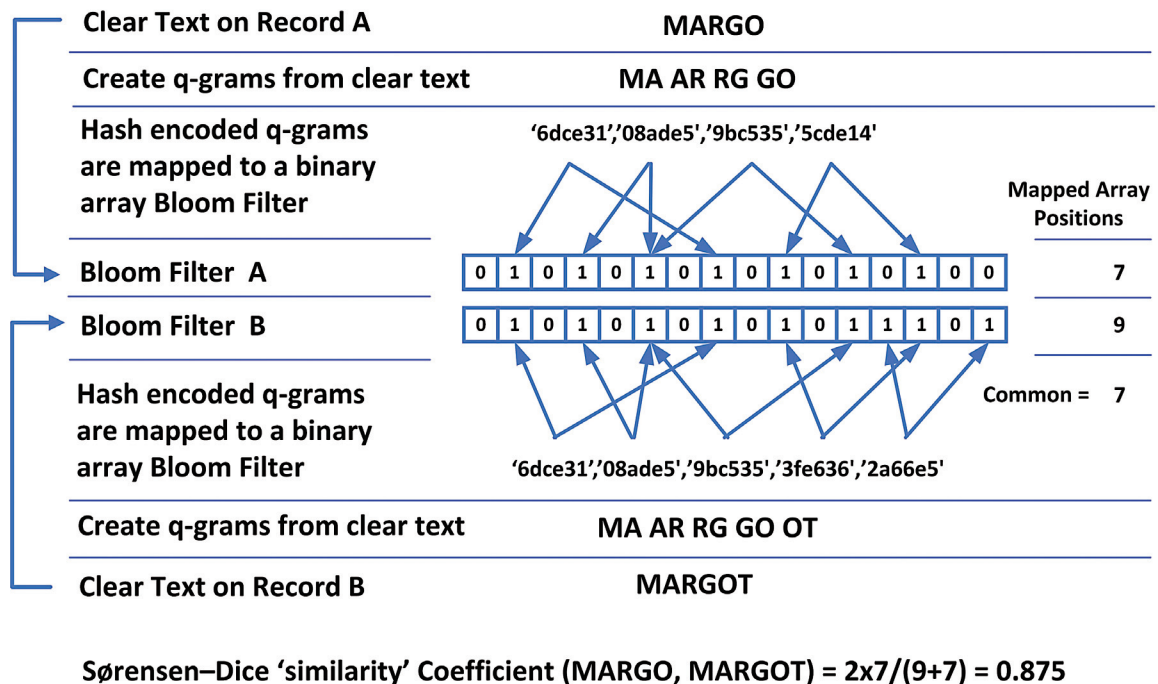


Fig. 3 Bloom filter generation and comparison.



cryptographic hash functions, i.e., letter/number combinations within the PII change positions in the binary array from 0 to 1. As Bloom filters are efficient at determining whether components of the PII appear within the binary array, two bloom filters can be compared with each other based on the number of positions within the Bloom filter that match. The similarity of two Bloom filters (i.e., how many of the array positions match between two Bloom filters) allows us to decide whether records belong to the same individual or not. Further detailed technical information on Bloom filters implementation have been described by Randall et al.<sup>31</sup>

The data was encoded for privacy preserving linkage using field-based Bloom filters; each identifier was encoded into a separate Bloom filter, with a standard probabilistic record linkage method used on these encoded identifiers. This approach has been documented in the literature previously,<sup>31</sup> showing higher linkage quality than other privacy preserving methods.<sup>32</sup>

Date of birth, sex, and postcode were encoded with a single hash value, while names, address, and suburb were encoded into Bloom filters. The Bloom filter encoding used bigrams with no padding, a Bloom filter length of 512, with 30 bit positions set per bigram for names, 20 for address, and 25 for suburb. The number of bit positions set was lower for suburb and address fields as the average length of these fields was longer. Only the first 20 characters of the input fields were used in creating the encoded data. Some basic pre-processing steps occurred as part of the encoding process to standardize the data formats. These included removing whitespace, converting all values to lower case, removing non-alphanumeric characters, converting to a standard format ("Street" to "ST"), and remove placeholder values ("9999" for postcode etc.).

