

Using the Problem Diagnosis Reference Set

in a secondary data use scenario

v1.0

Citation

Truran D, Lawley M, Zhang M, Kemp M, Hansen D. 2015. Using the Problem Diagnosis Reference Set in a Secondary Data Use Scenario. CSIRO Report EP162403, Australia.

Acknowledgements

Council of Australian Governments

The National E-Health Transition Authority is jointly funded by the Australian Government and all State and Territory Governments.

IHTSDO (SNOMED CT)

This material includes SNOMED Clinical Terms™ (SNOMED CT®) which is used by permission of the International Health Terminology Standards Development Organisation (IHTSDO). All rights reserved. SNOMED CT® was originally created by The College of American Pathologists. “SNOMED” and “SNOMED CT” are registered trademarks of the IHTSDO, (<http://www.ihtsdo.org/>).

Disclaimer

The National E-Health Transition Authority Ltd (NEHTA) makes the information and other material (‘Information’) in this document available in good faith but without any representation or warranty as to its accuracy or completeness. NEHTA cannot accept any responsibility for the consequences of any use of the Information. As the Information is of a general nature only, it is up to any person using or relying on the Information to ensure that it is accurate, complete and suitable for the circumstances of its use.

Document control

This document is maintained in electronic form and is uncontrolled in printed form. It is the responsibility of the user to verify that this copy is the latest revision.

Copyright © 2015 National E-Health Transition Authority Ltd

This document contains information which is protected by copyright. All Rights Reserved. No part of this work may be reproduced or used in any form or by any means – graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems – without the permission of NEHTA. All copies of this document must include the copyright and other information contained on this page.

Contents

Executive summary.....	iii
Recommendations.....	iv

Section A vii

Summary.....	viii
--------------	------

Section B xxi

1	Introduction	1
2	Background	2
3	Use cases	4
3.1	Dynamic use cases	4
3.2	Static use cases	5
4	Problem Diagnosis Reference Set	7
4.1	Scope.....	7
4.2	Clinical Finding Grouper Exclusion RefSet	8
4.3	Findings versus disorders.....	9
5	Example use case	11
5.1	Characteristics of APNMDS reports	12
6	Major challenge and opportunity	15
6.1	Over-counting, under-counting	15
6.2	Reporting by specification	15
6.3	Report generation using maps between SNOMED CT and ICD	16
6.4	Proposal for hybrid approach	16
7	Aggregation technique.....	18
7.1	Evaluating an Aggregation RefSet.....	18
7.2	Methods for content management	19
7.3	Methods for technical management	21
8	Results	24
8.1	Other specified.....	24
8.2	Problem-Diagnosis RefSet members and APNMDS report categories	25
8.3	Other	26
8.4	Overall.....	28
8.5	Further detailed examples.....	28
8.6	Efficiency and reproducibility gains	32
8.7	Implications.....	32
9	Conclusion	34

Figures

Figure 1: Operating environment, clinical documentation to statistical reports.....	xii
Figure 2: Problem Diagnosis RefSet relationships and pathways, heading towards the aggregated report categories	xiii
Figure 3: Choose the wrong aggregation pathway and lose patient data from secondary reports	xiv
Figure 4: Take a different pathway and get a different outcome.....	xv
Figure 5: Specifying the correct pathway, correct aggregation of patient case counts to achieve the correct Reporting category outputs	xvi
Figure 6: Specifying the preferred aggregation pathway	xvii
Figure 7: Pathways that don't reach the required Reporting category	xix
Figure 8: Residuals and exclusions in ICD.....	2
Figure 9: Chart of Principal diagnoses for admitted patient episodes by ICD chapter categories 2008-2013 (data summarised from AIHW)	13
Figure 10: Aggregation primary path, after pruning; mutually exclusive and sensitive tree structure	16
Figure 11: Generic workflow and methodology	22
Figure 12: Summary of aggregations achieved	24
Figure 133: Other specified concepts required for APNMDS reporting (aggregation for complete sub-hierarchy shown)	25
Figure 14: Linearisation of "Other specified" concepts required for APNMDS reporting	25
Figure 15: Linearisation of Problem Diagnosis RefSet aggregation for APNMDS reports.....	26
Figure 16: Linearisation of "Other" non APNMDS concepts	27
Figure 17: Overall, Problem Diagnosis RefSet tops to APNMDS, highest level aggregation shown	28
Figure 18: Example of aggregation path for a single concept	28
Figure 19: Portion of Problem Diagnosis RefSet aggregation showing Disorder and Finding partitions; without modelling flaws	30

Tables

Table 1: Hierarchical distribution of Problem Diagnosis Reference Set concept members.....	7
Table 2: Principal Diagnoses attributed to admitted patient episodes by ICD chapter categories 2008-2013 (data summarised from AIHW)	12
Table 3: Evaluation criteria for Problem Diagnosis aggregation RefSet	18
Table 4: Top level alignment between SNOMED CT, ICD-10-AM chapters and APNMDS reporting categories	20
Table 5: Example (extract) of aggregation relationship table	30
Table 6: Aggregation technique and outcomes assessed against evaluation criteria	32

Executive summary

The work program was designed to investigate obstacles and options related to the use of SNOMED CT in a secondary data use scenario. SNOMED CT is designed to be used by clinicians to document patient medical records. It is not entirely suited, in its original form, for secondary data purposes, especially statistical use cases.

Our work here focuses on techniques that can transform the SNOMED CT Problem Diagnosis Reference Set into an annotated Reference Set (RefSet), so that SNOMED CT content can be re-purposed for statistical uses, without relying on the use of maps, which would impose additional burdens and costs on the user community for map production, maintenance, synchronisation between versions and deployment.

The secondary data scenario we have chosen for demonstration purposes is the Admitted Patient National Minimum DataSet (APNMDS), and in particular the annual statistical reports for Principal diagnoses of hospitalised patients. All States and Territories make contributions to this collection, and AIHW produce annual reports that utilise the information. This use case is broadly familiar and well understood by the health information practitioner community.

The use case always sets the parameters for aggregation and downstream use. It specifies the desired outputs, and the challenge is to adapt the inputs (Problem Diagnosis RefSet content) to suit those requirements.

There are many other secondary use cases. This one was chosen for convenience, because it has stable and published requirements. The methods under development here are likely to be useful for other secondary reporting use cases as well, but would need to be tailored for other RefSet inputs, and different reporting outputs.

The techniques used here work in a technical sense, and are reproducible. However, this is proof-of-concept work only. No real world data is available to test the precision or goodness-of-fit of the aggregation technique against SNOMED-aware reporting requirements.

Comparative and repeated measures are needed to ensure that these techniques are suitable and/or can be refined, thus providing objective evidence that stable and reliable outputs for secondary data users can be produced.

Recent and ongoing development in SNOMED CT query language and data analytic approaches have informed some aspects of these techniques, and as the work being undertaken by the IHTSDO progresses further options are sure to emerge.

Section A of this report is abstracted from the main document and provides a summary of our investigations and findings, reported in lay-person terms, avoiding the technical jargon and complexity often associated with SNOMED CT. This section is best suited for consultation and liaison with stakeholders who are not SNOMED CT experts.

Section B provides a more technical description of the proof-of-concept techniques and results, and is mostly suited to SNOMED CT technologists and analysts.

Recommendations

Overall, potential next steps

The work here represents proof of concept only, but offers substantial promise for re-purposing SNOMED CT encoded patient data for secondary data uses.

The next step would be to refine and test the aggregation techniques provided here on real patient data collections. However, there are currently too few implementation of the Problem Diagnosis RefSet, especially within the operational context of the APNMDS use case to allow this to happen immediately.

In the meantime, there are some Problem Diagnosis RefSet enhancements that could be undertaken, to increase its usability and appeal for implementers, and some further issues that could be clarified or be resolved through consultation with stakeholders.

For NEHTA's consideration

- 1 Immediate and modest work to augment Problem Diagnosis RefSet membership to serve the APNMDS use case only could be relatively easily and quite quickly accomplished. This would entail adding the concepts identified from the file Other Specified APNMDS concepts.

These are the procedural concepts aligned with Chapter 21 of ICD, and are routinely reported in APNMDS reports. Patient cases allocated to this reporting category routinely account for approximately 25% of the admitted patient population. Without the ability to include these concepts in Problem Diagnosis RefSet, only ~75% of the APNMDS use case can be served by SNOMED CT encoded data collections.

Some caution is needed however. The Problem Diagnosis RefSet is used to bind to information model specifications. These specifications serve a broad range of users and implementations, the majority work in settings other than hospitals. Adding in procedural concepts that are required for the Admitted Patient setting may cause difficulties for other users.

- 2 More extensive work could be considered to expand the Problem Diagnosis RefSet content to suit other secondary data uses, and to consolidate NEHTA's RefSet management approach.

It is clear that RefSets for particular health domains (such as Emergency Departments, sub-acute care and General Practice) will use a considerable portion of the SNOMED CT terminology. Given that these sectors share terminology requirements, it makes little sense to produce distinct RefSets that duplicate or partially replicate the same content. Separate RefSets for each clinical domain only serve to entrench the health data silos that currently exist; they act as 'reasons not to share' and are therefore barriers to interoperability.

We see this is the Emergency Department (ED) RefSet and its intersection with the Problem Diagnosis RefSet. There are only 188 concepts that are unique, occurring only in the ED RefSet. Of these 188, we believe that 102 would be usefully included in the Problem Diagnosis RefSet. Again, these are all procedure-like terms. The remaining 86 concepts could be reviewed and included in Problem Diagnosis RefSet, if judged to be useful.

However, there is a difference between concepts and codes that are useful within a clinical setting or practice, and those concepts or codes that can be used to exchange clinical data and documents between care sectors and practitioners. The former may have a more constrained scope, while the latter needs to be as broad as possible. Techniques (such as RefSet annotation, indexing by frequency of use per user and 'boosting') might help manage these scope differences.

For secondary data user consideration, current APNMDS requirements

It is apparent from this proof of concept study, examining the goodness of fit between terminology inputs and patient data outputs, that there is an established data flow, from end-to-end.

Essentially, there are interaction effects between both the Problem Diagnosis RefSet content and the APNMDS use case.

There are no main effects detected.

This means that enhancements and adjustments will need to be made to both ends of the data pipeline, to provide an accurate use and repurposing of patient data.

Recommendations 1 and 2 above outline some appropriate short and medium terms tasks to enhance the Problem Diagnosis RefSet content to better serve APNMDS purposes.

The following recommendations are offered to help guide NEHTA's consultation with NHSICC, DoH, AIHW and other state and territory data custodians, and to alert them to the options and considerations for revision of APNMDS specifications. This will enable APNMDS collections to prepare for the inclusion of SNOMED CT encoded patient data over the coming years.

3 Consider further or different aggregation requirements in light of the use of SNOMED CT at the point of care for clinical purposes

Immediate questions and considerations that might provide NEHTA with stakeholder requirements to act upon would be:

- Does the proof of concept approach demonstrated here approximate their expectations for high level, national, static reports of inpatient episodes?
- Would they support the expansion of the Problem Diagnosis RefSet to include APNMDS reporting categories?
- Do they find the existing reporting categories acceptable, and would they make use of this aggregation approach, with AU namespace aggregation concepts and aggregation paths in future reports?
- Are further or different inclusions or exclusions needed? Different category memberships?

4 Longer term, and more importantly, secondary data custodians should take into account the overall evolving operating environment, and consider the end-to-end use of patient data, encoded in various terminologies and vocabularies and derived from different clinical information systems, and how this might impact their own portfolio responsibilities.

- Can existing secondary data specifications, such as APNMDS, be more comprehensively specified to take account of the growing number of e health initiatives?
- Do other existing secondary use cases and their data protocols adequately account for clinical uses?
- Would secondary data use cases benefit from a broader scope of health information capture, accounting for patient information collected outside the admitted patient care sector?
- What legitimate secondary (aggregated) data use might eventually be made of patient data collections that reside within the PCEHR?
- Could such data capture be facilitated by the use of Problem Diagnosis RefSet and aggregation techniques?
- What additional knowledge, and therefore health system administration and policy benefits might be achieved by more comprehensive and descriptive data collections, data mining and data analytics (in the 'big data' sense)?

5 There are also some fundamental issues that need research, development and consideration at the national and international level.

These include:

- OWL formats of SNOMED CT releases, and how these might impact implementers and users, particularly secondary data users
- SNOMED CT query language developments and whether these will be capable of lending themselves to these sorts of aggregation requirements
- Whether mapping is worth the continued level of investment and whether there are more efficient and effective approaches to what might be regarded as ‘harmonisation’ when what seems to be most in demand is re-purposing.

The constitute dependencies that need to be considered by stakeholders. IHTSDO is currently developing guidance and methods to assist with data analytics, and these may influence secondary data users when they attempt to re-purpose SNOMED CT-encoded data.

Section A

CAVEATS

1. This report outlines **proof of concept** work, investigating the potential to re-purpose patient data encoded in SNOMED CT for secondary, aggregated and statistical reporting protocols.
2. Ongoing technical developments, guidance and advice by the IHTSDO and WHO will eventually have some influence on the techniques we use here. This international work provides **external dependencies** we need to be aware of, and it would be premature to recommend this proof of concept work as the ultimate technique of choice.
3. Attention is focused only on the use of the Problem-Diagnosis Reference Set (RefSet) released by NEHTA and its relationship to a single, prominent secondary data report produced by the AIHW, using patient data encoded in ICD-10-AM, representing Principal Diagnoses, and collected under the Admitted Patient National Minimum DataSet specifications (APNMDS).
4. These have been selected for our research purposes merely because they are **convenient**.
 - a. The Problem Diagnosis RefSet is convenient because it has broad coverage of patient health conditions, whether patients are hospitalised or not. This means that Problem Diagnosis RefSet content is suited for use in recording Principal Diagnosis information in a hospital context.
 - b. The APNMDS, Principal Diagnosis report is convenient because it is one of the few secondary and statistical use cases that has stable, publicly available, well-known and documented data specifications. These specifications are essential; we must know in advance what the data aggregation and reporting requirements are, if we are to ensure that SNOMED CT encoded patient data can meet those requirements. This report is a very good example of such specifications.

The APNMDS-Principal Diagnosis report has been established for many years, with annual reports routinely published, and it is familiar to most health information practitioners and hospital health system managers. The APNMDS is mandated for use by all state and territory health authorities and is governed by NHISSC.
5. There is no suggestion, real or implied, that this investigation recommends that the APNMDS data collection, encoding or reporting should change. Abstraction and clinical coding of medical records (in ICD-10-AM and related data elements) will persist, along with the use of ACHI and AR-DRG protocols. This is a **non-disruptive, technical approach** that we hope will eventually assist health data managers to use and re-purpose patient data encoded in SNOMED CT, in environments where SNOMED CT products are implemented in clinical systems, for use by clinicians who will document patient medical records electronically.

(Clinical Systems include those like Cerner, Cerner FirstNet, Epic, and other new generation Clinical Information Systems (CIS), as distinct from Patient Administration Systems (PAS) like HBCIS).

It may be the case in the future that patient data captured by hospital and non-hospital CIS, encoded in SNOMED CT, can be re-purposed for different kinds of secondary reports, perhaps relevant to high level overviews of Outpatients or Community Health practices.

SNOMED CT and ICD-10-AM.

SNOMED CT is a **clinical terminology**, used by clinicians, to document patient conditions at the point of care. It is the terminology that captures data at the **primary** source, the **original** medical record.

ICD-10-AM is a **statistical classification**, used by coders, data analysts and statisticians. It is applied **after** the medical record has been documented. It is the terminology used to abstract and capture and sometimes aggregate data; it is regarded as **secondary** data source (downstream from the point of care, after the patient has been discharged).

SNOMED CT and ICD-10-AM have each **been purposely designed** so that they each meet those distinct uses. Each of them is fit-for-purpose, but **the purposes are different**.

Different, not the same (and not better, and not worse)

Many people believe that SNOMED CT and ICD-10-AM are very alike, and perhaps can be used interchangeably, because they share a lot of words.

We know that SNOMED CT contains the concept	195967001	Asthma (disorder)
We know that ICD-10-AM contains the code	J45.x	Asthma

Unfortunately, focussing on the words that are common to both, disguises the structural differences.

The relationships and linkages between the words in each terminology are the important features that determine what the words mean; they provide context and definition.

Without these the words are just a laundry list.

SNOMED CT relationships are built on description logic

- it is **ontological**
- it is **poly**-hierarchical
- each concept can have **multiple high level classes**

ICD-10-AM linkages are built on statistical principles of mutual exclusivity and sensitivity

- it is **nosological**
- it is **mono**-hierarchical
- each concept can have **one high level class and one only**

Why worry if they are different?

The health information industry wants to ensure that health data collections are integrated, coherent and patient-centred.

We want to be able to **collect once and use many times** as outlined by Cimino¹ as this provides a greater assurance of accuracy in the data merely because it decreases the chance of transcription errors promises efficiency gains.

Earlier, van der Lei warned against the re-use of clinical data and proposed what he called the first law of informatics: "Data shall be used only for the purpose for which they were collected".²

If we take this advice, then SNOMED CT encoded data should be quarantined for clinical point of care purposes, for use in constructing an original medical record, and for this purpose only.

¹ Cimino, James J.. "Collect Once, Use Many: Enabling the Reuse of Clinical Data through Controlled Terminologies." *Journal of AHIMA* 78, no.2 (February 2007): 24-29.

