Data Resource Profile

Data Resource Profile: Victorian Comprehensive Cancer Centre Data Connect

Alex Lee,1,* Damien McCarthy,1 Rebecca J Bergin,1 Allison Drosdowsky,1 Javiera Martinez Gutierrez,1,3 Chris Kearney,1 Sally Philip,1 Meena Rafiq,1,2 Brent Venning,1 Olivia Wawryk,1 Jianrong Zhang,1 and Jon Emery 1

1Department of General Practice, Faculty of Medicine, University of Melbourne and Centre for Cancer Research, Parkville, VIC, Australia, 2Epidemiology of Cancer and Healthcare Outcomes (ECHO) Group, UCL, London, UK and 3Department of Family Medicine, School of Medicine, Pontificia Universidad Católica de Chile, Santiago, Chile

*Corresponding author: Department of General Practice, University of Melbourne, 780 Elizabeth St, Parkville, VIC 3010, Australia. E-mail: alexalee@unimelb.edu.au

Keywords: Primary care data, data linkage, cancer services research.

Key Features

• Data Connect has been established as a data linkage resource that brings together Australian General Practice (GP), Metropolitan Hospital and Cancer Registry data sources to enable researchers to address questions in health services research related to the entire patient continuum of care.
• This resource includes patient-level de-identified data on over 3 million Victorian primary care patients aged over 18 years of age, with GP data covering the period 1996–2021 and hospital data covering the years 2006–21.
• Data have been linked across health services as this is a requirement for understanding patients’ journey of care, in particular for cancer patients, who interact with many different parts of the health care system, often starting at primary care.
• The resource includes rich clinical and administrative data from hospital presentations and admissions, GP encounters, prescriptions, pathology test results and observations and cancer diagnoses.
• Researchers interested in using this resource should contact Data Connect at [https://vcccalliance.org.au/our-work/research-and-translation/data-connect].

Data resource basics

Victorian Comprehensive Cancer Centre Data Connect

Victorian Comprehensive Cancer Centre (VCCC) Data Connect is a single platform for facilitating access to a range of Victorian health data sources, allowing researchers, analysts and policy makers to bring a rigorous, data-driven lens to questions related to health services and patterns of cancer care. It has been developed in Australia since 2017 to connect general practice databases with data from hospitals and clinical cancer registries, and involves nearly 3 million patients spanning the years 1996–2021. It is updated on a yearly basis to incorporate more recent datasets, and is being expanded to include additional health services datasets across the state, including from non-metropolitan hospitals, VCCC Alliance members and clinical registries. In this paper we provide high-level descriptive statistics of the data sources, summarizing representativeness, data quality and overall sizes of patient cohorts. As a specific example we then focus on general cancer statistics, to illustrate the utility of this resource.

Cancer patients interact with the health system through using many different health services, with primary care playing a critical role from initial presentation through to survivorship and end-of-life care.1,2 Each health service has its own corresponding data collection systems, so linkage is critical to understand how patients interact with the system across their entire journey of care.3

In Australia, health data linkage is challenging due to different parts of the health system being the responsibility of different jurisdictions at state and federal levels, as well as due to differing data standards and systems.4 Since primary care is usually the first point of contact that patients have with the health system,5 including prior to a diagnosis of cancer, General Practice (GP) data are a key component of research in this area.

The intended audience of this paper is researchers interested in data, research and analysis related to health service use.
Research studies using these data are already under way, including in the study of times to diagnosis and treatment, pathways of care, early detection of cancer and cancer survivorship. Whereas our particular focus is on cancer, Data Connect contains a wealth of information on other cohorts of patients with broader applications, and the resource is of value for health services researchers in different clinical domains.

Other linked data resources

Linked data resources are becoming increasingly important for health research in many countries.6–8 In the UK, the Clinical Practice Research Datalink (CPRD) is one such resource, covering almost 7% of the UK population, and has been used in over 1000 studies.6 In Australia, state and federal governments have their own linked data infrastructure. Examples include the Centre for Victorian Data Linkage,9 Data Linkage Western Australia,10 the Australian Institute of Health and Welfare11 and also the Australian Bureau of Statistics’ own linkage capabilities.12 However in Australia, primary care is often overlooked in linked data research, despite its critical role in the health system.4,5 For example, in the 2020–21 financial year, over 85% of Australians had at least one contact with a GP service.13 A key source of general practice data is the Medicinsinl database, a representative database covering the entire country.14 Another is the University of Melbourne Patron Data for Decisions Program.4 Few studies have involved linking health services datasets to these types of general practice data sources that contain rich clinical information in addition to billing information.15

