

International Journal of Population Data Science

Journal Website: www.ijpds.org



Using graph theory to flexibly construct patient journeys in linked healthcare data

Ian Powell^{1,*}, Zhisheng Sa¹, Branislav Igic¹, Maria Alfaro-Ramirez¹, Rachel Farber¹, and Michael Nelson¹

Submission History

Submitted:	14/11/2024
Accepted:	07/05/2025
Published:	09/10/2025

¹NSW Ministry of Health, Locked Bag 2030, St Leonards NSW 1590, Australia

Abstract

Introduction

Studies of epidemiology and health system use that use linked admitted patient data benefit from understanding the patient journey, particularly when it spans multiple records within or across multiple datasets.

Objectives

To develop a flexible method for grouping together administrative admitted patient records into periods of hospital care that follow patients from admission to discharge.

Methods

We describe a flexible and generalisable graph theoretic algorithm for grouping patient records into periods of hospital care. The algorithm can account for a variety of complex hospitalisation patterns involving multiple transfers and overlapping records. An R package, *journeyer*, that implements this algorithm, is included in the Supplementary Material.

Results

This algorithm was applied to the New South Wales Admitted Patient Data Collection, finding 21,405,451 periods of hospital care from 22,794,746 hospital records. The parameters and decisions required for this algorithm were assessed and found appropriate for this dataset, but we offer some advice for generalisation to other datasets.

Conclusions

Our method assists in preparing data for epidemiological research in New South Wales and can be generalised to inpatient data in other jurisdictions. The method can be extended to include ambulance and emergency department data.

Keywords

linked data; health administrative data; linked records; software; period of hospital care; patient journey; inpatient; hospitalisation; graph theory

*Corresponding Author:

Email Address: Ian.Powell@liverpool.ac.uk (Ian Powell)



Introduction

Understanding the patient journey is important for epidemiological studies and health system planning to accurately capture all the treatment a patient receives for a condition, and relevant context to each health service interaction. The patient journey covers the experiences of a patient through a health complaint, including their interactions with health services. Administrative data are created in the day-to-day provision of services, and healthcare administrative data are a vital source of information for understanding the patient journey because they can capture dates, times, and locations of health service interactions. Furthermore, healthcare data can capture useful variables for epidemiology, such as the conditions or diseases that a patient presents with, or the treatment that a patient receives. Administrative datasets can also capture clinical outcomes such as death or referral to other services, which provide more insight into a patient's experiences. Analysis of the patient journey helps health services detect patients that benefit from continuity of care.

Administrative data can have limited information on a patient's journey, requiring deduction from other available information. For example, a record might indicate that a patient was transferred to another hospital but may not specify the destination hospital record. This is the case in the Admitted Patient Data Collection (APDC) of New South Wales (NSW), a state in Australia. When a patient is admitted to a NSW hospital, an admitted patient record is created in the patient administration system. When the patient is discharged, information from the medical record is coded into the admitted patient record. A patient journey can include stays at multiple hospitals and different types of care within the same hospital stay, such as a change from acute care to rehabilitative care (statistical discharge). Each of these episodes results in a new admitted patient record or *episode of care*. A patient's journey from admission to discharge can have multiple episodes of care, which may not be in a linear sequence. To analyse patient journeys using APDC data, continuous episodes need to be chained together using available limited information.

Previous studies have examined patient journeys in a variety of ways. Several recent studies using admitted patient data have not specified how transfers or overlapping episodes are handled, to the standard required for replication [1–4]. Other studies grouped together “contiguous” episodes allowing up to 24 hours [5, 6], or an unspecified time period [7, 8], between episodes. However, without a clear definition of contiguous, these methods are difficult to replicate. Gubbels et al. [9] provided an explicit algorithm to clean Danish National Patient Register (DNPR) inpatient data and construct patient journeys from multiple episodes. This algorithm defined transfers based on the time between episodes and could be improved upon using information on mode of separation and transfer origin/destination when available. However, this algorithm was designed specifically for the DNPR and was not applicable to other datasets. Elliss-Brookes et al. [10] used an algorithm to classify patient journeys leading to diagnosis of cancer across multiple administrative datasets. However, this algorithm was designed for a specific research question and may not be generally applicable.

Graph theory is an area of mathematics that examines relationships between pairs of arbitrary objects. In a graph,

each of these arbitrary objects is represented by a *node*, and if a pair of objects are related according to a given definition, then their nodes are joined by an *edge*. Network theory is a subdiscipline of graph theory that represents various networks as graphs, and is primarily interested in analysing the patterns that occur in a network. In healthcare settings, graph and network theory have been used to analyse patterns of referral [11], networks of patient exchange between institutions, and collaboration between physicians and disease co-occurrence [12]. In this paper, we use graph theory to formulate the problem of uncovering individual patient journeys from admission to discharge.

Our aim is to develop a method that enables researchers to group multiple records relating to the patient journey from formal admission to formal discharge together. When health-related events create multiple administrative inpatient data records, these groupings can be used to prevent overcounting or provide units for further statistical analysis. Our method uses graph theory to group all records from a patient's formal admission to formal discharge into a single *period of hospital care* (POHC). POHCs account for transfers within and between hospitals. This method can be flexibly applied to a wide range of administrative inpatient datasets. Researchers can also extend our method to aggregate patient journeys through other patient-related administrative datasets, including ambulance, emergency department, and general practitioner data, and incorporate contextual information not directly related to patient health, such as public transport or air travel.

Methods

In this section, we outline the algorithm used to build POHCs and describe how we selected the parameter values for the NSW APDC. Admitted patient data can be represented as a graph, where each node represents an episode of care, and edges represent direct connections from one episode to another without formal discharge, such as patient transfers. Then, if one episode can be reached by following connections from another episode, then these two episodes are part of the same patient journey from formal admission to formal discharge.

Data sources

NSW Admitted Patient Data Collection

The NSW Admitted Patient, Emergency Department and Deaths Register (APEDDR) is a public health register established in 2013 under the *Public Health Act 2010* (NSW). APEDDR contains administrative records of people attending hospital, and deaths in NSW. It captures data relating to emergency department presentations, inpatient admission episodes, death registrations and causes of death. Personal identifiers in these datasets are removed and replaced with a project person number created using probabilistic linkage by the Centre for Health Record Linkage (CHeReL). Data dictionaries for these datasets are available on the CHeReL website [13]. The focus of this article is on the APDC from 1 July 2014 to 30 June 2022.

A small, synthetic example of the relevant columns of the NSW APDC captures some of the patterns of care present in

this dataset (Table 1). We refer to this example dataset to aid understanding of the methods we use. This dataset contains the records of three fictional patients at fictional facilities A, B and C.