² van der Lei J. Use and abuse of computer-stored medical records. *Methods Inf Med* 1991;30:79–80

This might mean that we would be faced with silo-isation of health data, with separate and disconnected health data collections, each suited distinctly to clinical, clinical research or statistical and reporting purposes.

Instead, health informaticians are pursuing Cimino's recommendations, acknowledging the known problems and challenges of re-using health data, and working towards developing approaches that **re-purpose**, rather than merely re-use health data.

Re-purposing attempts so far

Mapping between SNOMED CT and ICD-10 has received a lot of attention and effort internationally. There have been some successful initiatives that have produced and maintained maps between the two terminologies, though implementations of maps remain sparse and with little quantification of map performance.

However, these maps use the standard international edition of ICD-10, and not ICD-10-AM, and the maps may not adequately account for Australian coding and statistical standards. Maps produced in Australia, for the Australian operating context and users, have not been adequately tested or maintained. For example, the Emergency Department Reference Set (EDRS) to ICD-10-AM maps were produced in 2009, using the 2009 release of EDRS and the 6th edition of ICD-10-AM. These are now 6 years out-of-date.

The key characteristic of these maps is that they are separate and distinct technical products that sit between SNOMED CT and ICD-10. This means that in order to re-purpose SNOMED CT encoded data into ICD-10 coded data, you need something other than, and additional to, SNOMED CT and ICD-10.

This caused both IHTSDO (SNOMED CT owners) and WHO (ICD-10 owners) to look more closely at the possibility of achieving a closer relationship between the two terminologies, rather than building and maintaining a separate map product.

Work on **harmonising SNOMED CT and ICD** was commenced when ICD-11 developments began. WHO was considering a major change in ICD (from 10 to 11), so it presented a good opportunity to consider changes to the very foundation of ICD at the same time.

The proposal was that ICD-11 and SNOMED CT would be aligned by sharing a Common Ontology Foundation layer very similar to SNOMED CT ontological structures, and that ICD-11 would then 'linearise' the content from that foundation to allow a more tree-like mono-hierarchical structure to be derived.

The theory was that SNOMED CT provided a clinically valid ontological base, and that ICD-11 provided the statistical and reporting layer over the top, so that existing ICD users could sustain their Business As Usual work program and maintain backward compatibility with their historical data collections.

To some extent, this also meant that something new and different had to be constructed; a new foundation. It also suggested that there would be a significant change in ICD-11, making it distinctly different to earlier ICD products. This caused some concern about the potential changeover and adoption of ICD-11 (given the USA experience in changeover from ICD-9-CM to ICD-10-CM, still not complete in 2014). It also caused some people to speculate that such a close alignment could be considered scope creep, or would mean that one or the other terminology would be effectively redundant.

More recent research and reports indicate that the Common Ontological Foundation approach may not be as viable as first believed, with Schulz et al³ offering lessons from the joint IHTSDO and WHO harmonisation efforts, saying:

"We provide evidence for our hypothesis that this cannot be appropriately done by simple ontology alignment, due to diverging ontological commitment between the two terminology systems."

³ Schulz, S; Rodrigues, J-M; Rector, A; Spackman, K; Campbell, J; Ustun, B; Chute, C.G; Solbrig, H; Della Mea, V; Millar, J and Persson, K. What's in a class? Lessons learnt from ICD-SNOMED CT harmonisation. Stud Health Technol Inform. 2014; 205:1038-44.

Re-purposing techniques being developed and explored

More recent work on data analytics and query languages⁴ have motivated NEHTA and AEHRC to examine a different approach to re-purposing SNOMED CT encoded patient data, to **mimic** a secondary, statistical report that relies on ICD encoded data.

We are being careful to state that these methods only mimic the outcomes we might want to see; we do not claim that they are equivalent to existing reporting protocols.

We cannot make this claim at this point in time because there are no suitable patient data collections, encoded in SNOMED CT, to provide a proper controlled and comparative test.

The other reason we are being modest in our claims, is that it is not our place to assert that the outcomes required by secondary data users have been met by our techniques; that is a matter for health data analysts, AIHW and NISCC to confirm.

It may also be the case that these existing reporting requirements look as they do solely because they have traditionally relied on ICD-10-AM encoded data. Our techniques may open up other possibilities that data analysts, AIHW or NHISCC would like to explore further.

Further evaluation will be needed, along with engagement and consultation with downstream health data experts and users.

But for now, the following diagrams, examples and explanations outline the

- Nature of the problems encountered when using SNOMED CT in a secondary data use scenario
- Potential under- and over- counting of patient cases if using SNOMED CT in its native form to aggregate counts of patient cases
- Our approach to re-purposing existing SNOMED CT relationships in order to specify a single pathway to reach mutually exclusive reporting categories (without inflating or excluding patient cases from the aggregation)
- Limitations of our approach
- Considerations for further development of either SNOMED CT or statistical reporting requirements

As stated in the introductory Caveats section, we focus here only on

- Problem Diagnosis Reference Set content (realised by NEHTA, derived from SNOMED CT)
- APNMDS Principal Diagnosis report (as published ^{5 6})

These are used as convenience samples that might reflect real world practices and requirements.

Please note also that all diagrams are merely **partial** representations of the content contained in both Problem Diagnosis RefSet and the Principal Diagnosis report. Only selected examples are shown because a full rendition would not fit on the page and be readable.

⁴ xdoc_SNOMEDCTDataAnalytics_Current-en-US_INT_20141

⁵ The Australian Institute of Health and Welfare, "Admitted patient care NMDS 2015-16." pp. 1–130, 2014.

⁶ The Australian Institute of Health and Welfare, "Principal diagnosis data cubes," 2008-2013. [Online]. Available: <http://www.aihw.gov.au/hospitals-data/principal-diagnosis-data-cubes/>. [Accessed: 23-Oct-2014].

Context

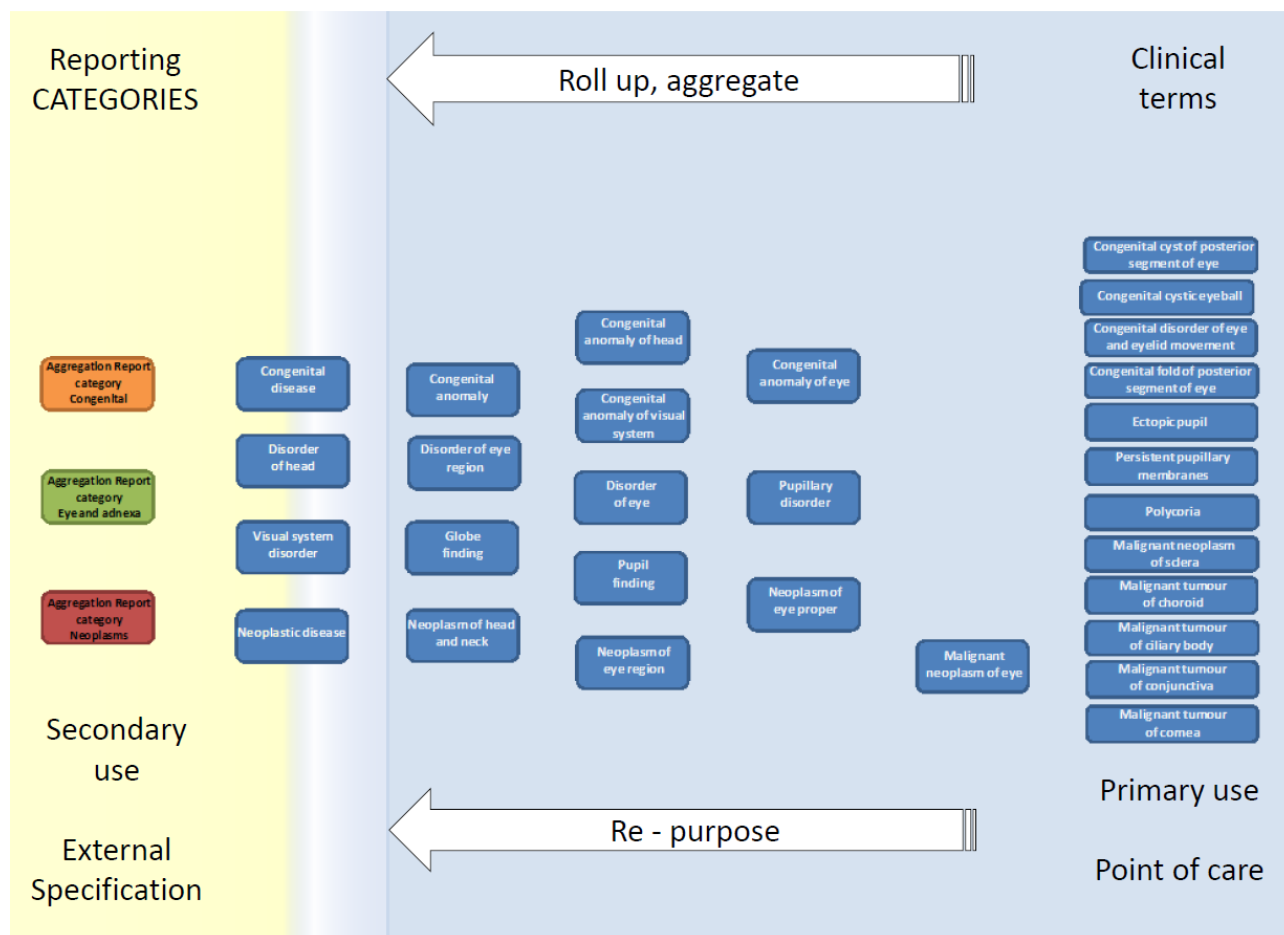


Figure 1: Operating environment, clinical documentation to statistical reports

Figure 1 shows examples of

Problem Diagnosis RefSet concepts that will be documented in medical records (by clinicians, point of care: Blue boxes, blue background)

These are the inputs (primary).

Reporting categories that are specified by the APNMDS Principal Diagnosis report. (Orange, green and red boxes, yellow background)

These are not specified with regard to SNOMED CT concepts, but rather (generally) represent ICD-10-AM chapters. Clinical coders and HIMs undertake the assignment of ICD-10-AM codes, and then data analysts write data queries that extract and assign patient cases according to code assignment and APNMDS report specifications.

These are the outputs (secondary).

Patient cases that are encoded with the Problem Diagnosis RefSet concepts (right hand side) need to be counted, accumulated and reported in the 'correct' reporting categories (left hand side).

The arrows show the direction of re-purposing and the direction in which patient case counts should be accumulated.

The question is how does, or should, that happen?

The nature of the problem

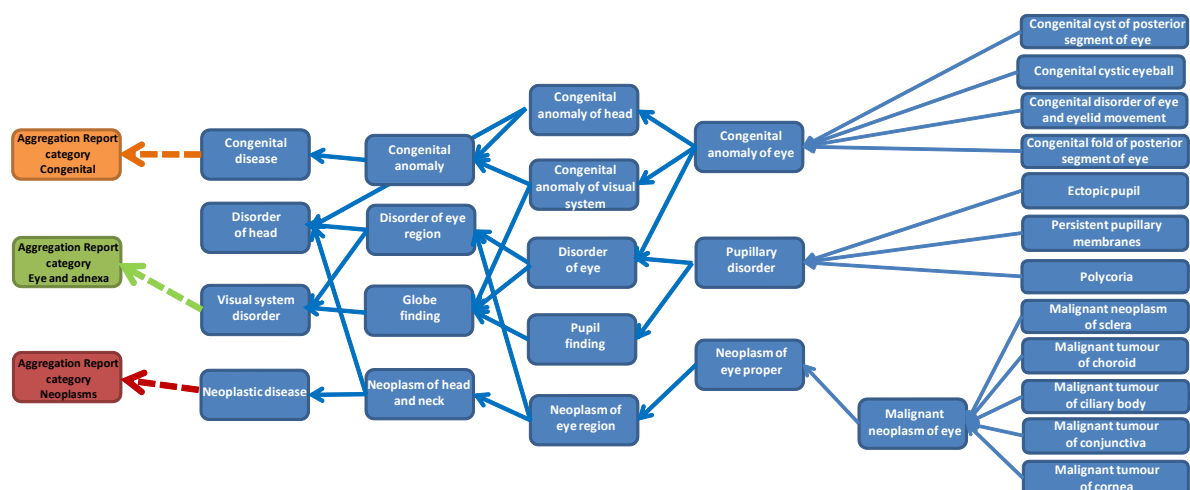


Figure 2: Problem Diagnosis RefSet relationships and pathways, heading towards the aggregated report categories

In Figure 2 the orange, green and red arrows show how we might bridge the gap between **internal** Problem Diagnosis RefSet concepts (at the highest level), and the **externally specified** (non SNOMED CT) Reporting categories and their labels.

These relationships do not currently exist within any SNOMED CT product; they would have to be authored and provided to users as metadata or specified in some other form and format.

All the blue arrows show the relationships between Problems Diagnosis RefSet concepts, as they currently exist in SNOMED CT.

This is what a poly-hierarchy looks like.

A poly-hierarchy provides numerous pathways (not just one way) of getting from input to output.

If we have numerous pathways, we can almost be certain that different users will each elect to take different pathways traversing the data, and counting patient cases from clinical input to reporting output.

Under those conditions, we could wave goodbye to standard and comparable national statistical reports.

Given that there are a number of pathways available in the Problem Diagnosis RefSet, is it the case that one pathway might be more suited to secondary data reporting requirements?

The trick is choosing 'the right one'.

Under-counting? Or perhaps or not counting at all

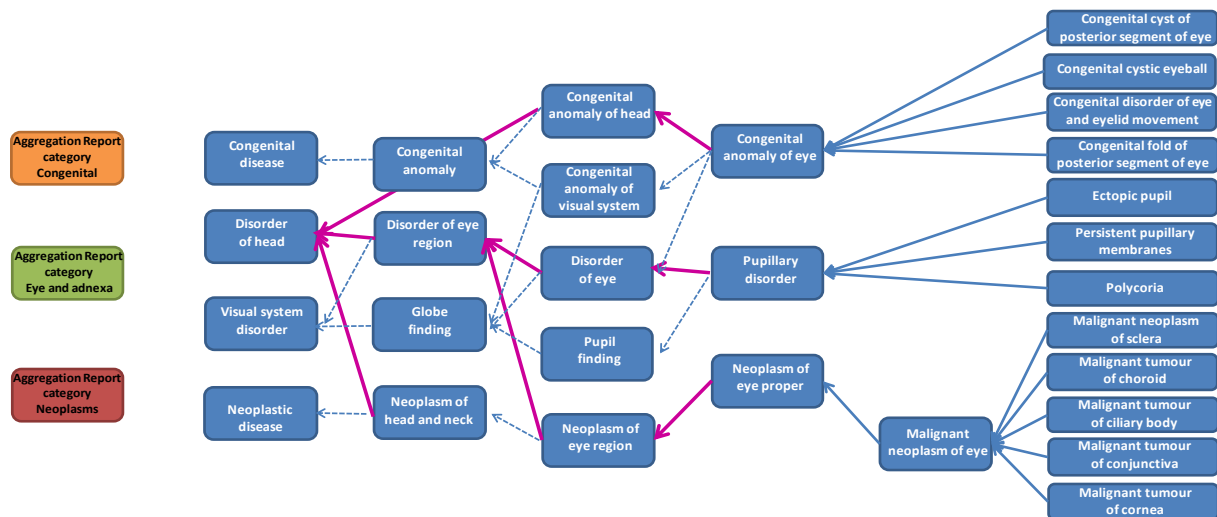


Figure 3: Choose the wrong aggregation pathway and lose patient data from secondary reports

Figure 3 shows one possible pathway that count aggregate counts of patient cases using SOME of the existing Problem Diagnosis RefSet relationships, and tracing through using a SINGLE pathway.

Patient cases would be counted along the paths shown by solid blue, and then solid pink arrows.

It is notable here that the solid pink arrows best serve statistical reporting purposes, and the blue arrows still allow some clinical descriptiveness to exist ‘close-to-the user’. Also note that there is only ONE solid pink arrow specifying a pathway from one concept to another. This mimics mutual exclusivity required for accurate statistically valid counts.

No patient cases (from left-most input, to right-most outputs) would be aggregated along the pathways shown with dashed blue arrows.

But there is a problem with this pathway.

All concepts aggregate to a high level Problem Diagnosis RefSet concept called “Disorder of the head”

The external Aggregation Report does NOT specify a category for all diseases that might manifest in a patient’s head.

All counts that were required by the secondary data output report are lost, because they stop within the Problem Diagnosis RefSet highest level, and have ‘nowhere to go’.

Concepts and patient case counts that SHOULD be reported in Congenital, Eye or Neoplastic categories do not appear; they are excluded from the secondary data report because we traversed the wrong pathway.