A probabilistic linkage framework was used, utilizing the traditional Fellegi-Sunter approach, whereby individual fields of record-pairs are compared, each resulting in a score based on their agreement or disagreement, with these scores summed and accepted as the same individual if reaching a set threshold. This method had been used previously and shown to achieve high quality linkage results. All available fields (names, sex, date of birth, and address) were used in the comparison process. Fields encoded using Bloom filters were compared using the Sørensen–Dice coefficient. A default set of weights (m and u probabilities) were used; these had been developed and validated through project-based linkages previously undertaken. To reduce the number of pair comparisons, a technique called blocking was used, whereby only record-pairs that had matching values in common were compared further. The blocking strategy employed only compared records further where they had the same date of birth, or the same Soundex value of surname and first initial. These blocks have been used and validated by the linkage team in previous privacy preserving and clear-text linkages.

The datasets from the three pathology providers were deduplicated, linked together, and linked against the WA emergency, morbidity, and mortality datasets. A probabilistic linkage framework was used, utilizing the traditional Fellegi–Sunter approach, whereby individual fields of re-

cord-pairs are compared, each resulting in a score based on their agreement or disagreement, with these scores summed and accepted as the same individual if reaching a set threshold. This method had been used previously and shown to achieve high quality linkage results.<sup>31</sup> All available fields (names, sex, date of birth, and address) were used in the comparison process. Fields encoded using Bloom filters were compared using the Sørensen–Dice coefficient. A default set of weights (m and u probabilities) were used; these had been developed and validated through project-based linkages previously undertaken. To reduce the number of pair comparisons, a technique called blocking was used, whereby only record-pairs that had matching values in common were compared further. The blocking strategy employed only compared records further where they had the same date of birth, or the same Soundex value of Surname and first initial. These blocks have been used and validated by the linkage team in previous privacy preserving and clear-text linkages. Encoding and linkage were performed using the LinXmart software suite<sup>33</sup>; linkage was performed sequentially and took approximately 10 hours to load and link all datasets to each other. Morbidity, emergency, and mortality data had been previously linked by the WA DoH's Data Linkage Branch; these links were supplied and honored within this study. The final extract from linkage included all pathology records, and any emergency, hospital, or mortality record that are linked to these pathology records. Descriptive statistics are presented outlining the links within and between these datasets.

### Ethics

The project has obtained ethics approval, (HRE2019–0303) from Curtin's Human Research Ethics Committee (HREC) for pathology datasets, and ethics approval from the DoH WA HREC for hospital datasets (RGS000000183).

### Results

Data was received from the WA DoH and three pathology providers at the time of writing, with data from the fourth pathology provider (Western Diagnostic Pathology) expected imminently. A summary description of provided datasets is shown in **Table 1**. Dataset sizes differed depending on how the data was stored; for data stored in a person-based format, each dataset likely contained few duplicates; other datasets were stored in a service-based format, with each episode of care for the same individual recorded separately. The data appeared of generally high quality, with only middle name showing high levels of incompleteness.

A probabilistic linkage was performed to join together records within and between these six datasets. The results of this linkage identified 2,007,309 individuals (excluding those with no pathology record who are not within the study cohort). Among those identified were 1,341,444 individuals with a PathWest pathology record, 1,378,507 individuals with a Clinipath pathology record, and 79,364 people with an ACL pathology record.