Defining periods of hospital care with graph theory

We developed a three-step algorithm to construct POHCs from a set of admitted patient records (Figure 1). First, a computational graph was initialised with nodes representing admitted patient records, and no edges. Second, every pair of records was compared and checked for evidence that the patient directly moved from one record to another without formal discharge; a pair of records related in this way was *connected*. Every connected pair was added to the list of edges in the graph. Finally, we assigned a unique POHC number to each group of records where one record in the group could be reached from another record in the group by following one or more connections. In graph theory terminology, POHCs are connected components.

Evidence of connections between episodes in the NSW APDC

To determine if a pair of episodes were directly connected in the NSW APDC, we used five connection predicates (Table 2, Figure 2). In these predicates, the *parent* episode represented an earlier or overarching episode of care. The *child* episode represented an episode of care that arose from the parent episode, for example through transfers to other hospitals. Typically, a parent episode would occur before a child episode, but there are some rare cases where this may not occur strictly, for example if two episodes begin at the same time.

Every episode was compared to every other episode pairwise. In each pair of episodes, one episode was arbitrarily selected as a candidate parent episode and the remaining as a candidate child episode. Then, each connection predicate in Table 2 compared this candidate parent/child pair, and the pair was connected if any of the predicates held true. Then, the candidate roles were swapped so that predicates for each pair were checked in both parent/child directions. This direction does not affect whether a connection is present or not, but does simplify interpretation of the predicates. Every predicate required that the probabilistically linked project person number matched between episodes.

Temporally overlapping episodes were captured in the first predicate. These overlaps could occur when a patient was admitted to one facility, but some of the patient's care was contracted to another facility. Note that episodes were also considered to overlap if the start time of the child episode was exactly equal to the end time of the parent episode.

Transfers, where a patient's care was continued at a different facility, were captured in three predicates. The "transfers in" predicate captured inter-hospital transfers where the later episode started by receiving a transfer, which was marked by a source of referral coded as another hospital, and the source facility in the later episode matched the facility of the earlier episode. The "transfers out" predicate captured similar transfers, except the earlier episode indicated the patient was transferred at the end of the episode, marked by a

mode of separation of transfer to palliative care unit, hospice, nursing home, public psychiatric hospital, other hospital, or other accommodation. This predicate also required that the earlier episode's destination facility matched the facility of the later episode. The "bidirectional transfers" predicate captured connections between parent episodes with transfer modes of separation and child episodes with transfer sources of referral, with facilities not required to match; this was designed to account for transfers where a patient was redirected in transit. In each case, the maximum amount of time allowed for a transfer was 9 hours.

The final predicate captured type changes or statistical discharges, which occurred when the type of care that a patient received changed during their stay at the same hospital. This predicate required the child episode to start with a type change admission, the parent episode to end with a type change separation, and the facility to be the same in both episodes. In this case, 30 minutes were allowed for a type change to occur.

Worked example

We applied this method to the dataset in Table 1 (Figure 3). First, we only compared episodes of the same patient for evidence of connections, resulting in pairwise comparisons for records (1)–(3) in patient 1000, (4) and (5) in patient 1001, and (6) and (7) for patient 1002. Record (2) started during record (1), satisfying the "overlapping episodes" predicate, and the two records were connected. Records (2) and (3) were roughly a week apart, and no predicates apply. However, records (1) and (3) simultaneously satisfied the "transfer in", "transfer out", and "bidirectional transfer" predicates with record (1) as the parent episode, so this pair was also connected. As record (1) was connected to both records (2) and (3), and these three records had no other connections to any other records, they all formed a single POHC. Notably, while the mode of separation for record (2) was "discharge", which generally indicates a patient returning home, this was not the end of the POHC. Record (5) started at the same time as record (4) ended, satisfying the "overlapping episodes" predicate. Records (4) and (5) also satisfied the "type change" predicate with record (4) as the parent. Because this pair satisfied at least one predicate, it was connected. Hence, records (4) and (5) form another POHC. Records (6) and (7) occurred roughly a week apart, and no predicates applied. In this case, record (6) formed its own POHC, as did record (7).

Selecting time allowed for transfers

The maximum time allowed for inter-hospital transfers was a key parameter in this algorithm, and so we examined transfer times to justify our choice of 9 hours. To do this, we defined *48-hour transfers* as connections between pairs of episodes that simultaneously satisfied the "transfer in" and "transfer out" predicates of Table 2, with the modification that 48 hours were allowed for a transfer to take place. This 48-hour limit was chosen arbitrarily as one that is longer than the time taken for any reasonable transfer. We then plotted the distribution of time between the end of the parent episode and the start of the child episode.

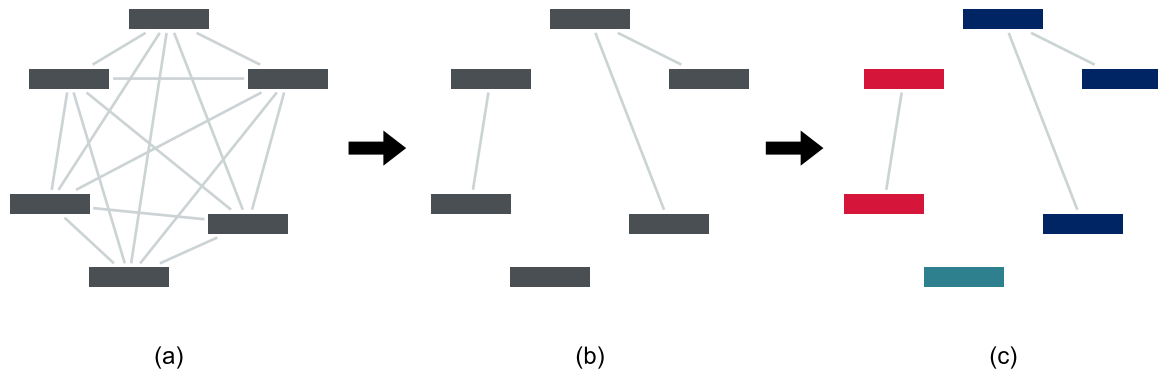
Transfers from regional and rural areas of NSW to Sydney, the state's capital and metropolitan hub, were expected

Table 1: Synthetic data imitating some admission patterns present in the New South Wales Admitted Patient Data Collection

Record ID	Patient number	Episode start	Episode end	Source of referral	Mode of separation	Facility	Facility to	Facility from
1	1000	2022-04-13 09:56	2022-04-21 10:15	Emergency department	Transfer	A	B	
2	1000	2022-04-14 10:45	2022-04-14 15:35	Other Health Agency	Discharge	C		
3	1000	2022-04-21 11:40	2022-04-25 12:30	Transfer	Discharge	B		A
4	1001	2022-09-26 14:20	2022-09-29 10:00	Emergency department	Type change separation	B		
5	1001	2022-09-29 10:00	2022-10-02 14:15	Type change admission	Discharge	B		
6	1002	2022-01-25 17:25	2022-01-26 08:15	Emergency department	Discharge	A		
7	1002	2022-02-01 21:05	2022-02-03 09:10	Emergency department	Discharge	A		

Each record has a unique identifier (ID) and a patient ID. For simplicity, records are ordered by patient ID and episode start time. Blank cells should be regarded as missing.