Data governance, ethics, privacy and security

Data access is provided by BioGrid Australia,16–18 a not-for-profit organization that provides an online platform for data linkage and management, so data storage and linkage infrastructure are handled by a single entity. Data linkage in Data Connect is carried out through the use of a Unique Subject Identifier (USI), a unique encrypted key for each patient. For hospital data these USI values are generated by processing patient-identifying information through BioGrid’s probabilistic record linkage key software and using the GRHANITE™ data linkage software,19 whereas for GP data these are generated at the practice using GRHANITE™. The USI effectively anonymizes patient records as no identifying information is released to researchers, protecting patient privacy and ensuring that data governance, ethics and privacy requirements of data custodians that are met.

Data Connect uses the BioGrid federated data infrastructure, which means that it is covered by BioGrid’s data governance model.18 A collaboration agreement among BioGrid member institutions ensures that there is no need to create multiple agreements for different health service sites in order to apply for data access.16 In addition, the human research ethics framework is managed by BioGrid, so that data access, governance and management across sites are covered by a single agreement, removing the need for researchers to approach data custodians individually and ensuring more timely access to data for research. It is important to note that data custodian and ethical approval is required for each research project prior to linked data being made available to researchers through BioGrid.

Data linkage allows researchers, analysts and policy makers to capture a more complete picture of patients’ interactions with the health care system.3 Primary care has a critical role through the patient’s continuum of care,1,2 as illustrated in the Figure 1, and linking data from general practice, hospitals and clinical registries allows tracking of patients along this continuum.

Figure 1. Stages in the continuum of care and relevant data sources. Patron, University of Melbourne general practice database; VAED, Victorian Admitted Episodes Dataset; VINAH, Victorian Integrated Non-Admitted Health dataset

Figure 2. Distribution of patients by age and sex in Medicinsinl vs 2016 Census (left) and Patron vs 2016 Census (right). ABS, Australian Bureau of Statistics; Patron, University of Melbourne general practice database
<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source of data</th>
<th>Variables available (not exhaustive)</th>
<th>Variables not available (not exhaustive)</th>
</tr>
</thead>
</table>
| MedicineInsight (General Practice database, Victorian subset) | MedicineWise | • Encounter date  
• Patient residence (at SA3)  
• Patient age  
• Patient sex  
• Health care card status  
• Alcohol status  
• Smoking status  
• ATSI status  
• Requested tests  
• Test results  
• Prescriptions  
• Billing  
• Observations such as weight and height measurements  
• Reason for encounter, including signs and symptoms  
• Conditions  
• Diagnoses  
• Comorbidities are captured in the reason for encounter, conditions and diagnoses free-text fields | • Clinical notes  
• Time of encounter  
• Practice location  
• Patient ethnicity  
• Patient country of birth |
| Patron (University of Melbourne general practice database) | Practices recruited via University of Melbourne Data for Decisions Program | • Encounter date  
• Patient residence (at SA3)  
• Patient age  
• Patient sex  
• Health care card status  
• Alcohol status  
• Smoking status  
• ATSI status  
• Requested tests  
• Test results  
• Prescriptions  
• Billing  
• Observations such as weight and height measurements  
• Reason for encounter, including signs and symptoms  
• Conditions  
• Diagnoses  
• Comorbidities are captured in the reason for encounter, conditions and diagnoses free-text fields | • Clinical notes  
• Time of encounter  
• Practice location  
• Patient ethnicity  
• Patient country of birth |
| VAED (Victorian Admitted Episodes dataset) | Victorian public hospitals via BioGrid | • Date of admission  
• Admission site  
• ICD-10 diagnosis  
• Diagnosis date  
• Separation dates  
• Patient age  
• Patient sex  
• CALD indicators | • Clinical notes |
| VEMD (Victorian Emergency Minimum Dataset) | Victorian public hospitals via BioGrid | • Presentation date  
• Presentation site  
• Patient postcode  
• CALD indicators | • Clinical notes |
| VINAH (Victorian Integrated Non-Admitted Health dataset) | Victorian public hospitals via BioGrid | • Episode date  
• Episode site  
• Patient postcode | • Clinical notes |
| Victorian Cancer Registry Pilot (Sample of patients from the Victorian Cancer Registry) | Cancer Council Victoria | • Diagnosis date  
• Patient age  
• Patient sex  
• Patient date of birth  
• Patient country of birth  
• Patient residence (at SA2)  
• Cancer diagnosis code (ICD-10-AM)  
• Tumour site  
• Tumour stream | • SA1 geography  
• Limited stage data |
Figure 1 illustrates the relation between points along the cancer continuum of care and corresponding datasets that are available through Data Connect. In this case we provide the example of a patient presenting to general practice with a symptom, and points along their care pathway until diagnosis and treatment. Different events are captured in different electronic health records systems, so linkage is crucial to fully understand patients’ interactions.