Under- and over-counting

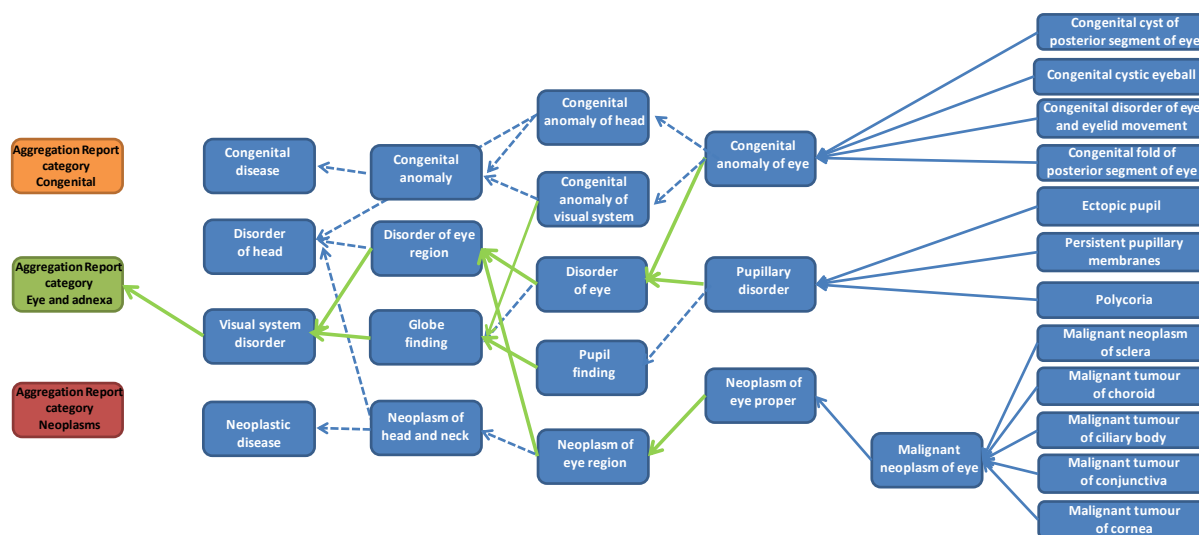


Figure 4: Take a different pathway and get a different outcome

Figure 4 shows what happens when we take a different pathway; we get a different result.

Again the solid blue and solid green arrows represent ONE of the available pathways from clinical input through to reporting outputs.

We are ignoring the existing Problem Diagnosis RefSet relationships shown with dashed blue arrows.

Here we can see that we reach ONE of the required Reporting Categories; it specifies a class for all Eye and Adnexa diseases.

It is a somewhat more circuitous route, but all roads lead to Eye and Adnexa disorders.

There are two non-optimal outcomes:

- (1) Eye and Adnexa disease counts are **inflated (this is over-counting)** because this category now also contains patient case counts for conditions relevant to Congenital and Neoplastic disorders.
- (2) Patient case counts for Neoplastic and Congenital conditions are **deflated (this is under-counting)** because these patient cases do not reach those categories but instead are now 'hiding' in the Eye and Adnexa disease category.

The more optional pathways that are 'allowed,' the more divergent and inaccurate the aggregated counts of patient cases will be given the secondary reporting categories that are specified.

Pruning: specifying the preferred pathway for aggregation

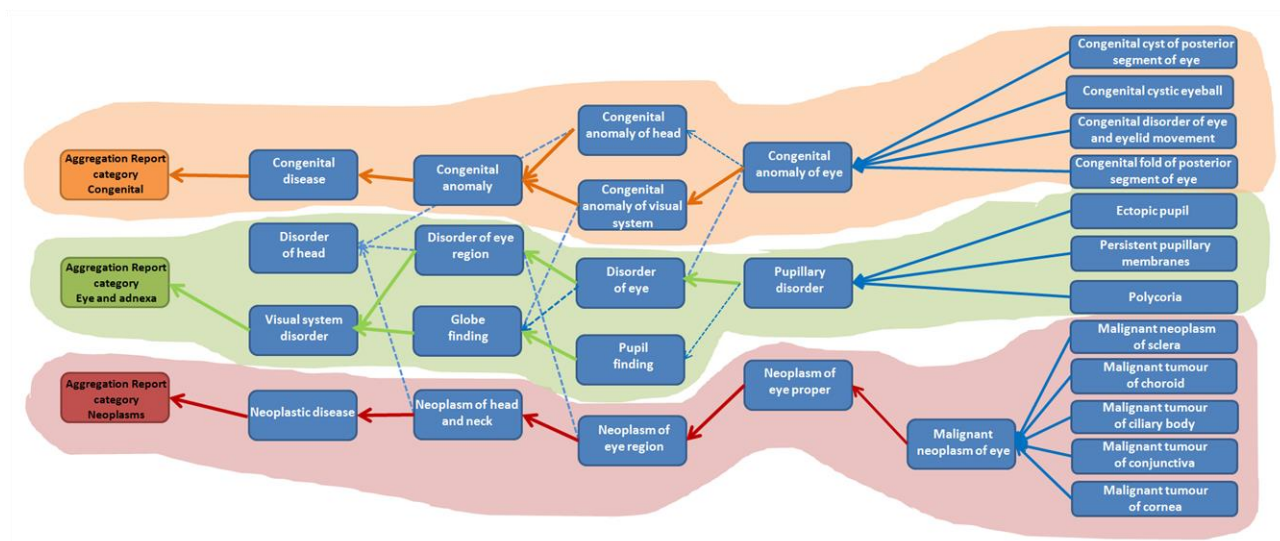


Figure 5: Specifying the correct pathway, correct aggregation of patient case counts to achieve the correct Reporting category outputs

Figure 5 shows the results of our attempts to carefully direct Problem Diagnosis concepts and their patient case counts along distinct, mutually exclusive, pathways to ensure that the right case counts end up in the right reporting categories; without diverging, without inflating or deflating the patient counts or hiding them away where they are not supposed to be.

Again here, we ignore any existing Problem Diagnosis relationship shown with a blue dashed arrow. We called this ‘ignoring’ of some relationships ‘pruning’.

It’s clear that many of these ‘cross-over’ the coloured backgrounds, and head off in directions we don’t want them to. That is they jump across the coloured backgrounds from red to green, from orange to green and from orange to nowhere. Following those dashed blue arrows ‘corrupts’ the possibility of accurate and required aggregated patient case counts.

We can also see in Figure 5 that concepts joined by solid orange relationship arrows, sit within an orange background, and arrive at their orange Reporting category – with patient case counts intact. Similarly, concepts connected with solid relationship arrows reach their green Reporting category, and the same for our red ones.

Why do we call it ‘pruning’?

Because when we ‘trim back’ all optional and possible relationships from the dense and bushy (ontological) SNOMED CT structure, we end up with a (ICD-like) nosological tree structure, with fewer twigs and branches to travel along, that would allow the patient case counts to be lost or diverted.

Simple example

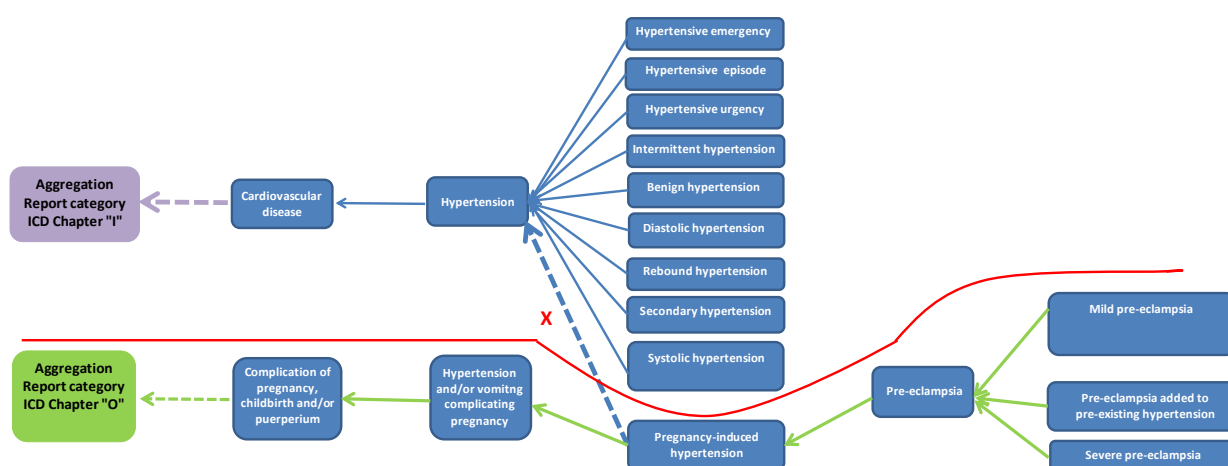


Figure 6: Specifying the preferred aggregation pathway

So how do we specify the correct pathway that patient cases should aggregate along?

The key point is that the external specifications of what the secondary report requires are paramount. Unless we have that, we don't know what to aim for and we only have an approximation at present. It is only an approximation because it deals exclusively with ICD encoded data, and is silent about what we might do with SNOMED CT encoded data.

There are hundreds of pages of technical documents that tell us "how" to do this, (or at least *suggest* how we might proceed), and we're happy to refer the interested reader to those resources.

For now, Figure 6 shows that based on the specification of the secondary aggregation report, we **ignore the Problem Diagnosis relationship that does not meet the report specification**.

This example shows that the Problem Diagnosis RefSet relationships optionally direct Pregnancy-induced hypertension, to a Hypertension class (shown with a blue dashed arrow), and then onwards to Cardiovascular diseases and the Chapter I reporting category.

We 'prune' this relationship from the preferred aggregation pathway (marked with a red 'X') and the only direction that remains for this concept, and its patient case counts, is to head toward Complications of Pregnancy and thence to Chapter O reporting categories, where the secondary use case specifies that it should be counted.

This has the effect of quarantining patient case counts destined for the Chapter I reporting category from those that are meant to end up in the Chapter O Reporting category (the quarantine line is shown in red).

Technical details

The techniques we used to trial this approach are fairly well known in computer-science, and are regarded as generally useful and predictable, graph traversal methods.

These include the use of algorithms such as

- Most specific common ancestors
- Breadth first search
- Transitive closure reduction

These are coupled with SNOMED CT query language and analytics designed specifically to deal with the SNOMED CT structure, logic and content.

When the preferred aggregation pathways have been identified using these techniques (as shown in Figures 5 and 6), the relationships that we ignored are pruned from the standard release relationships table.

A new table is produced that we call the aggregation edges.

The standard release relationships table for Problem Diagnosis RefSet contains ~163 000 rows. This means that all possible pathways are represented for the (~95 000 concepts in the Problem Diagnosis RefSet);

The aggregation edge table produced by this pruning technique includes ~95 000 rows. This means that for each of the ~95 000 concepts in the Problem Diagnosis RefSet, they have one aggregation path and one only.

Using the aggregation edges table

The usual approach to health data aggregation and analysis applies. The necessary data collection containing patient case records, along with the aggregation edges table, is analysed using a query (as shown below).

```
public int calculateFreq (Concept c)
{
    int total = 0;
    if (originalFreq.containsKey(c))
    {
        total = originalFreq.get(c);
    }
    for (Concept child : c.getChildren())
    {
        total = total + calculateFreq(child);
    }
    return total;
}
```

Results

When we tested this technique using a synthetic dataset of SNOMED CT encoded patient cases, the data returned showed that all cases were preserved, allocated and counted cumulatively into 21 top level reporting categories. There was no loss of data.

We cannot claim that these results are definitive, or even acceptable. This is a proof of concept investigation, and it does not employ real patient case data. There are some secondary reporting specifications, but these are insufficiently SNOMED CT-aware, so we cannot measure the goodness-of-fit of these results against desired outputs.

There are some known limitations.

[illegible]

S

Sadly, no.

There are no existing SNOMED CT relationships that could possibly direct aggregated patient case counts to the desired and specified Reporting categories (existing relationships are shown with solid blue arrows; there are no solid blue arrows that reach the destination we want).

SNOMED CT commits *Diabetes during pregnancy* to two classes (*Diabetes mellitus and hypoglycaemia*); whereas ICD and our Reporting use case commits this concept to a different, single class (only) of *Complications of Pregnancy*.

It is also worth noting here in Figure 7, that the mapped ICD-10-AM codes do not perform much better for aggregation purposes. Some codes are unspecified (shown in yellow text with 'x' or 'xx'); some are ranges E10-E14 (with a lack of specific) commitment⁷.

The short answer is we don't know yet. And we can't know until such time as secondary data use case specifications are SNOMED CT-aware, and we have SNOMED CT encoded patient data to test, measure and evaluate these approaches.

What happens now?

There are ongoing developments in data analytic techniques being pursued by the IHTSDO, and with contributions from NEHTA and CSIRO representatives on those working groups.

It is expected that those developments will bring to light other technical options and refinements to the methods used here.

Beginning a conversation with secondary data custodians, those people who design secondary data collection protocols and specifications, is an essential first step. These consultations should offer an opportunity to explore options for SNOMED CT-encoded health data, and requirements for aggregation or reporting.

Further evaluation of these techniques also needs to be undertaken, subject to the availability of data collections of patient case records, encoded in SNOMED CT. Importantly, such an evaluation will also need to address whether these techniques provide the desired functionality and outcomes that secondary data users require.

Section B

1 Introduction

This report and related aggregation Reference Set files fulfil the work program specified in 001687.

The work program was designed to investigate obstacles and options related to the use of SNOMED CT in a secondary data use scenario. SNOMED CT is designed to be used by clinicians to document patient medical records. It is not entirely suited, in its original form, for secondary data purposes, especially statistical use cases.

Our work here focuses on techniques that can transform the SNOMED CT Problem Diagnosis Reference Set into an annotated Reference Set (RefSet), so that SNOMED CT content can be re-purposed for statistical uses, without relying on the use of maps, which would impose additional burdens and costs on the user community for map production, maintenance, synchronising between versions, and deployment.

The secondary data scenario we have adopted for demonstration purposes is the Admitted Patient National Minimum DataSet (APNMDS). All States and Territories make contributions to this collection, and AIHW produce annual reports that utilise the information. This use case is broadly familiar and well understood by the health information practitioner community.

The APNMDS has a range of data specifications, including demographics, episode data (admission and discharge types), insurance status data, length of stay, Procedures and DRGs.

However, the APNMDS also has a prominent requirement for diagnosis information, and it is here that the primary use of SNOMED CT terminology content to document in patient records intersects with secondary data uses.

The SNOMED CT terminology content that is most applicable to this use case is the Problem Diagnosis Reference Set (der2_RefSet_ProblemDiagnosisFull_AU1000036_20140531).

The report and proof-of-concept annotated RefSets will be suited for consultation with health data custodians and stakeholders. The examples provided here are particularly relevant to the Department of Health (DoH), as well as State and Territory Health Data Units and the Australian Institute of Health and Welfare (AIHW), as they have a major investment and interest in secondary and statistical data collections.

Their consideration and feedback about this approach and assessing whether native SNOMED CT aggregation does faithfully reflect the results attained using ICD-10-AM will be essential. At this time it is not possible to empirically measure the approach or outcomes because:

- Any reliance on maps to judge comparability between reports based on SNOMED CT and ICD-10-AM will not provide a valid measure because maps themselves potentially introduce gains or losses in data fidelity (via many-to-one, many-to-many maps, contextual or rule-based maps, no available maps or maps that either specialise or generalise the meaning of the mapped concepts).
- There are no SNOMED CT and ICD-10-AM (dual-coded⁸) patient data collections; this means that there is no unbiased ground truth that could be regarded as a comparative benchmark.
- There have been no blind or double-blind re-coding studies in SNOMED CT and ICD-10-AM to provide guidance on gaps or inaccuracies in medical record documentation practices.
- There are few relevant SNOMED CT encoded patient data collections available to data-mine, and to comparatively test.

Ongoing investigation and collaboration with secondary data users and custodians will help to better specify requirements and refine these technical approaches.

⁸ (i) by clinicians in SNOMED CT at the point of care, in the original medical record, and then (ii) by clinical coders or HIMs in ICD-10-AM for data collection purposes using APNMDS rules and conventions

2 Background

SNOMED CT is primarily designed to serve clinical use cases; it describes and defines concepts in a way that is clinically meaningful and precise. Its structure is poly-hierarchical and multi-parented. These features mean that SNOMED CT - out of the box - is not suited to secondary data purposes, which (in the main) demand mono-hierarchies and single parenting so as to avoid double counting patient cases, and a primary aggregation path to cumulatively count and categorise patient cases into the desired statistically valid reporting classes.

Also SNOMED CT, unlike ICD products, does not contain many concepts that represent secondary reporting categories. ICD includes Not Otherwise Specified (NOS) and Not Elsewhere Classified (NEC) codes; the word **not** gives the game away and highlights that ICD has 'residual' or catch-all categories. These concepts also mean that patient cases recorded with these codes may not necessarily have experienced the condition described by any other code in that rubric as shown in Figure 8.

We see here that A41.1 means *Sepsis due to other specified staphylococcus*, but **not** *Staphylococcus aureus* (A41.0), and that the difference between code assignment of A41.1 and A41.2 is whether the medical records documentation is specific enough to differentiate the organism species that has caused the sepsis in this patient case. Codes A41.1 and A41.2 **cannot** be interpreted as asserting sepsis due to *Staphylococcus aureus*; these exclude *aureus*. The concept for code A41.8 means that the patient case of sepsis is **not** attributed to *Staphylococcus aureus*, *other specified staphylococcus*, *unspecified staphylococcus*, *Haemophilus influenza*, *anaerobes*, *any other Gram-negative organisms*. The use of the code A41.8 literally means that the patient case of sepsis is **not due to any of these things**. Similarly, but differently, the concept for A41.9 denotes that the patient case of sepsis is **not any of the above concepts**, but there is too little clinical documentation to assign a more accurate code.