Figure 1: Representation of method to construct periods of hospital care. Each box represents an admitted patient data record and corresponds to a node in the graph



(a) All pairs of records are compared to one another. (b) Each pair is checked for evidence of a connection and rejected if no evidence is found. (c) Each record is assigned a unique period of hospital care number that is shared by all connected records, represented by colour.

Table 2: Five types of evidence of a connection between episodes in the NSW Admitted Patient Data Collection

Predicate	Criterion
1 Overlapping episodes	Child episode started during parent episode
2 Transfer in	Child episode started within 9 hours of parent episode end AND source of referral in child episode indicated transfer AND child episode "facility transferred from" column matched parent episode facility
3 Transfer out	Child episode started within 9 hours of parent episode AND parent episode mode of separation was a transfer to another facility AND parent episode "facility transferred to" column matched child episode facility
4 Bidirectional transfer	Child episode started within 9 hours of parent episode AND source of referral in child episode indicated transfer AND parent episode mode of separation was a transfer to another facility
5 Type change	Child episode started within 30 minutes of the parent episode AND both episodes occurred in the same facility AND parent episode mode of separation was a type change separation AND child episode source of referral was a type change admission

If a candidate parent/child pair of episodes satisfies any of these criteria, then the two episodes are connected.

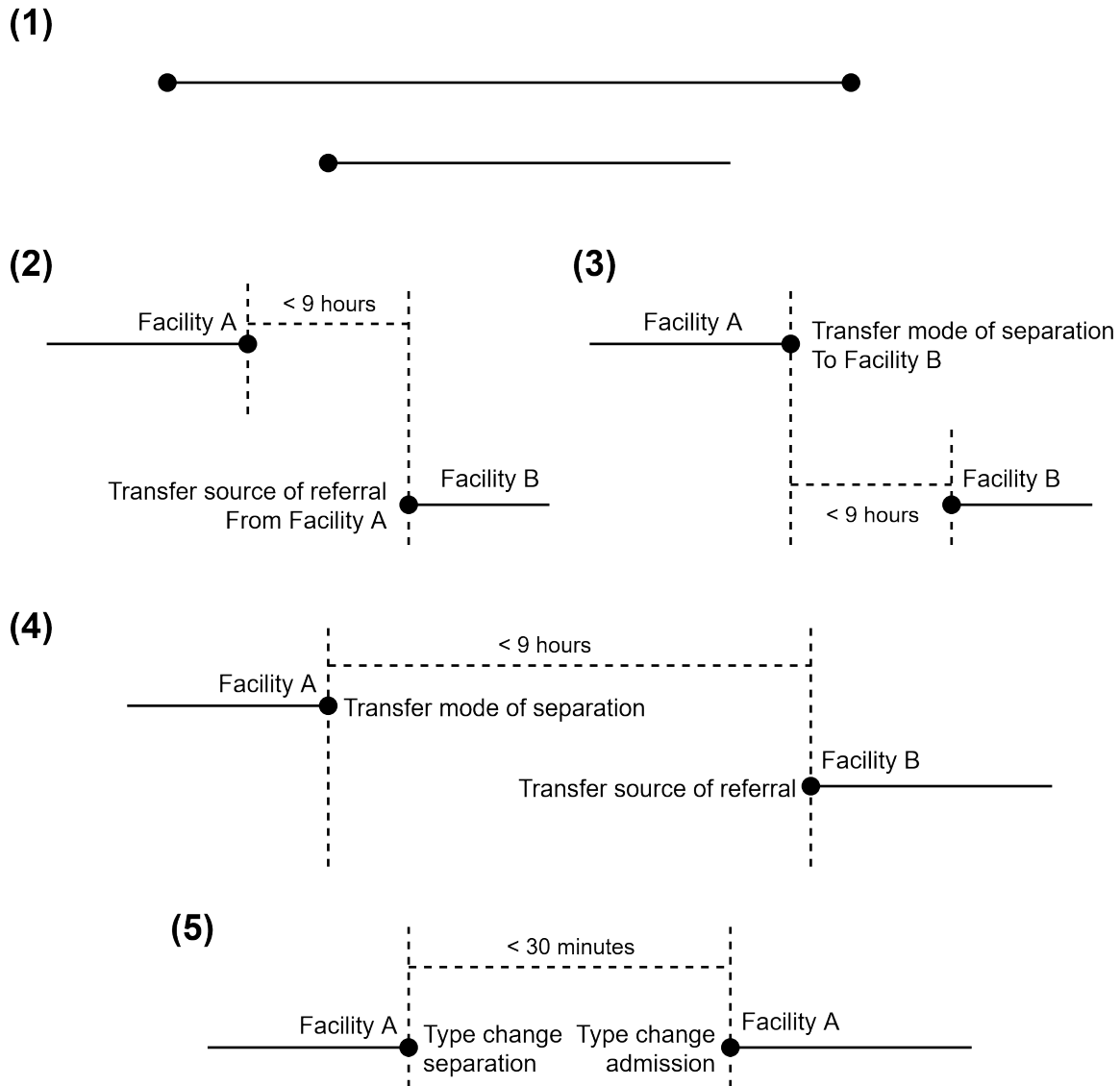
to take longer and be less common than transfers within metropolitan Sydney. To account for this, we also examined times of 48-hour transfers from regional and rural NSW to any Sydney hospital. This analysis was broken down by local health districts (LHD), which are administrative areas that partition the state for health service delivery [14]. For this stratification, LHDs were ranked by the proportion of their

population living in remote, outer regional or inner regional areas [15].

Assessing chosen predicates

To evaluate our chosen predicates in the NSW APDC, we examined trends in the type of connections discovered, and

Figure 2: Visual representation of predicates that determine if two episodes are part of the same period of hospital care in the NSW APDC



(1) Overlapping episodes. Note that the child episode, underneath, may end during or after the parent episode above it. (2) Transfer in. (3) Transfer out. (4) Bidirectional transfer. (5) Type change.

how the graph theory method applied to our dataset. For the transfer predicates, we counted episodes with a transfer mode of separation by the presence of a child episode connected by each type of predicate. Likewise, we counted episodes with a transfer source of referral by the presence of a parent episode connected by each predicate. For the type change predicate, we counted episodes with type changes recorded in the mode of separation (respectively source of referral) field by the presence of a connected child (respectively parent) episode by the “type change” predicate.