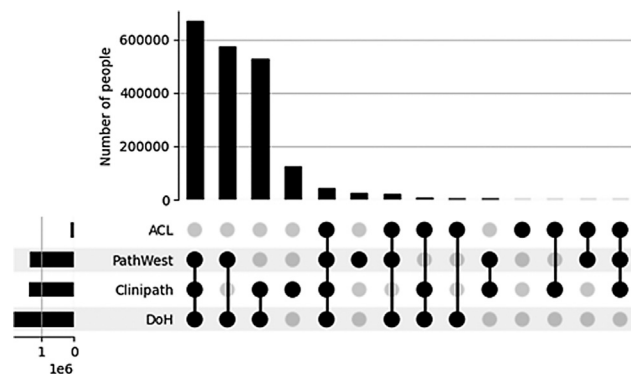
**Table 1** Summary of datasets received and percentage of missing information

	Department of Health			Pathology providers		
	Hospital	Emergency	Mortality	PathWest	Clinipath	ACL
Number of records	19,916,472	18,229,362	270,724	1,727,205	11,352,665	431,169
<b>Fields (% missing)</b>						
Given name	0.0%	0.0%	0.0%	0.1%	0.0%	0.0%
Middle name	53.0%	87.5%	23.1%	33.9%	98.6%	—
Surname	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Sex	0.0%	0.0%	0.0%	0.0%	0.0%	5.0%
Date of birth	0.0%	0.0%	0.2%	0.0%	0.0%	0.0%
Address	0.3%	1.3%	0.3%	0.4%	0.4%	0.0%
Suburb	0.1%	1.1%	0.3%	0.3%	0.4%	0.0%
Postcode	0.0%	0.9%	1.1%	0.0%	0.4%	0.0%

—Fig. 4 shows the number of individuals with records in different combinations of datasets. For instance, over 600,000 individuals in this cohort had data within the PathWest, Clinipath, and DOH collections (and no others), while just under 600,000 individuals had data within just PathWest and DOH collections. Over 55% of individuals with PathWest records had pathology records with other providers; similar results were seen for Clinipath (53% of individuals had records with other pathology providers) and ACL (91% of individuals had records with other pathology providers).

## Discussion

The main challenges to data linkage around the world are not related to the technical aspects of integrating data. They are often centered on more practical elements of data sharing, especially regarding of the release of personal information across organizations. Some barriers to data sharing are related to legislative restrictions but are more often rooted in the organization's security concerns to share data for research, where avoiding data sharing mitigates any associated privacy risk.

**Fig. 4** The number of individuals with records in different combinations of datasets.

Hesitancy toward data sharing is being challenged through open data policies and a need for government to work with private industry to maximize information available on community needs. It is also clear that linkage systems using encryption and other advanced techniques to mask data prior to linkage will make the process of data sharing and linkage easier. Barriers can be further reduced by mitigating risks associated with the data sharing process (through careful planning, secure protocols, legal agreements, and encoding techniques).

In this study, the research team aimed to maximize the use of available health data on CKD by integrating data resources from pathology providers and the WA DoH. By removing the need to use unmasked identifiers for data linkage, PPRL systems have been adapted and scaled with sufficient quality to augment current approaches to data linkage.

The linkage results found here demonstrate the benefits of record linkage approach when utilizing data sourced from pathology providers. Over half of individuals within the cohort had pathology records from more than one provider. Consequently, sourcing and utilizing data from a single provider would provide an incomplete pathological picture for over half of our cohort. This study is the first that used privacy preserving linkage to link data obtained from these pathology providers, whereas previous studies provided datasets that were unlinked and analyzed in isolation. This study contributes to greater acceptance of data linkage techniques among pathology providers and an expected key outcome from this study will be the increased availability of pathology data for linkage for other research studies.

Through the use of privacy preserving linkage methods, we are able to satisfy custodian concerns around privacy of their patient data and join these datasets together, providing a more complete picture necessary for accurate research. By obfuscating personally identifying information before data release, privacy preserving linkage increases privacy as the personal identifiers used for linkage are no longer visible to the third party carrying out linkage and remain solely with

the data custodian. These privacy preserving methods are not impenetrable, with potential attacks<sup>34</sup> on the encodings and their mitigations<sup>35</sup> documented in the literature. In discussions of privacy, it is important to consider the particular context in which the methods are being used. Our use of Bloom filter encodings involves their release to a trusted third party, which has strong information governance standards in place, and with significant legal and contractual safeguards around data transfer and use. Bloom filters are a tool used to reduce the risk of accidental or purposeful re-identification of individuals by the small number of designated users with access to the encoded data. For this purpose, they meet the required security level of data custodians who utilize them.