- ▼ A41 Other sepsis
 - A41.0 Sepsis due to *Staphylococcus aureus*
 - A41.1 Sepsis due to other specified *staphylococcus*
 - A41.2 Sepsis due to unspecified *staphylococcus*
 - A41.3 Sepsis due to *Haemophilus influenzae*
 - A41.4 Sepsis due to anaerobes
 - A41.5 Sepsis due to other Gram-negative organisms
 - A41.8 Other specified sepsis
 - A41.9 Sepsis, unspecified

Figure 8: Residuals and exclusions in ICD

There are distinct structural differences with SNOMED CT providing specificity and ICD providing sensitivity. The availability of residual categories means that any patient case, no matter how under-specified or vague the clinical documentation, can be included in a data collection, in a mutually exclusive fashion.

The key idea here is that of categories over concepts.

Australia has a long and successful history of establishing and maintaining secondary health data collections. Considerable investment in these ICD encoded data stores, and their support systems and expertise precedes the implementation of SNOMED CT. Arguably, the entrenched use of statistical classifications, a trained HIM workforce and built-for-purpose data warehouses all function as barriers to SNOMED CT adoption.

Mapping between SNOMED CT and existing classifications is regarded as a way to overcome those perceived barriers. However, after more than a decade of effort, mapping strategies have not significantly influenced health information management practice in Australia, nor have they succeeded in accelerating SNOMED CT adoption. In some ways, the development, use and maintenance of maps increases burden and costs to practitioners, managing complex and variable data specifications and protocols, and so engenders even further resistance to adoption.

Obviously there are numerous secondary and downstream data uses and each will have distinct requirements. This will mean that the re-use and re-purposing of SNOMED CT encoded patient data will also have distinct specifications.

Like maps, aggregation techniques will not be a one-size-fits-all proposition; each secondary use case will need a separate aggregation layer, a reference set of SNOMED CT input concepts, tailored to suit the required secondary data reporting outputs.

However, the use of SNOMED CT at least standardises the data inputs, and the technical approach to aggregation, while avoiding the need to map to or reference other vocabularies and non-standard legacy termsets. That is, the re-purposing of SNOMED CT, in an aggregation RefSet format, removes externalities and dependencies on other health information instruments, leaving implementers and users to deal only with SNOMED CT RefSets that have stable and predictable product formats and release cycles.

It is also apparent that deployments of SNOMED CT in different clinical systems will be influenced by the architecture and the information models inherent in those systems. Features like searching and indexing functionality, as well as data element definitions, will mediate or moderate the use of any terminology, not just SNOMED CT. However, these factors cannot be manipulated, interpreted or taken into account by terminology content alone; SNOMED CT cannot account for information that exists outside its scope, such as demographic data, patient encounter information or billing schemes. Hence, we are focussing only on SNOMED CT products, and their ability to be used downstream; by necessity, we regard information model variables as out-of-scope for the purpose of this investigation.

NT Cheung⁹ spoke recently at the IHTSDO Showcase in Amsterdam. One key principle for the HKHA was that clinicians only record clinical data and thus downstream data and analytics should be a by-product of the use of a clinical terminology. This will require that terminology inputs from the primary collection point are re-purposed and refined for those additional purposes as we propose here.

For a fully interoperable and standardised approach to be realised, further consideration of secondary data scenarios that utilise SNOMED CT will be needed, with particular attention given to connectivity between health care sectors, data flows and transformation specifications that account for both the clinical and statistical nature of health information.

⁹ N. T. Cheung, "Making Health Records Make Sense in Hong Kong," in *James Read Memorial Lecture 2014*, 2014.

3 Use cases

Use cases specify the purpose of the health data collection. Re-purposing techniques must be designed and deployed to serve the use case.

Most secondary use cases demand counting of patient cases in a way that is statistically valid, providing both power and sensitivity¹⁰. However, different use cases may have different levels of tolerance for sensitivity and specificity, and for false positive or true negative signals within the data.

In his recent presentation about SNOMED CT induced classifications, Jeremy Rogers¹¹ characterises these two types as ‘static’ or 20th century approaches, and ‘dynamic’ or 21st century approaches.

While we agree with the labelling of these different use cases as static and dynamic, we do not agree that one is old and the other new, or that either can or should replace the other. Rather, we believe both are valid, currently used, and will continue to be required as essential secondary data analysis techniques.

Both these types of secondary data use cases are prominent in the Australian health information domain.

3.1 Dynamic use cases

Dynamic use cases involve allowing SNOMED CT concepts to be aggregated and counted based on their rate of incidence in a patient population. These are signal detection use cases, where threshold measures of incidence and prevalence of particular diseases, conditions, symptoms or patient presentations can and will change over time. These may be seasonal changes, where we might see certain influenza conditions come to prominence during winter months.

Essentially aggregation here is used for surveillance purposes, and it is necessary that any ‘induced classification’ is capable of distinguishing signals from noise in the data. The aggregation needs to be dynamic and changeable, so that diseases or conditions that are not routinely seen in the population can be detected if there is a ‘spike’ in the patient community.

Many state and territory health departments have recently introduced new algorithms, new thresholds and new techniques in order to be alerted to any small increase in patient experiences of symptoms or signs related to Ebola, six months ago the terminology related to Ebola would not have been flagged as a ‘trigger’ concept for these public health monitoring systems.

Hence, as the population health changes, so does the terminology, the aggregation categories and surveillance protocols. They still require some form of aggregation in order to filter signal from noise and to set the detection levels or thresholds. A key requirement here is for epidemiological and statistical (Bayesian) measures and detection thresholds to be specified so that terminology aggregation techniques are capable of providing the correct trigger words at the required level of specificity.

We can speculate that aggregation techniques to serve this type of use case would provide greater numbers of categories containing smaller numbers of more specific terminology concepts. Obviously, the aggregation technique itself also needs to be agile and dynamic to keep pace with epidemiological investigations. This means that aggregation categories for high level reporting will be different seasonally,

¹⁰ Statistical power is affected chiefly by the size of the effect and the size of the sample used to detect it. Bigger effects are easier to detect than smaller effects, while large samples offer greater test sensitivity than small samples. This means that small, heterogeneous and specific categories are less likely to have the power or sensitivity to detect significant differences; it means that larger categories of a fairly homogenous nature, will be required.

¹¹ J. Rogers, “Using SNOMED CT to induce classifications for casemix analytics,” 2014.

or as a result of pandemic, epidemics or changes in (acute) population health. This is much more difficult to achieve, especially given the paucity of epidemiological and surveillance specifications currently available.

Interestingly, Rogers found that using maps between SNOMED CT and ICD for these purposes corrupted the scope of terminology that could be included in signal detection protocols. It is not yet established if this occurs because of the quality of the maps themselves, the variability of the granularity of maps (1:1 or 1:N or N:1) or the fact that Roger's experimental aggregation and induction method did not control for the known issue of 'double counting'.

3.2 Static use cases

Static use cases are those where the aggregation categories are stable over time. These provide a high level and comparable view of health statistics longitudinally. Regardless of seasonal variations in disease incidence and prevalence, the reporting categories remain the same. This allows comparisons of diagnoses over time and the ability to measure health service utilisation, major health challenges, the effectiveness of preventive health programs, or the influence of lifestyle factors on population health and health service utilisation.

Essentially longitudinal or trend analyses requires comparable categories and reports over time, the same variables measured perhaps annually as in the case of the AIHW Australian Hospital Statistics and APNMDS reports or the World Health Organization's Global Burden of Disease reports¹² (see also Section 6 for more detailed information).

Such a high level and long term comparison affords health system managers a meaningful snapshot view, as well as the ability to discern any major shifts in identified variables of interest over regular time periods, usually encapsulated in a single table or chart. Such overviews function as an executive summary of health system utilisation and population health in a readily accessible and digestible format.

These static reporting categories have a limited scope for specificity; we say they are quite lumpy. At most, the drill down and roll up functions within the data do not extend by more than 3 levels.

This is because ICD products are traditionally used for these static secondary use purposes. The full name of ICD is the International *Statistical* Classification of Diseases and Related Health Problems. The word statistical is apposite. ICD is built for these statistical purposes (only).

It goes without saying that once static, high level and general categories are defined and established they cannot be re-purposed for other clinical uses; once collected and encoded it is not possible to refine lumpy and general data into its constituent parts when those constituent parts were never originally captured. That is, we cannot 'unpack' ICD encoded data and attribute the patient case counts to SNOMED CT concepts. But the reverse is possible; aggregation from primary information source, richly clinically descriptive and specific SNOMED CT concepts can be achieved to build those higher level, more general reporting categories.

¹² Assessment of Global Burden of Disease 2010 methods for the Australian context Australian Burden of Disease Study Working paper No. 1
<http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=60129547710>

p15: **Cause (or condition) list**

Underpinning any burden of disease analysis is a cause (or condition) list. This is an exhaustive list of diseases based on International Classification of Diseases (ICD) codes for which analysis is meaningful and possible. The cause list also forms the basis for risk factor analysis, where the burden attributed to a condition can be related to a particular risk factor—for example, a portion of all lung cancer is caused by smoking (the risk factor). The 2003 Australian burden of disease study included estimates for 186 diseases and 14 risk factors. In contrast, GBD 2010 reported estimates for 241 individual diseases and causes of injury (described by 1,160 sequelae), and 57 individual risk factors.

ICD products themselves are relatively static and stable (compared to SNOMED CT). ICD editions are revised and released every two years, compared to the SNOMED CT release cycle of every 6 months. This ICD stability means that once we re-purpose SNOMED CT to aggregate and mimic APNMDS reporting, changes to these aggregation protocols are also predictable and fairly stable (every two years).

4 Problem Diagnosis Reference Set

The Problem Diagnosis Reference Set was created with the intention to have the broadest possible use across varying care sectors. It contains 98 610 active concepts from the various SNOMED CT hierarchies. Table 1 shows an overview of the hierarchical makeup of the reference set.

Table 1: Hierarchical distribution of Problem Diagnosis Reference Set concept members

SNOMED CT hierarchy	Count
Clinical finding	97457
Event	996
Situation with explicit context	141

The Problem Diagnosis RefSet is (mostly) an intensional RefSet and has been defined and produced by using a reusable query that identifies and includes a subset of concepts from SNOMED CT.

As shown in Table 1, the vast majority of SNOMED CT content is drawn from the Clinical Finding hierarchy. However, additional concepts from other hierarchies are also present. This means that there is some extensional content included, by specification.

The effect is that the Problem Diagnosis RefSet is largely reproducible by query with some, though minimal, hand crafting. This provides efficiencies for the National Release Centre, and some predictability for implementers.

It is clear from the RefSet membership that consideration has been given to the ‘kinds’ of concepts that might be routinely required by clinical users in order to document patient problems and disorders, as well as the events and situations that may motivate patients to seek care services. Events, such as motor vehicle accidents or a fall from a ladder will cause patients to attend the Emergency Department. Patients might also find themselves in situations where they perceive that they are at risk of domestic violence, and will seek advice and assistance from their General Practitioner.

Taking a strict and purist view, such concepts might be considered out-of-scope for documenting conditions in data fields specified as Presenting Problem or Diagnosis. Nonetheless these remain valid and are broadly used by the clinical community in these data fields. Information models, data custodians, software engineers and indeed paper based forms routinely seek to restrict clinical users to a more disciplined use of terminology, but such efforts have never been successful.

Additional extensional content may need to be included in the Problem Diagnosis RefSet in future releases. These issues are discussed further in Sections 5.1 and 7.2.

4.1 Scope

Because the Problem Diagnosis RefSet is constructed via intensional techniques (with some extensional content), it can be regarded as comprehensive. Its size (98 610 concepts) certainly indicates that most SNOMED CT content that is relevant for documenting patient conditions is included in this RefSet.

With the inclusion of concepts from varying hierarchies other than Clinical Findings the reference set is intended to suit use cases within and across care sectors such as hospital (acute), sub-acute, rehabilitation, mental health services, community care, primary care (GP), aged care and specialists.

Although the Problem Diagnosis RefSet content is broadly applicable to most clinical users, this will not mean that each user needs all the concepts. Their own care sector will determine which concepts will be selected to describe and document their patients and their clinical conditions.

This means that there will be differences in the frequency of use of Problem Diagnosis RefSet members between carers in different health service sectors.

It is obvious that:

- patients who suffer gunshot wounds or severe multiple injuries in a motor vehicle accident will never attend a GP clinic; they will be transported to an Emergency Department
- patients who are experiencing transplanted organ failure are most likely to be admitted to hospital and/or an intensive care unit; they are unlikely to be seen in a community or ambulatory care setting
- patients who have a normal delivery of a child may be admitted to hospital, or may be cared for by a midwife at home
- patients who suffer (only) “sniffles” will not be admitted to hospital; they are most likely to be treated by their GP
- patients who have a glue sniffing dependence are most likely to receive a program of treatment in community care settings and may only be seen infrequently in an acute setting (hospital) or GP clinic
- patients who have cataracts may be treated in a hospital or by an ophthalmology specialist in a day surgery clinic

The range and frequency of use of terminology content that is included in collections from these different sectors will only ever reflect the patient experience and the care delivery that actually occurs in those sectors.

The material effect is that different sectors will require different aggregation techniques specifically designed to reflect the casemix of patients routinely seen in each sector. This is unsurprising.

However, even though there will be distinctive statistical counts and preferred aggregation techniques across care sectors, the underlying Problem Diagnosis RefSet content is the same, and commonly available to all users. This foundation enables HL7, CDA or FHIR messaging and exchange, where the entire community of care can receive and interpret clinical information, referrals, follow up care and discharge summaries. The common use of a broad RefSet connects care sectors, follows the patient journey across multiple care providers and supports semantic interoperability.

4.2 Clinical Finding Grouper Exclusion RefSet

Consideration was given to the intersection between the Problem Diagnosis RefSet and the Clinical Finding Grouper Exclusion RefSet (CFGE RefSet). These two RefSet share 3781 concepts; ie 3781 CFGE RefSet members are also members of the Problem Diagnosis RefSet.

The CFGE RefSet identifies 4011 concepts from the Clinical Finding hierarchy that are ‘placeholder’ concepts, designed to assist navigation, indexing and retrieval of SNOMED CT concepts. They do not carry sufficient clinical meaning that could motivate or justify their use at the point of care for clinical documentation purposes.

Concepts like *Finding by Site*, *Finding by Method*, *Clinical history and observation findings*, *Abnormal histology findings* do not describe or define patient conditions. These sorts of concepts are parent (or ancestor) concepts that name and label a group of descendants (note the plurals). To some extent we might characterise these sorts of concepts as ‘structural’ rather than meaningful.

Because these 4011 concepts carry little clinical meaning, the CFGE RefSet was developed as a way of excluding them in clinician facing systems. This prevents them from ‘clogging up’ search and search return

functionality. It also allows clinicians to favour the selection of appropriately specific and meaningful terms for documentation, hence improving accuracy and validity of patient records.¹³

If implementers properly deploy SNOMED CT-AU content and conform to RefSet guidance, then these concepts will never be selectable for clinical documentation purposes.

This means that a patient case count will never be attributed to these concepts.

We contemplated removing the CFGE concepts from the Problem Diagnosis RefSet before we attempted to define an aggregation technique. However, because these CFGE concepts do provide some level of identifying similarly defined conditions we first tested this notion, and then elected to allow them to remain in scope. These CFGE concepts may later be useful for drilling down or rolling up purposes or for retrieval queries. That is, for secondary data use purposes CFGE RefSet concepts continue to be useful, even though they will not be available for clinical documentation purposes and will not have attributed patient case counts.

4.3 Findings versus disorders

All health vocabularies have some peculiar features, and all of them are peculiar in different ways.

There is a long running debate between some terminologists about disentangling concepts that are truly symptoms from concepts that are truly disorders.

Unfortunately this debate will not be resolved, and attempts to model symptoms and disorder as disjoint sets are doomed to failure; it entails a referential fallacy.

We see this in ICD products; the classification itself is called the International Statistical Classification of **Diseases** and Related Health **Problems**.

Both disorders and symptoms are included in the classification, although some ‘symptom’ concepts are quarantined in separate chapters (18 or 21). This is possible in ICD because of mutual exclusivity. It should be noted that such a distinction is ‘forced choice’ by the structural conventions of ICD.

Similarly SNOMED CT includes a range of health conditions including symptoms, findings, observations and evaluations as well as diseases. But SNOMED CT does not distinguish between symptoms and diseases. Rather, SNOMED CT has ‘types’ of concepts labelled Findings and Disorders. These are not entirely disjoint.

The rule of thumb adopted by SNOMED CT is that Findings can be normal conditions or abnormal conditions, but Disorders are always defined as abnormal. Findings therefore **do not** equate in any meaningful way with symptoms. All Disorders are descendants of Findings.

An illustrative example of the different structural approaches is how ICD and SNOMED CT represent the concept of 372070002 *Gangrenous disorder*.

In ICD, R02 *Gangrene not elsewhere classified* is placed within the *Signs and Symptoms* chapter indicating that it is considered a symptom.

In SNOMED CT the equivalent concept is *Gangrenous disorder* and is placed in Clinical Findings/Disease hierarchy, and defined as a *Traumatic and/or non-traumatic injury* indicating it is a *disorder*, and not a normal patient condition.