Questions related to early diagnosis of cancer involve general practice and hospital data, and questions related to survivorship relate to both earlier and later stages.

Data collected

Data Connect currently contains the following datasets:

- MedicineInsight: General Practice database (Victorian subset); MedicineWise14,20;
- VAED: Victorian Admitted Episodes Dataset (four hospital networks);
- VEMD: Victorian Emergency Minimum Dataset (three hospital networks);
- VINAH: Victorian Non-Admitted Episodes Health dataset (two hospital networks);
- Patron: University of Melbourne general practice database4;
- VCR Pilot: A sample of patients from the Victorian Cancer Registry;
- ACCORD: Australian Comprehensive Cancer Outcomes and Research Database (colorectal and sarcoma subsets);
- AURORA: AUstralian Registry and biObank of thoRAcic cancers.

Additional work is under way to link radiotherapy data, additional cancer registry data and genomics datasets. The datasets contain patient demographic information as well as

Table 1. Continued

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Source of data</th>
<th>Variables available (not exhaustive)</th>
<th>Variables not available (not exhaustive)</th>
</tr>
</thead>
<tbody>
<tr>
<td>AURORA</td>
<td>Peter MacCallum Cancer Centre</td>
<td>• Diagnosis date • Treatment date • Patient age • Patient sex • Tumour stage • Treatment type and usage • Comorbidities • Date of death</td>
<td>• Patient residence • Limited symptom data</td>
</tr>
<tr>
<td>ACCORD Colorectal</td>
<td>Royal Melbourne Hospital and Western Hospital</td>
<td>• Diagnosis and treatment dates • Patient age • Patient sex • Patient weight • Tumour stage • Chemotherapy drugs used • Dose • Radiotherapy fractions • Reasons for changes in dose • Resections • Comorbidities • Cancer history</td>
<td>• Patient residence • Limited symptom data</td>
</tr>
<tr>
<td>ACCORD Sarcoma</td>
<td>Peter MacCallum Cancer Centre</td>
<td>• Patient age • Patient sex • Cancer history • Death date • Referring specialty • First consultation details • Symptom history • Pre-treatment imaging status • Diagnosis and treatment dates • Tumour stage • Biopsy details • Resection details • Surgery details • Radiotherapy details • Chemotherapy details</td>
<td>• Limited symptom data</td>
</tr>
</tbody>
</table>


4 These are recorded in free-text and mapped by MedicineWise. They are not based on the International Classification of Primary Care (ICPC) standard.
Data related to patient interactions with health services. Table 1 provides a more detailed (though not exhaustive) list of information available. We also highlight certain items that are not available, mainly due to privacy. Further detailed information on each of these is available in the references. Note that the National Prescribing Service ceased operation at the end of 2022 and the primary care database is now managed by the Australian Commission on Safety and Quality in Health Care.

Data quality

Table 2 summarizes some general data quality issues in the datasets that make up Data Connect. More detailed data quality work for each of the data sources can be found in the references. In any case, for a given project, data quality assessment should be first carried out based on the specific research question as part of feasibility assessment. The statistics reported here should help researchers understand some of the major data quality issues that occur generally. Note that both MedicineInsight and Patron are drawn from similar Electronic Medical Record (EMR) systems—Best Practice and Medical Director—with Patron also containing data from ZedMed. As a result, the data are similar, with similar issues and limitations. Around 580,000 patients are common to Patron and MedicineInsight.