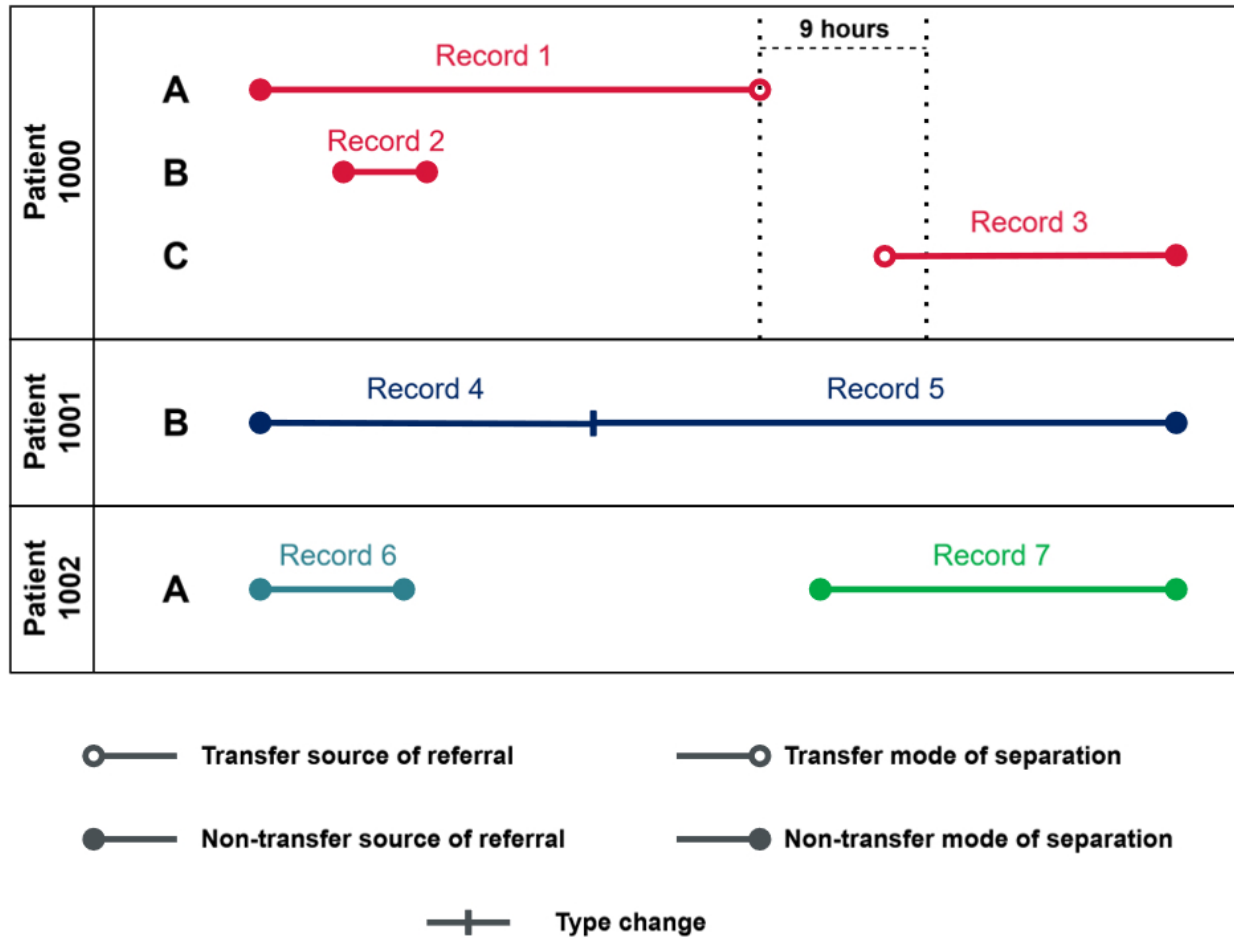
Comparing to a linear approach

To compare the results of the graph theory method, we applied a linear approach for constructing POHCs using a subset of the APDC records from June 2021 (295,044 episodes of care),

using the same predicates. We used a smaller subset of the APDC to allow for manual review of missed connections. In the linear approach, episodes were ordered by personal identifier, episode start date and time, and episode end date and time. An episode was considered part of the same POHC as the previous episode for the same patient if they satisfied any of the predicates in Table 2, with the earlier episode treated as the parent and the later episode treated as the child.

Using the example dataset in Table 1, record (2) overlaps with record (1), so is part of the same linear POHC. However, record (2) does not overlap with record (3), nor ends with a transfer, so record (3) marks the start of a new POHC. Records (4) and (5) together form a linear POHC, and records (6) and (7) each form their own linear POHC. We compared the POHCs constructed using the two methods to examine the performance of each method.

Figure 3: Periods of hospital care (POHC) arising from synthetic admitted patient data for three patients, as in Table 1



Letters represent the facility recorded for each episode. Each POHC is represented by a different colour.

Implementing this algorithm as an R package

This algorithm has been implemented in the R package *journeyer* (Supplementary Appendix 1). For the NSW APDC, connections were found by joining the APDC to itself by the patient identifier, excluding joins from each record to itself, and selecting only pairs where any of the chosen predicates hold true. This gave a list of edges in the form of a two-column dataset with record identifiers in each column. A graph was initialised with nodes identified by record identifiers, then connected components were found from the nodes and edge list and identified with an incremental journey number.

The implementation of this algorithm was largely compatible with *dbplyr* [16], which could divert the self-join and predicate evaluations to a database management system. One advantage of this was that it allowed all episodes to be compared without necessarily materialising each pair of records simultaneously in memory. This allowed connections to be found on datasets that were larger than memory allowed or stored in a remote database. However, the connected components algorithm required the computational graph to fit in memory. Other than this memory requirement, there were no strict computational requirements.

We included a README file in the supplementary source code for *journeyer* to describe the use of the package in more detail. In brief, an admitted patient dataset may be

joined to itself on patient identifier and filtered using the *find_links()* function to keep only pairs of episodes that satisfy the chosen predicates. Then POHC identifiers can be assigned to each episode using the *collect_journeys()* function, and the resulting dataset can be aggregated by the POHC identifier to create analysis variables for downstream epidemiological analysis.

Results

Characteristics of connections between episodes in the NSW APDC

The graph theory algorithm applied to the full NSW APDC dataset grouped 22,794,746 NSW APDC episodes of care into 21,405,451 periods of hospital care using 1,447,799 direct connections. The predicate involved in the most connections was overlapping episodes (Table 3). The most common types of connections were bidirectional transfers with matching facilities and no overlap in time (predicates 2–4; 481,108 pairs, 33.2%), overlapping episodes with no evidence of a transfer or type change (predicate 1 only; 341,197 pairs, 23.6%), and overlapping episodes with evidence of a type change (predicates 1 and 5; 316,330 pairs, 21.9%).