The results also highlight the importance of linking information from private health care providers. A significantly large percentage of individuals (47%) from Clinipath, a private pathology provider, did not have records with other pathology providers, hence demonstrate private pathology provider's central role in routine health care provision and could therefore provide greater insight into other areas of health in the community.

As all identifying information was encoded, it was not possible to directly check whether a particular pair of records belonged to the same individual, and so it was not possible to calculate quality metrics to gain an indication of the accuracy of linkage. However, the linkage method used here has been evaluated for similar linkages, achieving very high results<sup>31,32,36</sup> including real-world conditions,<sup>30</sup> with false positive rates below 1%. In addition, the linkage was designed to make use of the results of previous clear-text linkage of the Western Australia health data by the WA data linkage branch, which has been linked to a very high standard. Honoring and leveraging these previous links help improve overall linkage quality.<sup>30,37</sup> As such, the resulting linkage quality in this study is expected to be high.

The results presented in this paper provide only a preliminary insight into the potency of these datasets. Data from Western Diagnostic Pathology, the remaining WA pathology provider, was excluded with data arrival for this project expected imminently. Therefore, results provide an incomplete picture, and initial analysis suggests a greater level of overlap between pathology datasets once remaining data are included.

## Conclusion

Privacy preserving linkage gives confidence to data custodians through its ability to provide a secure and efficient access to large integrated datasets. It presents a useful and practical tool to enable integration and access to datasets which was otherwise not available, unlocking potential within these data resources to improve public health.

## Clinical Relevance Statement

PPRL is demonstrated as a viable method of linking pathology data between different pathology providers. Given the

longitudinal nature of CKD, the linked pathology datasets enable the identification of disease which was previously impossible to identify from siloed sets of records from a single provider. PPRL also enabled linkage to hospital records that contain clinical procedural and diagnosis codes in addition to administrative demographic details, which further improve the identification of risk factors associated with CKD and the burden of disease on the health system. This information will help clinicians to prioritize early detection intervention and optimum management for patients, and improve resource allocation within health systems.

## Multiple Choice Questions

1. What is data linkage (also known as medical record linkage)?
  - a. A process of promoting medical research partnerships.
  - b. Process of identifying, matching, and merging records that correspond to the same person/entity from within and/or across several datasets.
  - c. Process that feeds data sandbox environments.
  - d. Process that offers low-cost data storage.

**Correct Answer:** The correct answer is option b. Record linkage is the process of determining records that belong to the same person/entity within and/or between datasets.

2. What is the purpose of Privacy-Preserving Record Linkage (PPRL)?
  - a. To check and ensure the integrity of unit level Personally Identified Data (PII) for linkage.
  - b. To reduce the number of records that need to be released to a Trusted Third Party for linkage.
  - c. To integrate data without the need to release Personally Identified Data (PII) to a Trusted Third Party for linkage.
  - d. To reduce the need for Trusted Third Party organizations to link data.

**Correct Answer:** The correct answer is option c. Privacy preserving record linkage methods allow data to be linked without the release of personally identifying information.

3. What are the benefits of PPRL?
  - a. Reduces reliance on large databases and data lakes.
  - b. Creates a data streaming channel.
  - c. Improves the opportunities to share data across data providers.
  - d. Provides off-load low-usage data availability.

**Correct Answer:** The correct answer is option c. By reducing the privacy risk of record linkage, PPRL can improve opportunities for data to be shared and linked across providers.

### Note

All authors also showed consent for publication and the supplementary data that supports the findings of this review will be available upon submission and publication.

### Protection of Human and Animal Subjects

This project does not involve any animal subjects, and has obtained ethics approval from Curtin's Human Research Ethics Committee (HREC) for pathology datasets (HRE2019-0303), and ethics approval from the DOH WA HREC for hospital datasets (RGS000000183).

### Conflict of Interest

None declared.