Coverage and representativeness

More detailed statistics of datasets in Data Connect are shown in Table 3. Geographical coverage, described in Table 4, is greater for the general practice datasets, covering all parts of Victoria, with less coverage in the eastern part of metropolitan Melbourne. In 2019 there were 1985 general practices in Victoria, so approximately 10% of general practices in Victoria are represented. For the hospital datasets, no hospitals outside the metropolitan area are currently available.

Figure 2 shows the age distribution of patient populations in Patron and MedicineInsight. MedicineInsight data closely matches the broader Victorian population in terms of distribution of age and sex, though there are a larger proportion of younger patients, particularly female patients. This is consistent with other studies showing that women are more likely to visit a GP. Ages below 18 are not shown as there is no current ethical approval to include under-18s in Data. 

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Variable</th>
<th>Result</th>
<th>Specific data quality issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>MedicineInsight, Patron</td>
<td>Reason for encounter</td>
<td>A large proportion of these are free-text; drop-down boxes are from bespoke medical terminology lists, rather than standardized ones</td>
<td>Work needs to be done to map the ‘reason for encounter’ fields to standard medical terminology, this is under way for the Patron database</td>
</tr>
<tr>
<td>Age</td>
<td>Differences in age distribution between GP data and census</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sex</td>
<td>More women than men represented</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SEIFA</td>
<td>Distribution of patients agrees well with Census 2016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Indigenous status</td>
<td>Not a reliable variable; does not agree with 2016 Census</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SA3 of patient residence</td>
<td>For MedicineInsight counts agree well with those reported in Census 2016</td>
<td></td>
<td>Corresponds to SA3 boundaries for 2020; patient may have changed residence</td>
</tr>
<tr>
<td>Conditions</td>
<td>Data regarding time a condition was diagnosed is often unreliable</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAED</td>
<td>Local Government Area catchment</td>
<td>Not representative</td>
<td>Currently concentrated in metropolitan Melbourne</td>
</tr>
<tr>
<td>VEMD</td>
<td>Local Government Area catchment</td>
<td>Not representative</td>
<td>Currently concentrated in metropolitan Melbourne</td>
</tr>
<tr>
<td>VINAH</td>
<td>Local Government Area catchment</td>
<td>Not representative</td>
<td>Currently concentrated in metropolitan Melbourne</td>
</tr>
<tr>
<td>VCR Pilot</td>
<td>SA2</td>
<td>Not representative</td>
<td>Contains only a random sample of data that can be linked to Royal Melbourne Hospital and no other hospitals</td>
</tr>
<tr>
<td>AURORA</td>
<td>Stage at diagnosis</td>
<td>Not representative if only using the data from one centre</td>
<td>Data from Peter MacCallum Cancer Centre are predominantly for late-stage disease; data from St Vincent’s Hospital are predominantly for early-stage disease</td>
</tr>
<tr>
<td>ACCORD Colorectal</td>
<td>Cancer history</td>
<td>High percentage of unknown and/or missing data</td>
<td>Previous cancer diagnosis information not available</td>
</tr>
</tbody>
</table>

ACCORD, Australian Comprehensive Cancer Outcomes and Research Database (colorectal subset); AURORA, AUstralian Registry and biObank of thoRAcic cancers; MedicineInsight, General Practice database (Victorian subset); Patron, University of Melbourne general practice database; SA3, Statistical Area Level 3; SA2, Statistical Area Level 2; SEIFA, Socio-Economic Index For Areas; VAED, Victorian Admitted Episodes dataset; VCR Pilot, a sample of patients from the Victorian Cancer Registry; VEMD, Victorian Emergency Minimum Dataset; VINAH, Victorian Integrated Non-Admitted Health dataset.
Connect. For example, BioGrid receives under 18s data from MedicineInsight, but we filter these patients out as there is no current ethical approval to include children.

**Types of cancers and prevalence within the linked data**

Cancer diagnoses are recorded in registries as well as hospital admission episodes data. From this, the prevalence of different types of cancers in a specific patient cohort can be calculated: in this case, patients who were admitted to one of the four metropolitan hospitals for which data are available through Data Connect. A total of 79 277 patients are diagnosed with cancer in VAED data, 8.9% of the total.26 The ICD-10-AM (International Classification of Diseases, Tenth Revision, Australian Modification) codes have been used to filter to the specific cancer stream. The results are shown in Table 5.