Table 3: Observed connections involving each type of predicate in the NSW APDC

	Predicate	Connections satisfying predicate	%
1	Overlapping episodes	665,011	45.9
2	Transfer in	607,360	42.0
3	Transfer out	627,812	43.4
4	Bidirectional transfer	544,032	37.6
5	Type change	352,133	24.3
	Total	1,447,799	100.0

These predicates are not mutually exclusive, so connections can satisfy more than one predicate and counts of connections for each predicate can sum to more than the total.

Assessment of predicates

In the complete NSW APDC dataset, episodes with a mode of separation indicating a transfer had a connected child episode in 66% of cases (Table 4). The most common transfer type was “Transfer to other Hospital”, with 873,251 episodes with this mode of separation; it has the highest rate of connection to a child episode at 81%. Transfers to nursing homes and other accommodation had considerably lower connection rates than other modes of separation. Generally, there were no predicates that independently accounted for all episodes found; connections found by one predicate were supplemented by connections found by other predicates.

Episodes with a source of referral indicating they arose from transfers had matching parent episodes in 72% of cases (Table 5). Transfers from a hospital in the same LHD were recorded as sources of referral more than transfers from other health service areas, but connection rates were similar between sources of referral. Similarly to the episodes with transfer modes of separation in Table 4, predicates appear to supplement each other by finding a variety of connections.

There were 407,510 episodes of care in the NSW APDC that ended with a type change separation. The type change predicate found child episodes for 352,039 (86.4%) of these parent episodes. Conversely, there were 360,740 episodes of care with a type change admission source of referral; the type change predicate found connected parent episodes of care for 352,049 (97.6%) of these child episodes.

Comparison to linear approach

In the June 2021 subset of the NSW APDC, the linear approach missed some of the connections found by the graph theory method, resulting in some linear POHCs not accurately representing complete patient journeys from admission to discharge. From 295,044 episodes, the linear method found 281,603 POHCs, whereas the graph theory method found 281,315 POHCs. The linear method missed connections in 174 graph theory POHCs (0.1% of POHCs), which resulted in 288 extra linear method POHCs. These missed connections affected 700 episodes of care (0.2% of episodes). Of the 288 extra linear POHCs, 228 (79%) arose because an index episode had two or more nested episodes; the linear method compared subsequent nested episodes to the previous nested episode

and missed the relationship with the index episode. A further 47 (16%) arose due to a similar situation where a transfer occurred from an index episode, but was not discovered by the linear method due to an interleaving nested episode. The remaining 13 (5%) extra linear POHCs arose due to various uncommon reasons, including record duplication, other potential data quality issues, and a new episode interrupting a transfer.

Time allowed between transfers

The cumulative proportion of 48-hour transfers in the full NSW APDC dataset stabilises at approximately 6 hours from the end of the parent episode, with approximately 98% of transfers completed (Figure 4). However, at 6 hours there are still transfers to be completed from some regional and rural LHDs to metropolitan Sydney LHDs (Figure 5). The cumulative proportions of all 48-hour transfer types appear stable at 9 hours.

Discussion

We present a graph theory based method to reconstruct patient journeys from complex admitted patient data. This method deduces connections between inpatient records and groups them into cohesive units. It outperforms a method using the same comparisons that only considers the temporally previous episode.

Compared to a linear method based off chronological ordering using the same predicates, our graph theory method compares pairs of episodes that the linear method does not consider. The extra comparisons allow the graph theory method to identify complex, nonlinear patient journeys. For example, this sort of journey can occur when a patient is admitted to one facility, has a procedure performed at second facility, and then is transferred to a third facility.

The graph theory method requires researchers to decide how episodes should be classed as connected. These rules or predicates can be chosen using features that would be expected in arbitrary inpatient administrative data, such as overlapping episodes, transfers, and type changes. Using these features, we developed rules for finding connections in the NSW APDC (Table 1).

Table 4: Candidate parent episodes with a transfer mode of separation, and matching child episode by various types of connection

Mode of separation in candidate parent episode	Child episodes found					Total candidate parent episodes
	Overlapping episodes	Transfer in	Transfer out	Bidirectional transfer	Any connection	
Transfer to Palliative Care Unit or Hospice	216 (6.0)	1,854 (51.5)	827 (23.0)	1,962 (54.5)	2,187 (60.7)	3,601 (100.0)
Transfer to Nursing Home	2,329 (1.5)	3,304 (2.1)	2,870 (1.8)	3,510 (2.2)	5,993 (3.8)	158,579 (100.0)
Transfer to Public Psychiatric Hospital	1,364 (12.3)	5,319 (47.8)	6,187 (55.6)	5,608 (50.4)	7,828 (70.4)	11,122 (100.0)
Transfer to other Hospital	72,112 (8.3)	507,751 (58.1)	616,153 (70.6)	526,825 (60.3)	707,396 (81.0)	873,251 (100.0)
Transfer to Other Accommodation	1,584 (2.8)	3,973 (6.9)	514 (0.9)	4,199 (7.3)	5,809 (10.1)	57,432 (100.0)
Any transfer	77,605 (7.0)	522,201 (47.3)	626,551 (56.8)	542,104 (49.1)	729,213 (66.1)	1,103,985 (100.0)

Pairs of episodes where the candidate parent episode had a transfer mode of separation were tested for connections with each of the predicates in our method. Type change connections were impossible from candidate parent episodes with a transfer mode of separation, so were not included in this table. Percentages are in parentheses.

Table 5: Candidate child episodes with a source of referral indicating transfer, and matching parent episodes by various types of connection

Source of referral in candidate child episode	Parent episodes found					Total candidate child episodes
	Overlapping episodes	Transfer in	Transfer out	Bidirectional transfer	Any connection	
Hospital in same Local Health District	75,018 (10.4)	428,893 (59.4)	358,291 (49.6)	382,136 (52.9)	508,259 (70.3)	722,504 (100.0)
Hospital in another Local Health District	55,562 (17.8)	177,114 (56.7)	143,049 (45.8)	159,444 (51.1)	237,671 (76.1)	312,152 (100.0)
Any transfer	130,580 (12.6)	606,007 (58.6)	501,340 (48.5)	541,580 (52.3)	745,930 (72.1)	1,034,656 (100.0)

Pairs of episodes where the candidate child episode had a transfer source of referral were tested with each of the predicates in our method. Type change connections are impossible to candidate child episodes with a transfer source of referral, so are not included in this table. Percentages are in parentheses.