Further, some ICD chapters are agnostic about whether the concept is a symptom or a disorder. For example Chapter 15 *Pregnancy, Childbirth and the Puerperium* does not have a label that indicates disorder or symptom as is apparent in other chapters such as Chapter 11 ***Diseases of the Digestive System*** or Chapter 6 ***Diseases of the nervous system***.

¹³ The CFGE refset is now (November 2014) applied to the Clinical finding foundation refset, which in turn is used by the Problem/diagnosis refset.

The material effect is that users can, will and should be able to select concepts from the Problem Diagnosis RefSet as suits their clinical documentation requirements; these may be Findings or Disorders and all concept selections should qualify for inclusion in any aggregation technique. This is especially true if the Problem-Diagnosis RefSet is destined to be implemented across health care sectors; interoperability between health care sectors is served if all users can select concepts from the same RefSet, exchange them and understand the context of use. This may vary between users, where documenting problems, symptoms, findings, reasons for encounter and diseases are all legitimate requirements.

5 Example use case

This report explores whether SNOMED CT encoded patient data can be re-purposed for a single secondary use scenario – the national minimum dataset specification for admitted patients¹⁴, (APNMDS). Other use cases are known to exist, for example, communicable disease notifications, adverse drug reaction notifications, injury surveillance, but these are not examined here.

The Admitted Patient Care National Minimal DataSet (APNMDS) is used for the purposes of the collection of data for ‘episodes of care for admitted patients’ principally in public and private acute and psychiatric hospitals as well as ‘free standing’ day facilities and alcohol and drug treatment centres, dental hospitals and other specialised acute care facilities. Data is collected and collated on an annual basis per financial year.

Currently this data collection is used in routine publication of national statistical reports by relying on the use of ICD-10-AM 8th Edition (9th Edition pending), a classification purposely developed to serve statistical use^{15 16}. Diagnosis data is published as separation statistics and separated into categories by ICD -10-AM chapter. Diagnosis data is also available by ICD sub chapter as well as to ICD 3rd and 4th and 5th character code level if relevant¹⁷. Existing APNMDS data collections allow drill down and roll up of data into more specific or more general categories, accumulating patient cases in a mutually exclusive fashion.

While SNOMED CT is not currently mandated as a valid value domain attribute of the Data Elements ‘Episode of Care – Principal Diagnosis or Episode of care – Additional diagnosis’¹⁸, it is likely that the increasing use of PCEHR and Discharge Summary information will provide an alternate means of data mining and of collecting and reporting similar information.

The existing disconnect between primary health data collections, where SNOMED CT is recommended for encoding, and secondary data collections that mandate ICD-10-AM encoding, provides yet another barrier to SNOMED CT adoption; it is seen as disruptive to existing practice, procedures and policies.

We speculate that by aggregating SNOMED CT encoded patient data, it is possible to mimic the results and reports that are currently attained using only ICD-10-AM encoded patient data. We hypothesise that it is possible to use SNOMED CT content and structures to produce a comparable aggregation protocol, without resorting to the use of maps between SNOMED CT and ICD-10-AM.

The results and outcomes here are not yet ready to be used in anger or deployed by stakeholders; they are based only on contemporary secondary use specifications, and these rely entirely on ICD-10-AM encoding. Secondary data custodian and mandated health information specifications have not taken account of SNOMED CT encoded patient data¹⁹. Further refinement of NMDS specifications and secondary data use requirements, and extensive testing against real world data would be necessary to ensure that the aggregation RefSet layer provides the necessary precision.

¹⁴ The Australian Institute of Health and Welfare, “Admitted patient care NMDS 2015-16.” pp. 1–130, 2014.

¹⁵ The Australian Institute of Health and Welfare, “Episode of care — principal diagnosis , code (ICD-10-AM 8th edn) ANN { . N [N] }.” pp. 1–4, 2014

¹⁶ World Health Organisation, “International Classification of Diseases (ICD),” 2013. [Online]. Available: <http://www.who.int/classifications/icd/en/>. [Accessed: 25-Oct-2013].

¹⁷ The Australian Institute of Health and Welfare, “Principal diagnosis data cubes,” 2008-2013. [Online]. Available: <http://www.aihw.gov.au/hospitals-data/principal-diagnosis-data-cubes/>. [Accessed: 23-Oct-2014].

¹⁸ The Independent Hospital Pricing Authority, “Episode of care — additional diagnosis , code (ICD-10-AM 9th edn) ANN { . N [N] }.” pp. 1–2, 2014.

¹⁹ The Australian Institute of Health and Welfare, “Activity Based Funding: Emergency service care DSS 2013-2014.” The Australian Institute of Health and Welfare, pp. 1–11, 2012.

5.1 Characteristics of APNMDS reports

The annual reports produced by the AIHW are based on hospital inpatient data collections. Because these are constrained by the APNMDS, the reports themselves reflect both a national and minimalist view of patients admitted to hospitals. That is, these provide a very high level overview of all episodes of care based on principal diagnoses.

The reports generally fit on a single page, in a table or chart as shown below. These provide health system administrators with a meaningful, 'at a glance' perspective, and stable comparative data over many years.

Table 2: Principal Diagnoses attributed to admitted patient episodes by ICD chapter categories 2008-2013 (data summarised from AIHW)

Category	Chapter	08-09		09-10		10-11		11-12		12-13	
		Separations	% of Total Separations	Separations	% of Total Separations	Separations	% of Total Separations	Separations	% of Total Separations	Separations	% of Total Separations
1	Certain infectious and parasitic diseases	118,835	1.5	127,878	1.5	135,670	1.5	141,236	1.5	141,762	1.5
2	Neoplasms	553,564	6.8	579,699	6.8	582,263	6.6	595,279	6.6	600,812	6.4
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	99,736	1.2	102,743	1.2	124,322	1.4	134,106	1.4	139,049	1.5
4	Endocrine, nutritional, and metabolic diseases	164,230	2	167,689	2	121,851	1.4	127,749	1.4	136,913	1.5
5	Mental and behavioural disorders	324,064	4	341,581	4	329,049	3.7	345,119	3.7	357,095	3.8
6	Diseases of the nervous system	197,945	2.4	209,548	2.5	222,343	2.5	236,760	2.5	246,609	2.6
7	Diseases of the eye and adnexa	243,655	3	263,450	3.1	301,731	3.4	323,512	3.4	334,635	3.6
8	Diseases of the ear and mastoid process	55,903	0.7	56,019	0.7	60,678	0.7	60,692	0.7	60,660	0.6
9	Diseases of the circulatory system	474,171	5.8	482,859	5.7	510,752	5.8	523,805	5.8	518,702	5.5
10	Diseases of the respiratory system	369,333	4.5	375,706	4.4	391,398	4.4	404,005	4.4	400,803	4.3
11	Diseases of the digestive system	839,244	10.3	870,708	10.2	890,000	10.1	920,801	10.1	920,728	9.8
12	Diseases of the skin and subcutaneous tissue	132,387	1.6	139,537	1.6	150,138	1.7	154,228	1.7	155,233	1.7
13	Diseases of the musculoskeletal system and connective tissue	435,791	5.3	459,916	5.4	476,628	5.4	494,228	5.4	499,279	5.3
14	Diseases of the genitourinary system	379,754	4.7	395,489	4.6	415,893	4.7	430,881	4.7	434,316	4.6
15	Pregnancy, childbirth, and the puerperium	482,440	5.9	482,195	5.6	477,119	5.4	490,907	5.4	493,667	5.3
16	Certain conditions originating in the perinatal period	56,727	0.7	55,815	0.7	54,788	0.6	63,558	0.6	65,131	0.7
17	Congenital malformations, deformations and chromosomal abnormalities	35,182	0.4	35,130	0.4	34,558	0.4	36,261	0.4	36,987	0.4
18	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	543,751	6.7	579,578	6.8	626,782	7.1	656,522	7.1	655,618	7
19	Injury, poisoning and certain other consequences of external causes	543,229	6.7	557,689	6.5	580,494	6.6	603,992	6.6	601,760	6.4
21	Factors influencing health status and contact with health services	2,094,787	25.7	2,246,878	26.3	2,361,905	26.7	2,508,676	26.7	2,569,254	27.4
	Not reported	3,720	0	5,039	0.1	4,188	0	3,884	0	4,513	0
	Total	8,148,448	100	8,535,146	100	8,852,550	100	9,256,169	100	9,373,526	100

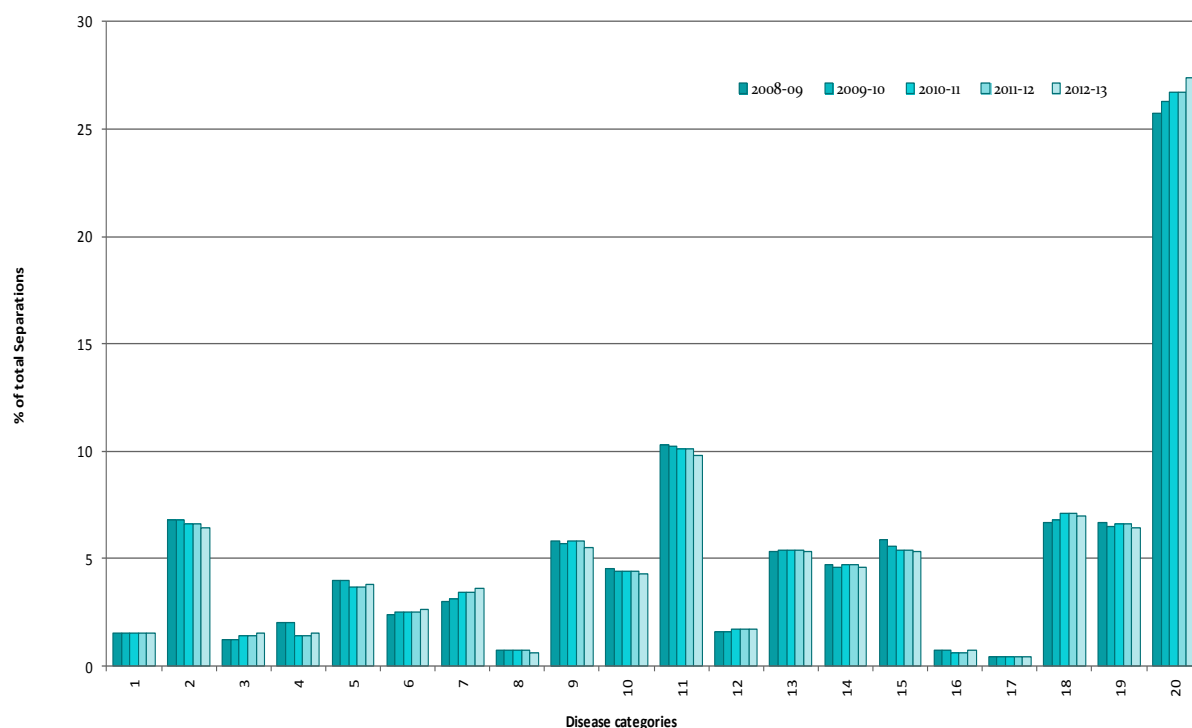


Figure 9: Chart of Principal diagnoses for admitted patient episodes by ICD chapter categories 2008-2013 (data summarised from AIHW)

The noteworthy features include:

- There are only 20 top level categories of diseases reported; these mostly reflect ICD chapters
- Some diseases are responsible for higher volumes of hospital admissions Chapter 2 *Neoplasms*, Chapter 9 *Diseases of the circulatory system*, and Chapter 11 *Diseases of the digestive system* compared to others such as Chapter 8 *Diseases of the ear and mastoid*, Chapter 16 *Certain conditions originating in the perinatal period* or Chapter 17 *Congenital malformations, deformations and chromosomal abnormalities* indicating that some health conditions are more likely to require treatment and management as an inpatient; this might also indicate that conditions (such as those from 8, 16, or 17) are more frequently managed by non-acute, non-admitted care services
- The counts of patient cases are relatively stable over a five year period, and this is attributable to the stable inpatient operating environment, the use of static reporting categories (ICD-10-AM products, which have been revised only twice during this five year period), persistent and pervasive data specifications (from NHISSC and AIHW) and routine collection and reporting protocols by the health information management workforce.

The aggregated counts for patient admissions attributed to conditions coded to Chapter 21 *Factors influencing health status and contact with health services* are interesting, and the disproportionate percentage of patient cases requires explanation.

The conditions reported in this category are mostly procedures (or similar therapeutic interventions). These are present in ICD-10-AM in order to record a hospital admission for procedural purposes, and to avoid attributing a disease concept to patients who only visit a hospital for treatment purposes. In the main, these might be regarded as 'reason for admission' concepts.

There are a large number of patients who are frequently admitted to hospital, mostly for day-stays. If the underlying conditions or disorders that they have were recorded and reported instead, the statistics for cancer and renal failure (for instance) would be enormously inflated. This is because patients who have certain cancers routinely have 12 x 1 day-stay admissions for a 12 week cycle of chemotherapy; they don't have cancer 12 times, they are diagnosed once. The same scenario applies for dialysis patients, they really only have one renal condition, but visit weekly for dialysis, they don't have 52 renal conditions.

That is, the inclusion of procedural terms in ICD provides a method for counting episodes of care, while not spoiling the accuracy of disease counts. Isolating these terms in Chapter 21 enables an immediate recognition that "... these ones are not like the other ones..."

The inclusion of procedural concepts in ICD and hence the APNMDS reports reveals a scope gap or a mismatch in the Problem Diagnosis RefSet content, which does not have SNOMED CT concepts drawn from the Procedure hierarchy.

NEHTA may optionally choose to include relevant SNOMED CT concepts from the Procedure hierarchy using the same technique that incorporates concepts from the Event or Situation with Explicit Context hierarchies (extensional inclusion). This is discussed further in Section 7.2 where our approach has taken account of these Procedures and Interventions.

6 Major challenge and opportunity

SNOMED CT has no intrinsic notion of "level" or granularity associated with the IS-A hierarchy, which is in contrast to ICD, which has a formal notion of "level". Given this and the structural and scope differences (outlined in Sections 4.1, 4.3 and 5.1), it will always be necessary to develop a hybrid approach to aggregation of SNOMED CT for secondary data purposes.

In order to provide accurate cumulative counts of patient cases, we need to ensure that the hierarchical structure of SNOMED CT does not lead to any degradation of the true count of patient cases, while at the same time addressing requirements for aggregating in fairly homogenous categories.

This is a challenge if we use the native structure of SNOMED CT because it defines clinical conditions using description logic axioms, and these allow dual or multi parenting. Direct native use of SNOMED CT for statistical and reporting purposes would inflate, deflate or otherwise 'disguise' an accurate count of patient cases.

Early SNOMED CT adopters have tried to replicate traditional ICD and APNMDS data management, retrieval and reporting by using SNOMED CT hierarchy and subsumption, and they found that

- i. it didn't work reliably and results were not comparable with existing report protocols
- ii. they needed to make arbitrary decisions about 'categorisation' (what are reasonable and sensible categories in SNOMED CT)
- iii. they needed to make arbitrary decisions about 'when' to stop aggregating patient case counts (how to stop the 'counting by subsumption' or inheritance) before reaching ancestor concepts such as 118234003 *Finding by site*)
- iv. maps between SNOMED CT and ICD were also inadequate.

6.1 Over-counting, under-counting

This phenomena is explained more fully in Section A, and in Figures 3 and 4.

Briefly, the poly-hierarchical (multi-parented) structure of SNOMED CT allows multiple aggregation paths that might be used to roll-up patient encoded data and provide cumulative case counts.

The multi-path structure of SNOMED CT means that different users can, and probably will, choose a different aggregation strategy. Some of those approaches will result in inflated, deflated or missing patient case counts.

6.2 Reporting by specification

Constructing secondary reports via a query specification leads to similar outcomes. If users want to report on patient data, they pretty much have to be fully aware of what SNOMED CT encoded data is contained within the collection in order to construct a query that adequately accounts for true negatives.

If, for example, the query specification resembled:

Find and return all cases encoded with 6142004 *Influenza* and any descendant, then report as "influenza count"

The report would not include patient cases assigned 95891005 *Influenza-like illness*, nor would it include any patient case assigned with an organism identifier (as a proxy).

Secondary reports constructed by query specifications are useful for retrospective studies, and when the use case tolerates a dynamic, non-reproducible approach.

6.3 Report generation using maps between SNOMED CT and ICD

Other secondary data users, notably the Independent Hospital Pricing Authority (IHPA) has attempted to use maps between SNOMED CT and ICD-10-AM to manage data collections for Activity Based Funding (ABF) purposes. This approach has proven to be unreliable. Any SNOMED CT concept that does not have an explicit map to ICD-10-AM is excluded from ABF data protocols, and the patient case does not qualify for resource funding as would be expected. Work-arounds are required to ensure that excluded patient cases can be resourced.

The lack of synchronisation between versions and map maintenance has been shown to be problematic as well, and this means that continued investment in map building and revision will be necessary, across versions, and as ABF protocols evolve.

We investigated the potential use of SNOMED CT to ICD-10 maps to determine whether the maps, available nationally in ICD-10-AM (EDRS, CSIRO) or internationally in ICD-10 (IHTSDO, NHS) would serve aggregation of the Problem Diagnosis RefSet for APNMDS purposes.

There are 31 041 concepts in the Problem Diagnosis RefSet that are not mapped to ICD-10 at all (by AU, UK or IHTSDO). Of these 30 427 are drawn from the Clinical Finding hierarchy, 107 from the Event hierarchy, 5 from Observable entities and 21 from the Situation with explicit context hierarchy.