**Data structure and extraction**

The GP data is extracted from three Electronic Medical Record systems: Best Practice; Medical Director; and ZedMed. The data linkage software GRHANITETM is installed at each of the contributing practices. Hospital data sources are extracted from the EMR systems at Metropolitan Melbourne hospitals and are then provided to BioGrid for linkage.
Data resource use

There are many studies currently under way that make use of Data Connect. The focus of each involves a different part of the continuum of care for cancer patients. Two examples include the following.

- **Diagnostic intervals:** time to diagnosis is an important factor driving survival rates of cancer patients.\(^{27,28}\)
  
  Using data linked between MedicineInsight, Patron and the AURORA and ACCORD clinical registries, researchers have been studying diagnostic intervals for lung (268 patients) and colorectal cancers (273 patients).

- **Primary care blood test use and early detection of cancer:** in the 6 months leading up to a diagnosis of cancer, population-level trends in pathology tests and signs and symptoms become apparent. These can facilitate the early detection of cancer through assisting GPs to focus on those clinical features that are most likely to be predictive of cancer.

Finally, there is currently a lack of standards in Australia for defining phenotypes from electronic health records. Phenotyping libraries are publicly accessible catalogues of disease and other health-related definitions, which ensure reproducibility of research outputs. In the UK, Health Data Research (HDR) UK\(^{29}\) has produced an example of such a library. PhEKB (the Phenotype Knowledge Base) developed in the USA is another example.\(^{30}\) A recent review defined a set of desiderata for a high-quality phenotype library.\(^{31}\) Due to differences in the way that electronic health care data are recorded in Australia, a similar resource would be very valuable here as well. This would provide a way for researchers to share reusable definitions that appear in observational studies involving Victorian health records, and would better ensure reproducible, high-quality research outputs.

Strengths and weaknesses

**Strengths**

Key strengths of Data Connect include the detailed unit-record-level data on over 3 million patients in Victoria and its broad applicability to a range of questions in health services research. It brings together general practice, hospital and clinical registry data and is particularly valuable for studying the entire patient journey from a patient’s first presentation at a health service through to diagnosis, treatment, survivorship and end-of-life care. Rich clinical information is available about test results, observations, signs, symptoms, dates of events, locations and demographic information such as age, sex and place of residence, allowing study of many different questions involving all stages in the continuum of cancer care. The examples of types of questions we are examining in colorectal cancer demonstrates the utility of these linked datasets.

**Weaknesses**

Some limitations of Data Connect include the following.

- The quality and quantity of data varies across different variable types and between datasets, which may limit the feasibility of certain studies particularly, for example, less common cancers.
- Currently only some hospitals in metropolitan Melbourne are represented and over 200 general practice sites provide data. This is just over 10% of those in Victoria.\(^{24}\) There are two arising limitations: incomplete data on...
hospitalizations and emergency presentations, and limited size of cohorts for which general practice data can be linked.

- More generally in relation to cohort size, although there are many patients represented across all datasets, for a given cohort this number is significantly smaller, as demonstrated in the colorectal cancer example. In comparison with jurisdictions such as the UK, Victoria’s population is much smaller, so this can be a limitation to studies requiring larger sample sizes, for example where survival may be the outcome of interest for investigating specific subgroups.

Data resource access

For further information about access to Data Connect data, please contact Sally Philip, Program Manager, VCCC Data Connect, and Alex Lee, Research Fellow, via the Data Connect website at [vcccalliance.org.au/our-work/research-and-translation/data-connect/contact-us].

Ethics approval

This work was approved by Melbourne Health HREC.

Author contributions

A.L. wrote the initial manuscript and created the overall structure. D.M. provided summary statistics for the hospital and primary care data sources. S.P. provided input into the governance sections, and all authors reviewed the entire paper and provided statistics regarding various data sources presented through the paper, including the cancer registries.

Funding

This work was supported by the Bupa Foundation and Victorian Department of Health through the Victorian Comprehensive Cancer Centre Alliance.

Acknowledgements

The authors thank BioGrid Australia for their ongoing work in developing this very useful resource and providing valuable input to this paper, as well as the VCCC. This research used de-identified patient data from the Patron primary care data repository (extracted from consenting general practices), which has been created and is operated by the Department of General Practice, University of Melbourne [www.gp.unimelb.edu.au/datafordecisions].

Conflict of interest

The authors have no conflicts of interest.

References