Characteristics of periods of hospital care and assessment of predicates

Patient journeys in NSW admitted patient data are often simple, involving only a single record. This is demonstrated by the comparatively small number of connections (1,447,799) found in the NSW APDC compared to the number of episodes (22,794,746). However, complex inpatient journeys still occur, and these must be accounted for in health system planning and epidemiological studies to prevent overcounting.

The predicates we used to find connections in the APDC (Table 2) all work together. Alone, each predicate accounted for no more than half the connections found in the APDC (Table 3), suggesting that there is no single connection type that captures most hospitalisation patterns. Predicates to find inter-hospital transfers also improve capture in combination (Table 4, Table 5). Type changes, or statistical separations, have a different structure to inter-hospital

transfers in NSW APDC data and so benefit from a different predicate.

Transfer and type change predicates

To evaluate transfer predicates, we examined the proportion of episodes with a transfer mode of separation that had a connected child episode (Table 4). Among these episodes, 66% had a matching child episode. We also examined episodes that had a transfer source of referral, finding that 72% of these episodes had a connected parent episode (Table 5). However, we would not expect a child episode to be found in cases where the receiving facility is out-of-scope. In-scope facilities are public hospitals, public psychiatric hospitals, multi-purpose services, private hospitals and private day procedure centres in NSW [13], which excludes interstate hospitals.

When examining episodes with a type change admission source of referral, we found parent type change separations

Figure 4: Cumulative proportion of 48-hour transfers complete through time, in the NSW APDC

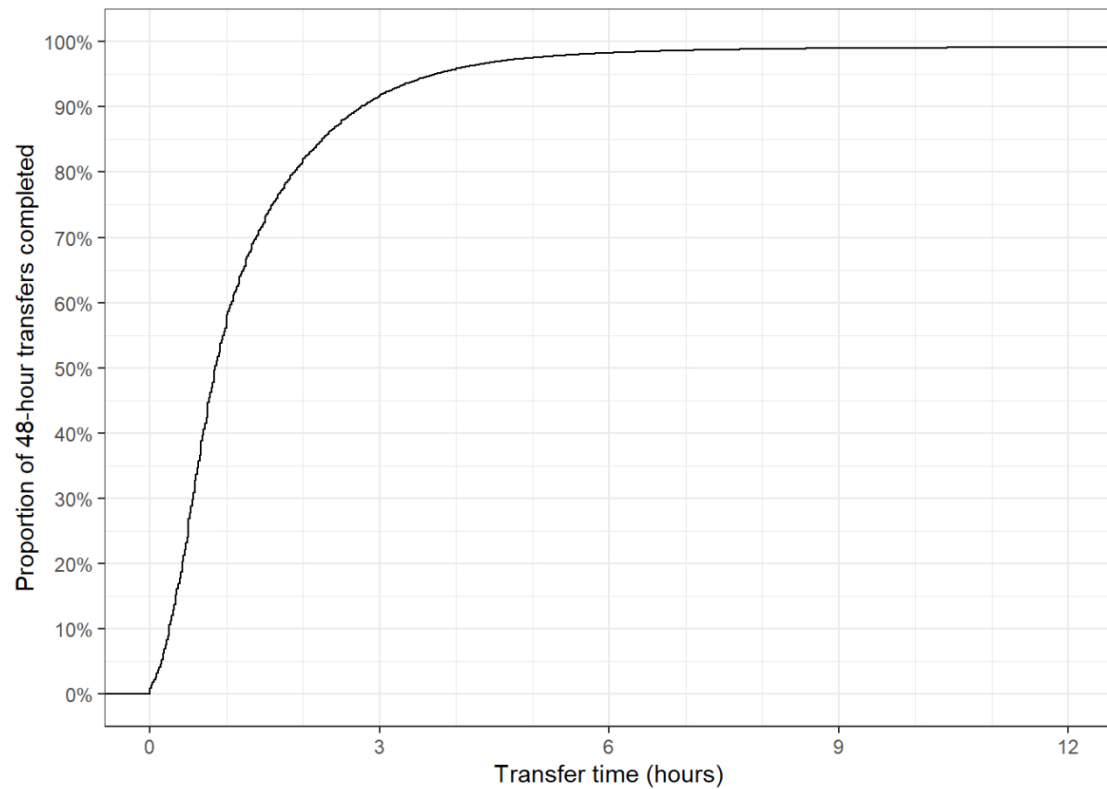
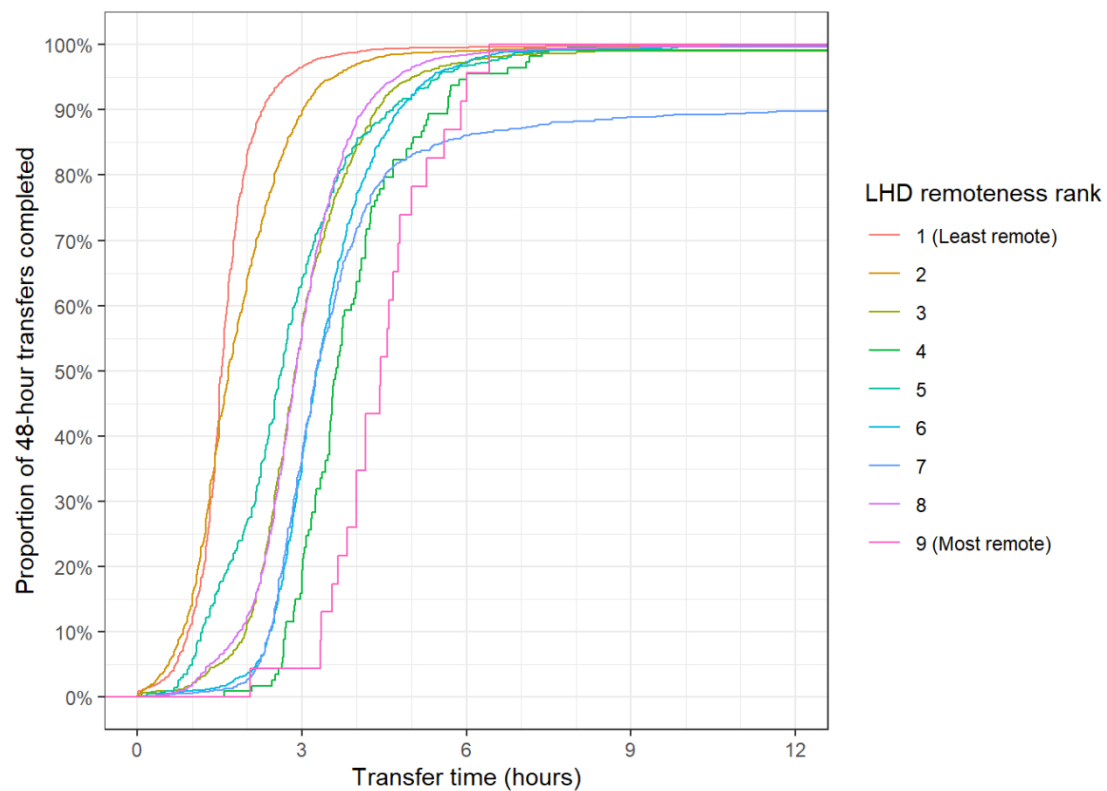