### Acknowledgments

This study is supported by the Digital Health CRC Limited (DHCRC) and is part of a larger 4-year collaborative partnership with Curtin University, La Trobe University, WA Department of Health, WA Country Health Service, WA Primary Health Alliance, and the DHCRC. The DHCRC is funded under the Commonwealth's Cooperative Research Centres (CRC) Program. The authors also thank each data provider, WA Department of Health, PathWest, Australian Clinical Laboratories, Clinipath Pathology and Western Diagnostic Pathology, and for their keen interest and support for this study.

### References

- Bikbov B, Purcell CA, Levey AS, et al; GBD Chronic Kidney Disease Collaboration. Global, regional, and national burden of chronic kidney disease, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* 2020;395(10225):709-733
- Luyckx VA, Tonelli M, Stanifer JW. The global burden of kidney disease and the sustainable development goals. *Bull World Health Organ* 2018;96(06):414-422D
- Foreman KJ, Marquez N, Dolgert A, et al. Forecasting life expectancy, years of life lost, and all-cause and cause-specific mortality for 250 causes of death: reference and alternative scenarios for 2016-40 for 195 countries and territories. *Lancet* 2018;392(10159):2052-2090
- Obrador GT, Schultheiss UT, Kretzler M, et al. Genetic and environmental risk factors for chronic kidney disease. *Kidney Int Suppl* (2011) 2017;7(02):88-106
- Chartier MJ, Tangri N, Komenda P, et al. Prevalence, socio-demographic characteristics, and comorbid health conditions in pre-dialysis chronic kidney disease: results from the Manitoba chronic kidney disease cohort. *BMC Nephrol* 2018;19(01):255
- Silva Junior GBD, Oliveira JGR, Oliveira MRB, Vieira LJES, Dias ER. Global costs attributed to chronic kidney disease: a systematic review. *Rev Assoc Med Bras* 2018;64(12):1108-1116
- Cisek K, Krochmal M, Klein J, Mischak H. The application of multi-omics and systems biology to identify therapeutic targets in chronic kidney disease. *Nephrol Dial Transplant* 2016;31(12):2003-2011
- Liyanage T, Ninomiya T, Jha V, et al. Worldwide access to treatment for end-stage kidney disease: a systematic review. *Lancet* 2015;385(9981):1975-1982
- Wong LY, Liew AST, Weng WT, Lim CK, Vathsala A, Toh MPH. Projecting the burden of chronic kidney disease in a developed country and its implications on public health. *Int J Nephrol* 2018; 2018:5196285
- Levin A, Tonelli M, Bonventre J, et al; ISN Global Kidney Health Summit participants. Global kidney health 2017 and beyond: a roadmap for closing gaps in care, research, and policy. *Lancet* 2017;390(10105):1888-1917
- Nøhr C, Parv L, Kink P, et al. Nationwide citizen access to their health data: analysing and comparing experiences in Denmark, Estonia and Australia. *BMC Health Serv Res* 2017; 17(01):534
- Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in health-care: management, analysis and future prospects. *J Big Data* 2019; 6(01):54
- Zeng X-X, Liu J, Ma L, Fu P. Big data research in chronic kidney disease. *Chin Med J (Engl)* 2018;131(22):2647-2650
- Mendu ML, Ahmed S, Maron JK, et al. Development of an electronic health record-based chronic kidney disease registry to promote population health management. *BMC Nephrol* 2019;20(01):72
- Kohli R, Tan SS-L. Electronic health records: how can IS researchers contribute to transforming healthcare? *Manage Inf Syst Q* 2016;40(03):553-573
- Hsieh C-Y, Su C-C, Shao S-C, et al. Taiwan's national health insurance research database: past and future. *Clin Epidemiol* 2019;11:349-358
- Häggglund M, DesRoches C, Petersen C, Scandurra I. Patients' access to health records. *BMJ* 2019;367:l5725
- Ishikawa H. [Japan] health database in an information society. *Japan Med Assoc J* 2014;57(04):192-195
- Australian Institute of Health and Welfare. 2018Australia's health 2018. 2018. . Australia's health series no 16 at: Accessed June 6, 2022 at: <https://www.aihw.gov.au/getmedia/7c42913d-295f-4bc9-9c24-4e44eff4a04a/aihw-aus-221.pdf>
- Hagar Y, Albers D, Pivovarov R, Chase H, Dukic V, Elhadad N. Survival analysis with electronic health record data: experiments with chronic kidney disease. *Stat Anal Data Min* 2014;7(05):385-403
- Royal Australian College of General Practitioners. Privacy and managing health information in general practice. 2017, at: <https://www.racgp.org.au/FSDEDEV/media/documents/Running%20a%20practice/Protecting%20practice%20-information/Privacy-and-managing-health-information-in-general-practice.pdf>
- Ferrante A, Eitelhuber T, Maller M, et al. Cooperation and innovation in data linkage creates a linked, multi-sectoral data repository for Western Australia – the SIDR project. *Int J Popul Data Sci* 2020;5(05):
- Brown AP, Borgs C, Randall SM, Schnell R. Evaluating privacy-preserving record linkage using cryptographic long-term keys and multibit trees on large medical datasets. *BMC Med Inform Decis Mak* 2017;17(01):83
- Brown AP, Ferrante AM, Randall SM, Boyd JH, Semmens JB. Ensuring Privacy When Integrating Patient-Based Datasets: New Methods and Developments in Record Linkage. *Perspective Front Publ Health*; 2017:5
- Vatsalan D, Christen P, Verykios VS. A taxonomy of privacy-preserving record linkage techniques. *Inf Syst* 2013;38(06): 946-969
- Nitzlader M, Schreier G. Patient identity management for secondary use of biomedical research data in a distributed computing environment. *Stud Health Technol Inform* 2014;198:211-218
- Correll P, Feyer A-M, Phan P-T, et al. Lumos: a statewide linkage programme in Australia integrating general practice data to guide system redesign. *Integrated Healthcare J* 2021;3(01):e000074
- Lopez-Vargas PA, Tong A, Phoon RKS, Chadban SJ, Shen Y, Craig JC. Knowledge deficit of patients with stage 1-4 CKD: a focus group study. *Nephrology (Carlton)* 2014;19(04):234-243
- Myers GL, Miller WG, Coresh J, et al; National Kidney Disease Education Program Laboratory Working Group. Recommendations for improving serum creatinine measurement: a report from the Laboratory Working Group of the National Kidney Disease Education Program. *Clin Chem* 2006;52(01):5-18
- Randall S, Wichmann H, Brown A, et al. A blinded evaluation of privacy preserving record linkage with Bloom filters. *BMC Med Res Methodol* 2022;22(01):22