Together, this means that almost one third of the Problem Diagnosis RefSet is not officially mapped, and patient cases encoded with the Problem Diagnosis RefSet would not be counted in an aggregation technique that relied upon maps.

Of course, there may be enthusiasts who would be willing to construct an additional 30 000 maps. This will be costly.

These maps would then have to be maintained, curated for quality, revised each SNOMED CT release cycle and synchronised with ICD-10-AM releases; cost unknown and not predictable.

It is not yet known whether the planned release of the 9th edition of ICD-10-AM will influence utility of existing maps between the Problem Diagnosis RefSet content and ICD-10-AM (~60 000).

6.4 Proposal for hybrid approach

Essentially we propose a hybrid approach²⁰ where we rely on the description logic and inheritance structure of SNOMED CT to define a primary aggregation path, while also relying on APNMDS and ICD-10-AM specifications to define the aggregation stopping points and report categories (See Figure 5, Section A for a larger rendition)

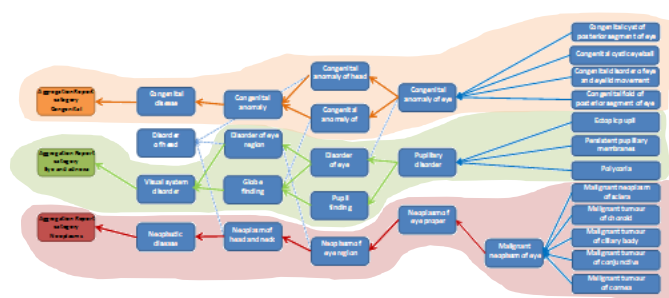


Figure 10: Aggregation primary path, after pruning; mutually exclusive and sensitive tree structure

²⁰ It should be noted that the hybrid approach examined here is only one technique. Other methods are being developed and investigated by AEHRC.

Importantly, any methods we develop and recommend must be reproducible by automated approaches. Any hand-crafting that relies on human manipulation and judgment will not provide stable and comparable results that adequately match static requirements. Our preference is also to resist mapping strategies wherever possible as these have already proven to be high cost, low value, tedious and error prone. Human judgments and maps will diverge and change over time, and this puts at risk the data stability and comparability needed for annual, bi-ennial or decadal reports.

This hybrid approach means that:

- The Problem Diagnosis RefSet will be exhaustively included
- The native SNOMED CT IS-A hierarchy will serve as a basis for a solution, but is not a solution in itself because the directed acyclic graph structure (and its description logic) is not mutually exclusive and will disperse patient case counts
- Some native SNOMED CT IS-A relationships will be favoured while others are ignored; when we ignore certain IS-A relationships we call this “pruning”; Pruning relationships in SNOMED CT transforms the structure into a simple hierarchical classification
- New namespace concepts (referred to as *Aggregation concepts*, similar to Navigational or Special concepts in SNOMED CT) may need to be introduced to appropriately label APNMDS categories and to allow a ‘container’ for those ICD-10-AM residual concepts (ie: “Other specified” or “Other”).
- While we will not define or specify a primary path for aggregation between existing concepts that does not currently exist as a SNOMED CT IS-A relationship, there will be a modest number of new *aggregation relationships* where these are required as link assertions with (*namespace*) *Aggregation concepts*.
- There will need to be explicit and measurable criteria for making the choice between which IS-A relationships are favoured as steps along a primary path, and which are ignored so that the technique is reproducible by programming and not reliant on hand-crafting
- The result will form a different, separate aggregation table derived from the original SNOMED CT relationships table. These aggregation relationships should *not* be thought of as representing the IS-A relationship (and not the same as and no longer related to the IS-A relationships).
- Secondary data users who are interested in aggregation (and who do not require specificity, description logic or ontological features) can elect to use the aggregation table instead of the standard SNOMED CT relationships table (as released).

7 Aggregation technique

This proposed approach for secondary data aggregation purposes involves the non-arbitrary specification of an alternate, prioritised, pathway through the existing IS-A relationships, to utilise those portions of the SNOMED CT hierarchy - preferentially - to define a single, non-duplicating inheritance structure.

This specified pathway takes the form of an aggregation RefSet where each RefSet concept member (source) is annotated with a single designated parent (target) that best serves as its class or category for case counting purposes. It is also required that the IS-A relationship holds (directly or transitively) between each source and target concept pair, except in the special case where either of these concepts is an aggregation concept (see Hybrid approach Section 6.4). The standard international or AU released relationships table (a) is not replaced by this aggregation table (b). Rather, secondary data managers will substitute (b) for (a) only when statistical aggregation tasks dictate.

7.1 Evaluating an Aggregation RefSet

The criteria against which an aggregation RefSet can be evaluated are shown in Table 3.

Table 3: Evaluation criteria for Problem Diagnosis aggregation RefSet

Criteria	Justification	Assessment method
<i>What we need</i>	<i>Why we need it</i>	<i>How we'll measure if we got it or not (quantifiable)</i>
Small number of mutually exclusive top level categories (disjoint)	The highest level categories with cumulative patient case counts should be able to be easily displayed on a one-page 'frequency report', bar chart or a table, suited to high level policy, strategy or epidemiological snapshot reports.	Less than 50 SNOMED CT high level concepts that act as report category labels and ancestors of the content of the Problem Diagnosis RefSet
Single primary inheritance path for aggregation purposes, no dual parenting	It is necessary to remove any native SNOMED CT IS-A relationship that would distribute or disseminate the count of patient cases across multiple parent concepts within the RefSet.	The number of relationships in the aggregation RefSet should approximate the number of concepts in the Problem Diagnosis RefSet (ie: nodes and edges should be roughly equal in number). If the number of edges exceeds the number of nodes in the aggregation RefSet then it is possible that dual or multi-parenting exists (and the aggregation methodology will have failed).
Minimise orphan concepts	Ensure that any Problem Diagnosis RefSet concept aggregates to a meaningful disjoint top level category. This might not be possible given the difference in scope and the removal of some IS-A relationships.	A residual ("other specified") to aggregate concepts that may or may not be relevant to the admitted patient sector but that can be used when required, recognised in exchange scenarios and interrogated or retrieved separately
Multiple levels of granularity	Arrangement of Problem Diagnosis RefSet members in discrete levels to allow discretionary and dynamic drill-down and roll-up options for different secondary data users or purposes.	Between 2 and 23 levels based on native IS-A steps between concepts. These levels are somewhat arbitrary (within SNOMED CT, but secondary data users may find them helpful for drill down and roll up purposes.
Maximise automation of aggregation RefSet development	A hand-crafted or map-based approach to aggregation will result in continuous hand-crafted maintenance for each release cycle; this is onerous, non-reproducible, error-prone and unsustainable. Some hand-crafting will always be necessary to accommodate orphans (see above).	Maximal automated techniques need to be developed using native SNOMED CT description logic, analytics and transitive closure reductions to provide ongoing development, update and release agility. Reliance on hand-crafting techniques should be less than 10% of RefSet volume.
Inclusion of all patient cases (sensitivity)	Every patient case recorded in the original data collection must be included in any aggregated report (somewhere).	The total number of patient cases attributed to concepts in the aggregation RefSet must equal the number of patient cases in the original (real world, relevant) data collection.

The above criteria will ensure the development of an aggregation layer that will enable secondary data users to achieve comparative results from the data. The reporting form and format will have a similar look and feel to that previously achieved with the added benefit of being able to drill down and report at finer detail if required.

7.2 Methods for content management

The APNMDS specification, as it stands, already stipulates the top level categories that are preferred for national, minimalist aggregated reporting purposes.

That is, we know from the outset the categories we hope to construct from the Problem Diagnosis RefSet content so that our aggregation technique mimics these secondary data and reporting requirements.

Despite our resistance, mapping still forms the best option to align SNOMED CT and ICD-10-AM concepts. But here we map only the top level concepts between each instrument, and map in a backward direction. This is a best match, and a categorical map as distinct from a ‘meaning’ map. It is not an exhaustive equivalence mapping strategy. Table 4 shows the top level category alignment.

Although there are only 20 top level aggregation categories specified by the APNMDS reports using ICD-10-AM, we see that SNOMED CT is much more specific and hence there are 62 SNOMED CT concepts that provide meaningful alignment. The specificity ratio is 1:N. Note also the modesty of the map. Based on Rogers’ experiments, we were aware that ‘over-mapping’ would introduce scope-creep issues. Care needs to be taken to select a SNOMED CT concept that does not inherit child concepts that fall outside the ICD-10-AM scope.

No further mapping was undertaken. No additional maps are required to construct the primary path that will serve aggregation.

7.2.1 Additional content requirements to serve this use case

Note that we began by mapping between ICD-10-AM and SNOMED CT, and we did not constrain that initial high level to Problem Diagnosis RefSet content. This is because we knew in advance that the APNMDS use case and routine reports included Procedures, and that the Problem Diagnosis RefSet did not. There was a known mis-match in the scope (See Section 5.1 above). The blue highlighted rows in Table 4 show the additional SNOMED CT content, not currently included in the Problem Diagnosis RefSet, that will be required if the aggregation technique we construct is to faithfully mimic the existing reporting protocols. We include these Procedural items here to demonstrate how the Problem Diagnosis RefSet might be expanded to better serve the APNMDS use case. We refer to these in a category called “Other specified”.

7.2.2 Partition of findings and disorders

Because the Problem Diagnosis RefSet is intended to be broadly used in various health information systems and care settings, it will be the case that the RefSet content will be available for clinicians to select and enter into patient records. It is unlikely that users will, or will need to, make distinctions between what SNOMED CT deems to be a Finding or a Disorder. Because the Problem Diagnosis RefSet contains Findings and Disorders, our methods deal with both.

However, the APNMDS use case, reliant as it is on ICD-10-AM and admitted patient episode records, favours Diagnoses (disorders). ICD-10-AM coding conventions stipulate that what is recorded as the principal diagnosis is the condition that is found ‘after study’ to be the most acute disease justifying the admission. This generally means that problems, symptoms and reasons for encounter do not qualify for assignment as a principal diagnosis.

Because the goal is to serve this use case, our techniques and decision criteria also favour Disorders. But Findings are still included in the aggregation. We have developed a partition so that secondary users can – if they require – distinguish between the two, either in their post hoc analyses or in standard reports. Alternatively, the inclusion of Findings in the aggregation RefSet provides secondary users with extra visibility of patient conditions that might not otherwise be revealed in ICD-10-AM data collections.

The green highlighted rows in Table 4 provide a single example that demonstrates that for each ICD-10-AM chapter we have included a SNOMED CT concept from both types, Findings and Disorders (where relevant).

Table 4: Top level alignment between SNOMED CT, ICD-10-AM chapters and APNMDS reporting categories

ICD	Chapter name	ID	Fully Specified Name
1	Certain infectious and parasitic diseases	40733004	Infection
2	Neoplasms	55342001	Neoplastic disease
3	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	414030009	Disorder of immune structure
3	Diseases of the blood and blood forming organs and certain disorders involving the immune mechanism	299691001	Finding of blood, lymphatics and immune system
3	Diseases of the blood and blood forming organs and certain disorders involving the immune mechanism	414022008	Blood disease
4	Endocrine, Nutritional and Metabolic Diseases	300893006	Nutritional finding
4	Endocrine, Nutritional and Metabolic Diseases	75934005	Metabolic disease
4	Endocrine, Nutritional and Metabolic Diseases	106176003	Endocrine finding
4	Endocrine, nutritional and metabolic diseases	362969004	Disorder of endocrine system
4	Endocrine, nutritional and metabolic diseases	2492009	Malnutrition
4	Endocrine, nutritional and metabolic diseases	106089007	Metabolic finding
5	Mental and behavioural Disorders	74732009	Mental illness
5	Mental and behavioural disorders	116367006	Psychological finding
6	Diseases of the nervous system	118940003	Neurological disorder
7	Diseases of the eye and adnexa	128127008	Visual system disorder
8	Diseases of the ear and mastoid process	118236001	Ear and auditory finding
8	Diseases of the ear and mastoid process	362966006	Disorder of auditory system
9	Diseases of the circulatory system	106063007	Cardiovascular finding
9	Diseases of the circulatory system	362971004	Disorder of lymphatic system
9	Diseases of the circulatory system	49601007	Cardiovascular disease
10	Diseases of the respiratory system	50043002	Respiratory disease
10	Diseases of the respiratory system	106048009	Respiratory finding
11	Diseases of the Digestive System	386617003	Digestive system finding
11	Diseases of the digestive system	53619000	Disorder of digestive system
12	Diseases of the skin and subcutaneous tissue	106077005	Integumentary system finding
12	Diseases of the skin and subcutaneous tissue	128598002	Disorder of integument
13	Diseases of the musculoskeletal system and connective tissue	928000	Musculoskeletal disorder
13	Diseases of the musculoskeletal system and connective tissue	106028002	Musculoskeletal finding
14	Diseases of the genitourinary System	118238000	Urogenital finding
14	Diseases of the genitourinary system	42030000	Disorder of the genitourinary system
15	Pregnancy, childbirth and the puerperium	173300003	Disorder of pregnancy
15	Pregnancy, childbirth and the puerperium	415073005	Perinatal disorder
15	Pregnancy, childbirth and the puerperium	118185001	Finding related to pregnancy
17	Congenital malformations, deformations and chromosomal abnormalities	417893002	Deformity
17	Congenital malformations, deformations and chromosomal abnormalities	409709004	Chromosomal disorder
17	Congenital malformations, deformations and chromosomal abnormalities	66091009	Congenital disease
18	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	365860008	General clinical state finding
18	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	250171008	Clinical history and observation findings
18	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	441742003	Evaluation finding
18	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	102957003	Neurological finding
19	Injury, poisoning and certain other consequences of external causes	75478009	Poisoning
19	Injury, poisoning and certain other consequences of external causes	419026008	Effect of exposure to physical force
19	Injury, poisoning and certain other consequences of external causes	417163006	Traumatic AND/OR non-traumatic injury
20	External causes of morbidity and mortality	409544000	Immediately dangerous to life and health condition
20	External causes of morbidity and mortality	418420002	Intentionally harming self
21	Factors influencing health status and contact with health services	367336001	Chemotherapy
21	Factors influencing health status and contact with health services	52052004	Rehabilitation therapy
21	Factors influencing health status and contact with health services	385785003	Chemotherapy care assessment
21	Factors influencing health status and contact with health services	399141000	Male sterilisation
21	Factors influencing health status and contact with health services	417928002	Abuse
21	Factors influencing health status and contact with health services	385971003	Dialysis care
21	Factors influencing health status and contact with health services	183505002	Non-urgent plastic surgery admission
21	Factors influencing health status and contact with health services	390906007	Follow-up encounter
21	Factors influencing health status and contact with health services	385788001	Chemotherapy care management
21	Factors influencing health status and contact with health services	386493006	Venous access device maintenance
21	Factors influencing health status and contact with health services	385786002	Chemotherapy care
21	Factors influencing health status and contact with health services	418799008	Finding reported by subject or history provider
21	Factors influencing health status and contact with health services	418715001	Exposure to potentially harmful entity
21	Factors influencing health status and contact with health services	52637005	In vitro fertilisation
21	Factors influencing health status and contact with health services	312851005	Screening for disorder
21	Factors influencing health status and contact with health services	108241001	Dialysis procedure
21	Factors influencing health status and contact with health services	60890002	Female sterilisation

7.2.3 Dealing with Problem - Diagnosis RefSet residuals

Again, the content and intended use for Problem-Diagnosis RefSet implementation is broader than this single APNMDS use case. It contains concepts describing health conditions and issues that will never be seen in an inpatient care setting. This means that these concepts will not aggregate to any of the nominated top level aggregation categories that we target; because they do not have a place in the APNMDS use case, they do not find a place in the APNMDS aggregation. We treat these as ‘residuals’ and direct their aggregation path to a synthetic category called “Other”. These concepts may be more likely to be used in community, social, residential aged care, sub-acute care sectors, or more relevant to GP (ICPC)

or laboratory use cases (LOINC), but should be viewable, selectable and exchangeable between all health care sectors.

7.2.4 Editorial considerations

It is apparent from Table 2 that the APNMDS use case routinely publishes the names of their report categories, and (as in Table 4) these names largely reflect ICD-10-AM chapter titles.

SNOMED CT authors have long resisted attempts to model and include ICD-10-AM names, especially categorical names, NEC and NOS concepts, because statistical conventions are not compatible with description logic axioms.

We are similarly reluctant to introduce new ICD-like aggregation concepts into SNOMED CT (international) itself. Instead we suggest that new concepts are authored, specifically designed for aggregation, with a NEHTA namespace and for SNOMED CT-AU releases only; we call these aggregation concepts.

These could be descendants of a new *|aggregation concept|* which itself would be a child of *|special concept|* or indeed *|navigational concept|*. Alternatively, these synthetic categorical concepts may be quarantined in metadata, or an aggregation RefSet.

These aggregation concepts would need to be ‘coupled’ with their Problem-Diagnosis Aggregation RefSet top level concepts in order to direct cumulative counts to the correct aggregation category. This will require a different sort of link assertion since they represent an aggregation relationship rather than IS-A semantics and thus the existing SNOMED CT IS-A relationship cannot be used. We call these aggregation relationships; they define the primary aggregation path.