Figure 5: Cumulative proportion of 48-hour transfers from a regional or rural local health district (LHD) to any metropolitan Sydney LHD complete through time. Each curve represents transfers out of each regional or rural LHD, ranked by remoteness [14, 15]



in 98% of cases. This high rate suggests that the predicate is appropriate to capture type changes where a type change admission is recorded, and that the 30 minute window for type changes is appropriate. When considering type changes in the other direction, where a type change mode of separation was present, 86% of type change separations had a matching type change admission. The number of episodes with type change separations was higher than the number of episodes with type change admissions (407,510 versus 360,740 respectively). The basis for our predicate is that every type change admission should have a matching type change separation, so we would expect these numbers to be approximately equal, allowing for mismatching episodes at the temporal boundary of the data period. However, this difference suggests that some episodes are coded to a type change separation without a matching type change admission.

Improvement on a linear approach

The graph theory method performed more comparisons than the linear method and was more capable of tracking complex patient journeys. When an episode had multiple nested episodes, and the nested episodes did not overlap with each other, the graph theory method was able to group them all together, while the linear method started a new POHC for each subsequent nested episode. The graph theory method also correctly connected transfers from an index episode following a nested episode, while the linear method would miss the connection from the index episode to the transfer. Hence, the graph theory method is more suitable for tracking patient journeys with complex hospitalisation patterns.

Time allowed for transfers

The time allowed for transfers is a key parameter for our predicates. Sufficient time must be allowed for a transfer to take place, but allowing too much time for transfers increases the risk of finding spurious connections. The cumulative proportion of transfers state-wide suggests that at least 6 hours is an appropriate time to choose (Figure 4). However, an allowance of 9 hours is more appropriate to capture transfers from regional and rural NSW to Sydney (Figure 5).

Generalisability

Other admitted patient datasets

Administrative inpatient data varies between jurisdictions, and our graph theory method can be adapted to a wide range of settings. Using this method, researchers are able to make their own decisions about what connection predicates between episodes are suitable for tracking patient journeys in their research. A one-size-fits-all solution is unlikely to be feasible [17], but the flexibility of this method allows it to be adapted to many researchers' needs. To use this method, researchers can develop tailored predicates for their dataset and research questions. Each dataset will have its own criteria for evidence of a connection between episodes and researchers may decide to account for other types of connections not mentioned below. However, fundamentally, connections are only possible between two episodes relating to the same patient.

There are several prerequisites for constructing POHCs from administrative inpatient data using our method. Each record must have a unique record identifier; a patient-specific personal identifier; dates and times for the start and end of each episode; identifiers for facilities or hospitals; a coded description of the provenance of the episode, e.g. from the hospital's emergency department or a transfer from another hospital; and a coded description of the outcome of the episode, such as discharged home or transfer to another hospital. If available, information on origin and destination facility for inter-hospital transfers improves the accuracy of the algorithm.

The predicates we chose reflect real-world processes captured in the NSW APDC; other admitted patient datasets may have analogous processes that allow these predicates to be tweaked and reused. Types of connections that could be considered in other admitted patient datasets include overlaps, inter-hospital transfers, and type changes or statistical discharges.

There are three main decision points available to researchers to develop flexible predicates for inter-hospital transfers. First, researchers may decide which codes for an episode's provenance indicate that it arose from a transfer, and similarly which codes for an episode's outcome indicate that it ended with a transfer. Second, researchers can decide which of these pieces of evidence is required, including if both are required. Thirdly, researchers can choose the maximum amount of time allowed for a transfer, considering the geography of their health system, and any clinical considerations that would affect the time taken to travel between hospitals. If origin and destination facilities for hospital transfers are available in the dataset, researchers may further use this information to improve their predicates.

For type changes, statistical discharges, or other intra-hospital transfers that result in a new record in the dataset for the same hospital, researchers have similar flexibility in determining their predicates. Similarly to inter-hospital transfers, researchers can consider which evidence can be considered from the provenance and outcome of a pair of episodes, and what evidence is sufficient. They may also decide how much time should be allowed for type changes to occur. Another decision point is whether the facilities of a pair of episodes in a candidate type change connection must be the same.

Other patient-related datasets and research questions

An advantage of the graph theory method is its adaptability to various datasets and research questions. When examining emergency department presentations and resulting admissions, nodes representing health service interactions (presentations or inpatient admissions) can be included. The researcher only needs to decide how connections between each type of interaction (emergency–emergency, emergency–inpatient, inpatient–inpatient) should be defined, and then periods of emergency or admitted care can be defined as the connected components of this new graph. For studies examining cancer treatment journeys or readmissions, the allowed time between subsequent episodes could be expanded to a scale of days or weeks as needed. Other patient-related data, such as public transport trips or primary data such as surveys, could also be

incorporated using this method to provide more context for each patient journey.

Limitations

There is no gold standard to which we can compare the results of our algorithm against, which Gubbels et al. [9] noted as a similar limitation to the DNPR patient journey method. While specific aspects of the method and results can be examined, assessment and evaluation are limited to heuristics built from patterns observed in the data, and a true validation is not feasible. Although we found the algorithm performs well for NSW APDC, its use for other types of admitted patient data will need to be assessed.

The algorithm we present assumes that patient identifiers are accurate and consistent. In many administrative datasets this is not guaranteed, particularly when probabilistic record linkage is required. In the case of a missed match, where two records relating to the same patient are given different patient identifiers, then these two records will never be part of the same patient journey. Conversely, in the case of a false match, where two records relating to different patients are given the same patient identifier, this algorithm could combine too many records into one patient journey if the records satisfied the required predicates.

In this method, predicates cannot use data from any records except for the pair of records they are comparing. This limits the type and nature of predicates that could be used in other datasets. For example, a predicate that requires the absence of any records occurring between the pair being compared is incompatible with this method.

Inpatient activity is only a part of the patient journey, and administrative data is unlikely to capture the complete patient experience. This is why a variety of patient journey mapping approaches have been used, including interviews of patients and case note audits [17]. Nevertheless, our method is a useful tool to analyse administrative data to understand broad patterns for epidemiological and health system planning purposes, and with the generalisation possibilities introduced above, it may be extended to capture other aspects of patient journeys.