- 31 Randall SM, Ferrante AM, Boyd JH, Bauer JK, Semmens JB. Privacy-preserving record linkage on large real world datasets. *J Biomed Inform* 2014;50:205–212
- 32 Randall SM, Brown AP, Ferrante AM, Boyd JH. Privacy preserving linkage using multiple match-keys. *Int J Popul Data Sci* 2019;4(01):1094–1094
- 33 Boyd JH, Randall SM, Brown AP, et al. Population data centre profiles: centre for data linkage. *Int J Popul Data Sci* 2020;4(02):1139
- 34 Christen P, Schnell R, Vatsalan D, Ranbaduge T. *Efficient Crypt-analysis of Bloom Filters for Privacy-Preserving Record Linkage*. Springer International Publishing; 2017:628–640
- 35 Ranbaduge T, Schnell R. Securing Bloom Filters for Privacy-preserving Record Linkage. Paper presented at: Proceedings of the 29th ACM International Conference on Information & Knowledge Management 2020; Virtual Event, Ireland
- 36 Irvine K, Smith M, Vos R, et al. Real world performance of privacy preserving record linkage. *Int J Popul Data Sci* 2018;3(04). Doi: 10.23889/ijpds.v3i4.990
- 37 Boyd JH, Randall SM, Ferrante AM, Bauer JK, Brown AP, Semmens JB. Technical challenges of providing record linkage services for research. *BMC Med Inform Decis Mak* 2014;14(01):23