We have not pre-empted those editorial decisions here. Rather we have provided a light-weight rendition of what these aggregation concepts and link assertions might look like, and how they might serve the APNMDS use case.

7.3 Methods for technical management

The techniques we use to determine the primary aggregation path are, for the most part, automated. They rely on some recognised, and some new and developing, SNOMED CT query languages. This approach means that the vast majority of aggregation RefSet construction and quality assurance can be run and re-run in an efficient and reproducible manner. Refinements can also be undertaken using this same approach.

Figure 11 shows how we approach RefSet construction through analysis, extension and relationship processing to serve a specified secondary data use. This is a generic description of the analytic workflow, and only covers the major steps in the technique; it is not specific to the Problem Diagnosis RefSet or the APNMDS use case. Other AEHRC clients are also using this same (generic) approach for their own different use cases.

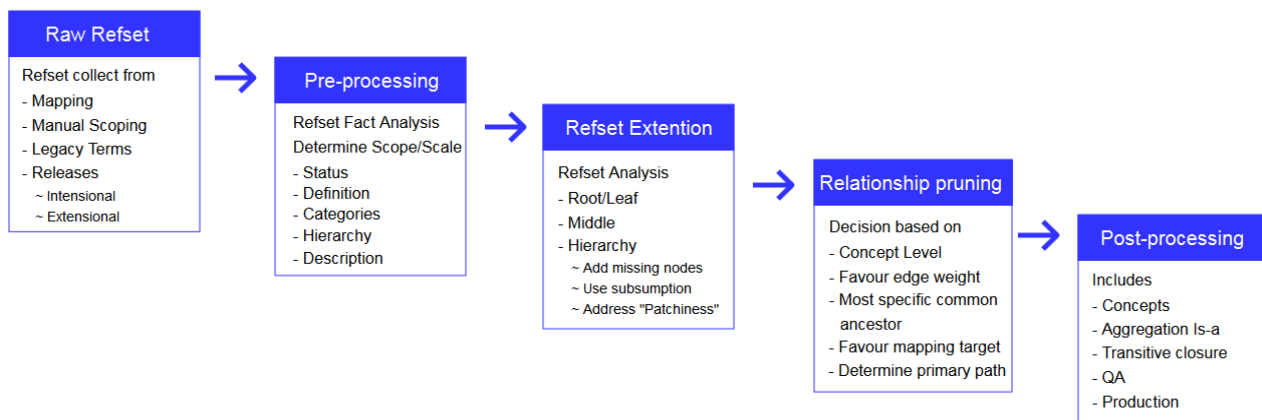


Figure 11: Generic workflow and methodology

Of course, there is some re-iteration and recursiveness in the processes, and some human review and intervention along the way.

Our experience is that the more dissimilar the scope and use cases, the more human judgement and intervention is necessary to resolve internal terminology and classification ‘collisions,’ in either meaning or categorisation. Nonetheless, any such judgments can be collated and referenced through subsequent queries and iterations, and so lend themselves to an almost entirely automated approach.

We use a variety of metrics that function as ‘decision criteria’ in order to identify terminology content, and select the single, preferred (native) IS-A relationship that will serve as a primary aggregation path.

Some of the metrics, techniques and decision criteria include:

- Scope analysis, including overlaps, gaps, required partitions
- The limited use of initial maps to nominate a starter (top level) aggregation set of targets (as shown in Table 4)
- Counts of child concepts, where greater number of child concepts is favoured
- Counts of parent and ancestor concepts where fewest number of ancestors is favoured
- Favour edge weight
- Most specific common ancestor algorithms (weighted for either specificity or commonality)
- Breadth first search algorithms
- Distance measures, noting that distance in SNOMED CT is somewhat arbitrary
- Iterate from top down and bottom up, triangulate the middle
- Level or layer analysis to ensure drill-down, roll up options
- Transitive closure and transitive closure reductions
- Analysis of orphans, short or small paths
- Where conflicted choices (multi-paths) remain, favour use case specifications (where available)
- Where conflicted choices (multi-paths) remain, favour mapped concepts (where available)
- Where conflicted choice (multi-paths) remain, favour frequency of use data (where available)
- Review and resolution of any gaps, conflicts, dual paths that remain (often forced choice human “decision by exclusion”)

For this investigation, it was necessary to use these decision criteria adaptively, and iteratively so that it would eventually:

- Be exhaustive of all Problem Diagnosis RefSet content
- Adequately mimic the APNMDS use case
- Provide a single primary aggregation path that was reproducible (efficiently), and simple to use

From the initial analyses we were aware that there were three different components required to build this aggregation RefSet.

This meant that our technical approaches were performed (at least) three times, separately for each component, because each had different features that had to be addressed:

1. Concepts that were **not in** Problem Diagnosis RefSet but were required for APNMDS use case purposes (“Other specified”)
2. Maximising the Problem Diagnosis RefSet content as participating concepts in the aggregation RefSet (“APNMDS reporting categories”)
3. Concepts that were **in** Problem Diagnosis RefSet, but **not in** APNMDS use case (“Other”)

8 Results

For convenience, we present results (at least initially) by dealing with each of these components separately. The number of concepts involved means that a detailed view of the whole is difficult to present visually with any clarity. A number of different visualisations are provided, and it is worth noting that some of them re-represent results in a different fashion, and not all are directly comparable. Some are extracts of a small number of concepts (to reveal some fine grained details), others are ‘aggregations of aggregations’ (to provide the bird’s-eye view). The actual data files provided should be regarded as the ‘source of truth’.

Overall results obtained from these content and technical management approaches are summarised, diagrammatically, in Figure 12.

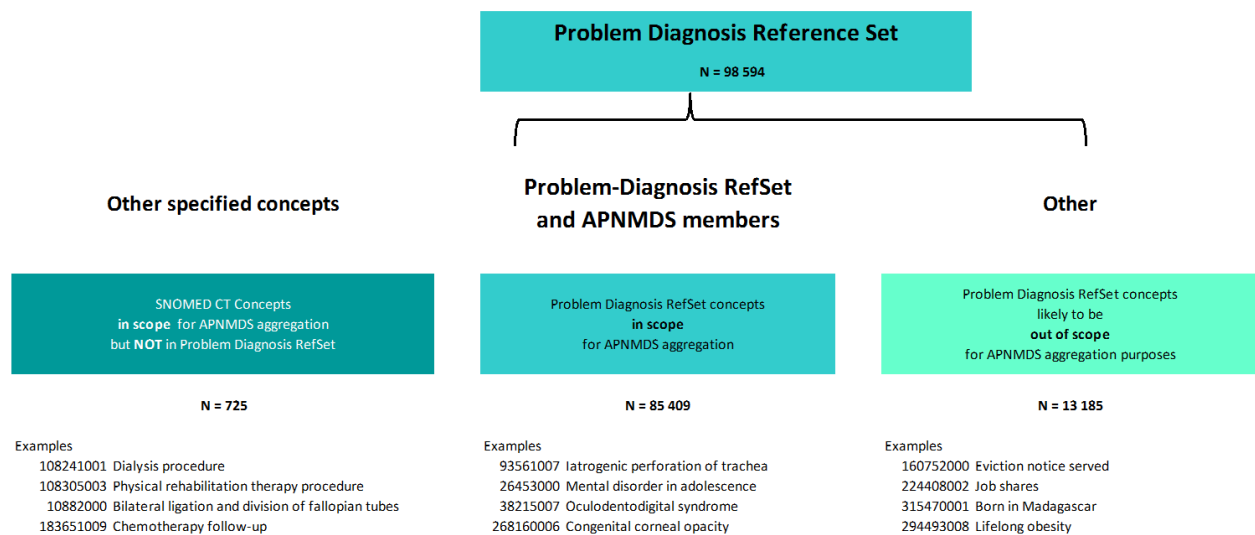


Figure 12: Summary of aggregations achieved

8.1 Other specified

These are the procedural and intervention concepts that APNMDS requires. They were identified via the initial top level category mapping using SNOMED CT (not the Problem Diagnosis RefSet), and then the algorithmic identification of relevant descendant concepts and their primary aggregation pathway.

Number of APNMDS top level categories	1
Number of Problem Diagnosis concepts, top aggregation level	19
Number of layers (drill down, roll up)	7
Number of concepts (nodes)	725
Number of aggregations relationships (edges)	706
Number of SNOMED CT (native) IS-A relationships	749
Number of pruned (SNOMED CT native) IS-A relationships	43

These 725 concepts can be optionally included in the Problem Diagnosis RefSet (future releases) to increase the ability of this RefSet to serve secondary data uses. Figure 13 is a visualisation of this portion of the Aggregation RefSet. It shows 725 nodes (concepts) 706 edges (aggregation relationships), 19 top level categories, and 7 layers, that have drill down or roll up functionality.

These 19 top level Problem Diagnosis concepts all aggregate to a single APNMDS reporting category of (Chapter 21) *Factors influencing health status and contact with health services*.

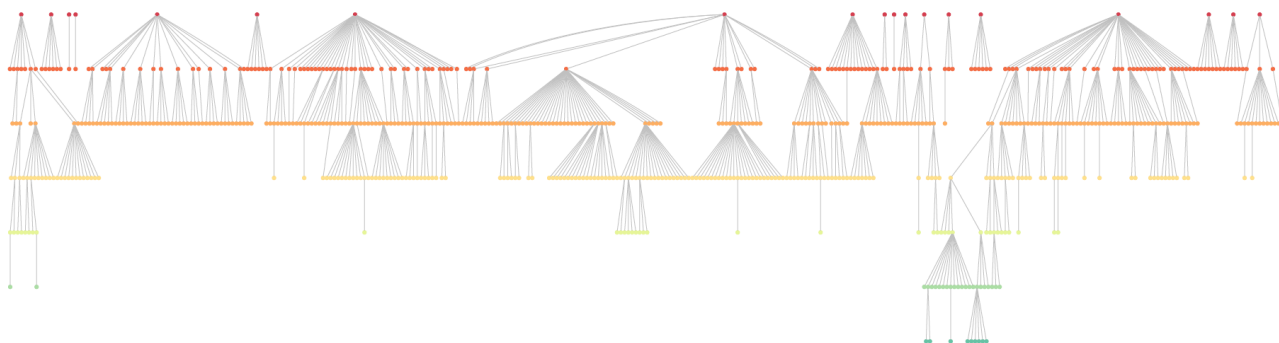


Figure 133: Other specified concepts required for APNMDS reporting (aggregation for complete sub-hierarchy shown)

Figure 14 shows a different representation of this same aggregation approach. Here, each of the 19 top level concepts and their descendant concepts are shown in distinct lines (from top to bottom); this is an aggregation of an aggregation.

We call this linearisation, reflecting the WHO ICD-11 description. This rendition will be more familiar to secondary data users who are most familiar with ICD encoded data which takes the form of disjoint, mono-hierarchies as depicted here. Seven (7) aggregation layers are provided, compared to the traditional 3 layers native to ICD-10-AM. The size of each bubble represents the number of concepts contained in each layer of the aggregation.

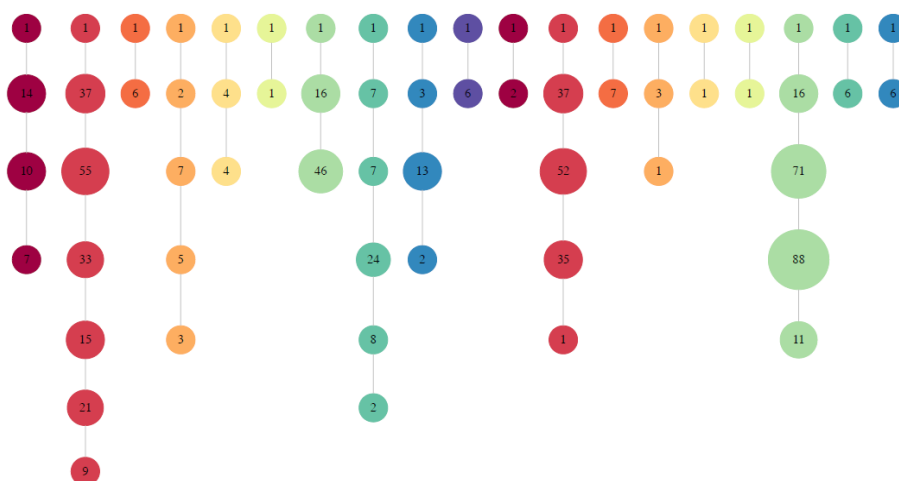


Figure 14: Linearisation of “Other specified” concepts required for APNMDS reporting

8.2 Problem-Diagnosis RefSet members and APNMDS report categories

These are the 85 409 concepts, members of the current Problem Diagnosis RefSet that are in scope and of interest to APNMDS reports. They form the vast majority of aggregated Problem Diagnosis concept members.

Number of APNMDS top level categories	20
Number of Problem Diagnosis concepts, top aggregation level	48
Number of layers (drill down, roll up)	15
Number of concepts (nodes)	85409
Number of aggregations relationships (edges)	85361
Number of SNOMED CT (native) IS-A relationships	135076
Number of pruned (SNOMED CT native) IS-A relationships	50714

The primary aggregation pathway leads to 48 top level concepts, reducing the originally mapped top level categories from 62. This aggregation pathway is mutually exclusive and sensitive to the statistical requirements for valid representation of patient cases where the patient was admitted to hospital.

Figure 15 shows a visualisation of this portion of the aggregation. This shows the vast majority of Problem Diagnosis RefSet content and content that is in scope for current APNMDS reporting purposes according to the current specifications of the APNMDS use case and the distribution of concepts among the disjoint reporting (mono) hierarchies.

Because it contains 85409 concepts, it is not suited to display the entire aggregation hierarchy (it will not fit on a single page); instead we provide the linearised rendition. The size of each bubble represents the number of concepts included in each layer, per disjoint mono-hierarchy. The full aggregated content, both concepts and aggregation primary pathway is provided in the files at Appendix A.

These top categories may be further aggregated, if required, to more accurately mimic the APNMDS traditional reporting categories, reducing the top level shown here from 48 to 20.

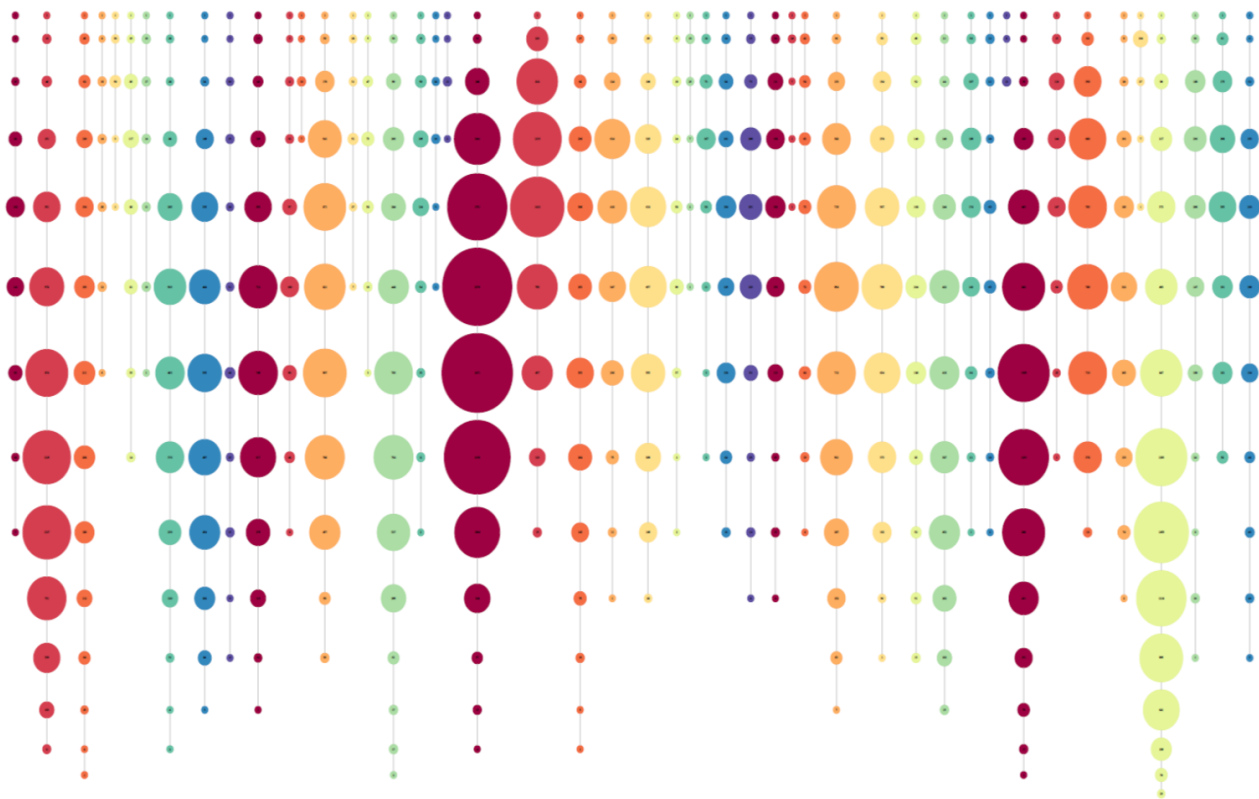


Figure 15: Linearisation of Problem Diagnosis RefSet aggregation for APNMDS reports

8.3 Other

These conditions and service encounters are outside of the current specification of APNMDS. However, they may still be of interest to secondary data users as a way of viewing and understanding which patient conditions are documented but not routinely reported under current protocols.