Finally, responsibility for cleaning the input data for this algorithm remains with the researcher, in a departure from the DNPR method [9]. A researcher should always consider their research question, and how policies and procedures have affected the data collection they use.

Conclusion

We present a method to track patients on their journey through admitted patient data, from formal admission to formal discharge. This method is flexible, allowing it to be used on a range of datasets; and generalisable, allowing modifications to capture other aspects of the patient journey. This method can assist in answering a wide range of research questions in epidemiology. To support ease of use of this method, it has been implemented in the R package *journeyer* (Supplementary Appendix 1).

Acknowledgements

Record linkage was carried out by the Centre for Health Record Linkage (www.CHeReL.org.au).

This work was conducted while Ian Powell and Zhisheng Sa were employed as trainees on the NSW Biostatistics Training Program funded by the NSW Ministry of Health.

Statement on conflicts of interest

No authors declare any conflicts of interest.

Ethics statement

This article shares methodology developed during the course of government work with the Admitted Patient, Emergency Department and Deaths Register (APEDDR) established under the *Public Health Act 2010* (NSW). Ethical approval has not been sought, although the results presented herein are shared with the approval of the APEDDR data custodian.

Data availability statement

The *Public Health Act 2010* (NSW) does not permit sharing APEDDR as used in this research. However, researchers interested in using linked NSW data, including parts of the NSW APDC, may apply for access to linked data through the CHeReL website (<https://www.cherel.org.au/>).

Supplementary Appendices

1. *journeyer* version 0.1.0 R package source

References

1. Ha NT, Harris M, Preen D, Moorin R. Time protective effect of contact with a general practitioner and its association with diabetes-related hospitalisations: a cohort study using the 45 and Up Study data in Australia. *BMJ Open* 2020 Apr.;10:e032790. <https://doi.org/10.1136/bmjopen-2019-032790>
2. Dinh MM, Bein KJ, Delaney J, Berendsen Russell S, Royle T. Incidence and outcomes of aortic dissection for emergency departments in New South Wales, Australia 2017–2018: A data linkage study. *Emerg. Med. Australas.* 2020;32:599–603. <https://doi.org/10.1111/1742-6723.13472>
3. Bell J, Lingam R, Wakefield CE, Fardell JE, Zeltzer J, Hu N, et al. Prevalence, hospital admissions and costs of child chronic conditions: A population-based study. *J. Paediatr. Child Health* 2020;56:1365–70. <https://doi.org/10.1111/jpc.14932>
4. Leong RNF, Wood JG, Liu B, McIntyre PB, Newall AT. High healthcare resource utilisation due to pertussis in Australian adults aged 65 years and

- over. *Vaccine* 2020 Apr.;38:3553–9. <https://doi.org/10.1016/j.vaccine.2020.03.021>
5. Worthington JM, Gattellari M, Goumas C, Jalaludin B. Differentiating Incident from Recurrent Stroke Using Administrative Data: The Impact of Varying Lengths of Look-Back Periods on the Risk of Misclassification. *Neuroepidemiology* 2017 Jun.;48:111–8. <https://doi.org/10.1159/000478016>
 6. Gattellari M, Goumas ,Chris, Jalaludin ,Bin, and Worthington JM. Population-based stroke surveillance using big data: state-wide epidemiological trends in admissions and mortality in New South Wales, Australia. *Neurol. Res.* 2020 Jul.;42:587–96. <https://doi.org/10.1080/01616412.2020.1766860>
 7. Nolan EK, Chen H-Y. A comparison of the Cox model to the Fine-Gray model for survival analyses of re-fracture rates. *Arch. Osteoporos.* 2020 Dec.;15:1–8. <https://doi.org/10.1007/s11657-020-00748-x>
 8. Mitchell R, Draper B, Brodaty H, Close J, Ting HP, Lystad R, et al. An 11-year review of hip fracture hospitalisations, health outcomes, and predictors of access to in-hospital rehabilitation for adults ≥ 65 years living with and without dementia: a population-based cohort study. *Osteoporos. Int.* 2020 Mar.;31:465–74. <https://doi.org/10.1007/s00198-019-05260-8>
 9. Gubbels S, Nielsen KS, Sandegaard J, Mølbak K, Nielsen J. The development and use of a new methodology to reconstruct courses of admission and ambulatory care based on the Danish National Patient Registry. *Int. J. Med. Inf.* 2016 Nov.;95:49–59. <https://doi.org/10.1016/j.ijmedinf.2016.08.003>
 10. Elliss-Brookes L, McPhail S, Ives A, Greenslade M, Shelton J, Hiom S, et al. Routes to diagnosis for cancer – determining the patient journey using multiple routine data sets. *Br. J. Cancer* 2012 Oct.;107:1220–6. <https://doi.org/10.1038/bjc.2012.408>
 11. Zand MS, Trayhan M, Farooq SA, Fucile C, Ghoshal G, White RJ, et al. Properties of healthcare teaming networks as a function of network construction algorithms. *PLOS ONE* 2017 Apr.;12:e0175876. <https://doi.org/10.1371/journal.pone.0175876>
 12. Brunson JC, Laubenbacher RC. Applications of network analysis to routinely collected health care data: a systematic review. *J. Am. Med. Inform. Assoc.* 2018 Feb.;25:210–21. <https://doi.org/10.1093/jamia/ocx052>
 13. Centre for Health Record Linkage, NSW Government. Home – CHeReL [Internet]. 2023 [cited 2023 24]; Available from: <https://www.cherel.org.au/datasets>
 14. NSW Ministry of Health. Local health districts and specialty networks [Internet]. 2023 [cited 2023 29]; Available from: <https://www.health.nsw.gov.au/443/lhd/Pages/default.aspx>.
 15. Bureau of Health Information. The Insights Series: Healthcare in rural, regional and remote NSW [Internet]. Sydney (NSW): BHI; 2016. Available from: https://www.bhi.nsw.gov.au/BHI_reports/Insights_Series/healthcare_in_rural,_regional_and_remote_nsw.
 16. Wickham H, Girlich M, Ruiz E, RStudio. dbplyr: A “dplyr” Back End for Databases [Internet]. 2022 Jun.; Available from: <https://cran.r-project.org/package=dbplyr>.
 17. Davies EL, Bulto LN, Walsh A, Pollock D, Langton VM, Laing RE, et al. Reporting and conducting patient journey mapping research in healthcare: A scoping review. *J. Adv. Nurs.* 2023;79:83–100. <https://doi.org/10.1111/jan.15479>

Abbreviations

NSW:	New South Wales
LHD:	local health district
POHC:	period of hospital care
APEDDR:	Admitted Patient, Emergency Department and Deaths Register
DNPR:	Danish National Patient Register
APDC:	Admitted Patient Data Collection