Number of APNMDS top level categories	nil
Number of Problem Diagnosis concepts, top aggregation level	20
Number of layers (drill down, roll up)	13
Number of concepts (nodes)	13185
Number of aggregation relationships (edges)	13165
Number of SNOMED CT (native) IS-A relationships	17707
Number of pruned (SNOMED CT native) IS-A relationships	4541

That is, these concepts may be used to document admitted patient cases, but both ICD and APNMDS neglect to encode or collect such concepts. Hence they have also been aggregated, but isolated, from the core APNMDS reporting categories. It remains possible that secondary data users might – in the future – prefer to include some of these concepts in routine reports, but this will require a revision of the current APNMDS specification. This is an option for NHISSC, DoH and AIHW consideration.

Some of these concepts may not be relevant for admitted patients, instead reflecting health conditions that are most likely to be treated in a non-hospital care setting. This would mean that some Problem Diagnosis RefSet members won't ever be relevant to or included in APNMDS report aggregations.

It is also the case that a few of these 13185 concepts reflect a collision between the medical record information model and the terminology model provided by the Problem Diagnosis RefSet. For example, this portion of the aggregation RefSet contains concepts related to *Death*. The APNMDS data specifications contain a separate pre-defined value set for recording death (Meteor data element id 270094). Therefore the inclusion of a terminology concept here is redundant for APNMDS purposes; this information is sourced from outside the terminology.

Figure 16 shows a visualisation of the aggregated content of the remaining concepts from the Problem Diagnosis RefSet. These concepts are quarantined from the majority of Problem Diagnosis aggregation because they are not required for APNMDS reporting under the current specifications defined by NHISCC, DoH or AIHW (See also Section 8, dot point 3 above).

Because it contains over 13 000 concepts, we provide the linearised rendition. The size of each bubble represents the number of concepts included in each layer, per disjoint mono-hierarchy. The full aggregated content, both concepts and aggregation primary pathway is provided in the files at Appendix A.

These top categories may be further aggregated, if required. This could be achieved if NEHTA chose to author an additional Metadata concept along with the necessary aggregation relationships (SNOMED CT-AU extensions) to provide a single synthetic high level grouper concept.

If, in the future, NHISCC, DoH or AIHW elect to include these patient conditions in routine APNMDS report protocols, the aggregation technique can be refined and re-run to pull these concepts out of the non-APNMDS quarantine aggregation and include them in the main APNMDS aggregation files. This will depend entirely upon re-specification of data management and report requirements by secondary data custodians.

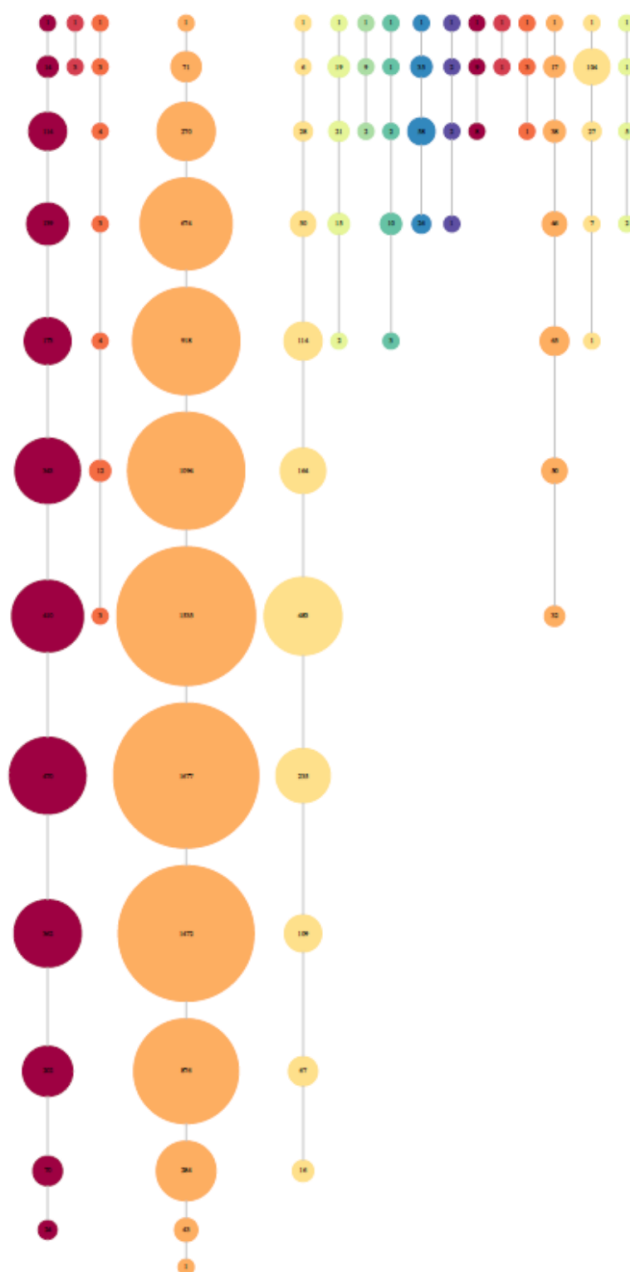


Figure 16: Linearisation of "Other" non APNMDS concepts

8.4 Overall

We now place these three separate components back into a cohesive whole, representing the aggregation of the entire Problem Diagnosis RefSet content.

Figure 17 shows the overall aggregation of SNOMED CT content re-purposed for APNMDS reporting purposes (see files: aggregation concepts nodes, aggregation links edges)

Only the SNOMED CT top level concepts in **black** bubbles, (from all components; 19 tops from “Other specified” 48 tops from APNMDS report categories and 20 tops from “Other”). No other Problem Diagnosis RefSet content is shown here, it is too big to display.

NEHTA and secondary data stakeholders may choose to author some additional aggregation concepts along with the necessary aggregation relationships (SNOMED CT-AU extensions) to provide even closer alignment with the current APNMDS reporting categories. Such synthetic namespace concepts are shown with **green** bubbles.

Note that “Other specified” (left hand side) does have an existing APNMDS category name. The **blue** bubbles signify that all concepts within this sub hierarchy are currently not members of the Problem Diagnosis RefSet, and NEHTA would (optionally) include these. The **pink** bubbles (and those that subsume) represent those Problem Diagnosis concepts that are currently out of scope for the APNMDS use case.

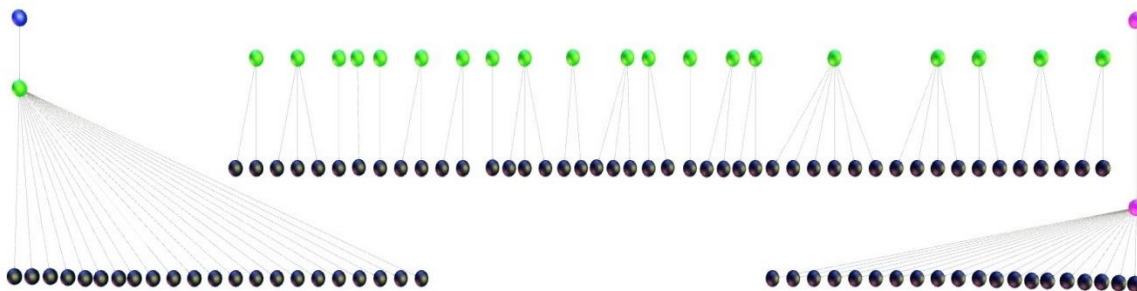


Figure 17: Overall, Problem Diagnosis RefSet tops to APNMDS, highest level aggregation shown

8.5 Further detailed examples

While these high level visualisations are potentially helpful, they disguise some of the fine grained detail of the aggregation RefSet outcomes. The following examples ‘unpack’ some of that hidden information, and reveal some features of aggregation that secondary data users may want to consider. Figure 18 shows an example of the primary path along which patient case counts would aggregate, without over or under counting, for a single concept, and Table 5 shows the same terminology content as an aggregation table.

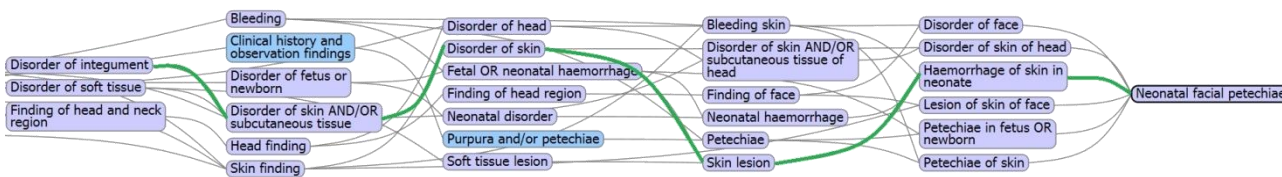


Figure 18: Example of aggregation path for a single concept

Viewed differently, we can see this small portion (single concept) in a slightly broader context, within a tree like structure with other concepts from Problem Diagnosis RefSet.

Below in Figure 19 (left panel), the black bubbles show the direct, primary aggregation path, without the maze of multi-parents for the same concepts as shown in Figure 18. The lowest black bubble is *Neonatal facial petechiae*, and the aggregation hierarchy stops at the highest, most sensible ancestor *Disorder of integument*. Concepts above this highest aggregation category are generally SNOMED CT ‘grouper’ concepts like *Disorder of body system*, *Disorder by body site*. These will be too high, and too heterogeneous to be meaningful for secondary data users. The layout here allows visualisation of the depth and layers (specificity) and the drill down and roll up potential.

The middle panel in Figure 19 shows the partitioning achieved via our aggregation method, where Findings are represented in green bubbles, and Disorders as blue bubbles. This assists secondary data users to distinguish abnormal patient conditions (Disorders, blue) from conditions that may be normal, or abnormal (and potentially not diagnoses, but rather symptoms, problems or reasons for encounter; green).

Note there is one aberration here, where a leaf disorder concept (blue) is directly inherited by a finding concept (marked with ★).

This is unavoidable because it is a modelling problem within SNOMED CT. It is always the case that - by design - conditions that have a suffix tag of (disorder) should **always** be linked via the IS-A hierarchy or transitively to their ancestor of Disease. This concept, 423716004 *Petechiae of skin*, lacks a Disease ancestor and the modelling needs to be fixed in the SNOMED CT release²¹.

The right panel shows the same portion of the aggregation hierarchy with some example patient case data attributed to each concept. Here the size of the bubble represents the (relative) number of patient cases that were assigned to each Problem Diagnosis RefSet concept. Note that the far left bubble represents a very general, high level, concept of representing the condition of 131148009 *Bleeding (finding)*.

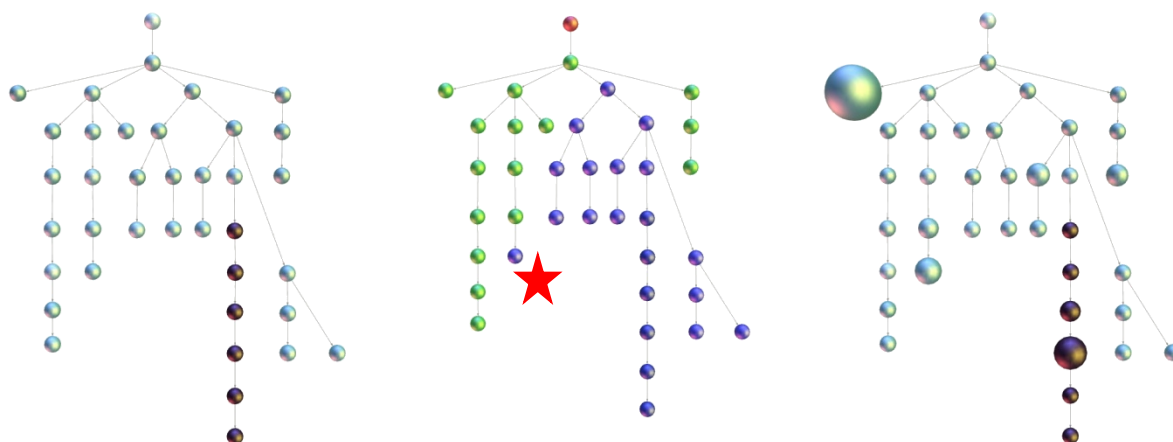


Figure 19: Example portion of aggregation RefSet, showing tree-like structure(left), partition of Disorders and Findings(centre), and patient case attributions (right)

Figure 20 shows another example of the partitioning between Findings and Disorders, without the influence of SNOMED CT modelling flaws. There will be, from time to time and here and there, a number of concepts that do not ‘behave’ as we would expect, and as our techniques are specified.

²¹ We are gratified that our analytic and aggregation technique, which relies on structural features of SNOMED CT, not only produces a valid primary aggregation path for APNMDs purposes, but it is also capable of identifying quality issues within SNOMED CT modelled content.

Any anomalies we have noticed so far have been attributed to peculiarities in the original SNOMED CT content itself.

This also shows that we have favoured concepts that are Disorders and existing IS-A relationships have been preferred as the primary path for aggregation.

That is, within SNOMED CT (and the Problem Diagnosis RefSet) there are concepts that are:

- purely Findings and only ever subsume to Clinical Finding via the IS-A relationships
- purely Disorders until they reach Disease which is a direct child of Clinical Finding
- a mix of both Findings and Disorders that could be aggregated along either route.

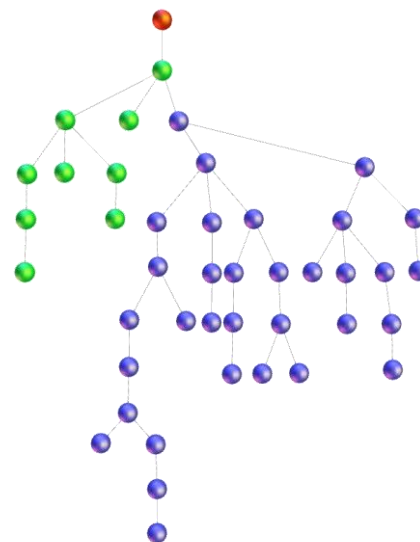


Figure 19: Portion of Problem Diagnosis RefSet aggregation showing Disorder and Finding partitions; without modelling flaws

It is these mixed Findings and Disorder concepts that have been re-routed through the Disorder partition only. That is, where aggregation *could* have occurred along either of the two pathways, we have *preferred* the Disorder path, and *pruned* the Finding path.

This best serves the ICD-10-AM and APNMDS use case, without excluding the meaningful and useful Problem Diagnosis RefSet content that is tagged as a (finding).

Table 5: Example (extract) of aggregation relationship table

Concept ID	Fully Specified Name	Concept ID	Fully Specified Name
406122000	Head finding	118254002	Finding of head and neck region
22925008	Neonatal disorder	414025005	Disorder of fetus or newborn
80659006	Disorder of skin AND/OR subcutaneous tissue	128598002	Disorder of integument
95320005	Disorder of skin	80659006	Disorder of skin AND/OR subcutaneous tissue
248402002	General finding of soft tissue	118234003	Finding by site
106076001	Skin finding	248402002	General finding of soft tissue
431268006	Haemorrhage of skin in neonate	95324001	Skin lesion
95324001	Skin lesion	95320005	Disorder of skin
271813007	Petechiae	423306009	Purpura and/or petechiae
106077005	Integumentary system finding	118234003	Finding by site
301310005	Finding of face	298364001	Finding of head region
423716004	Petechiae of skin	297968009	Bleeding skin
414025005	Disorder of fetus or newborn	64572001	Disease
239953001	Soft tissue lesion	19660004	Disorder of soft tissue
128598002	Disorder of integument	362965005	Disorder of body system
362965005	Disorder of body system	123946008	Disorder by body site
301857004	Finding of body region	118234003	Finding by site
121000119106	Lesion of skin of face	301310005	Finding of face
111467008	Fetal OR neonatal haemorrhage	414025005	Disorder of fetus or newborn
250171008	Clinical history and observation findings	404684003	Clinical finding
64572001	Disease	404684003	Clinical finding
276619008	Neonatal facial petechiae	431268006	Haemorrhage of skin in neonate
400082007	Disorder of skin of head	128217007	Disorder of skin AND/OR subcutaneous tissue of head
123946008	Disorder by body site	64572001	Disease
297968009	Bleeding skin	106076001	Skin finding
404684003	Clinical finding	138875005	SNOMED CT Concept
298364001	Finding of head region	406122000	Head finding
118254002	Finding of head and neck region	301857004	Finding of body region
19660004	Disorder of soft tissue	123946008	Disorder by body site
118930001	Disorder of face	301310005	Finding of face
128217007	Disorder of skin AND/OR subcutaneous tissue of head	118934005	Disorder of head
85539001	Neonatal haemorrhage	22925008	Neonatal disorder
118234003	Finding by site	404684003	Clinical finding
118934005	Disorder of head	406122000	Head finding
423306009	Purpura and/or petechiae	250171008	Clinical history and observation findings
131148009	Bleeding	404684003	Clinical finding
60822000	Petechiae in fetus OR newborn	111467008	Fetal OR neonatal haemorrhage

Proof of concept files are provided in Appendix A.

There are concept and aggregation files for each of these components of the Problem Diagnosis RefSet. This allows NEHTA and secondary data users to quarantine concepts and aggregated patient counts as required by current or developing APNMDS specifications. Alternatively, they may choose to deploy them in a cohesive single aggregation RefSet.

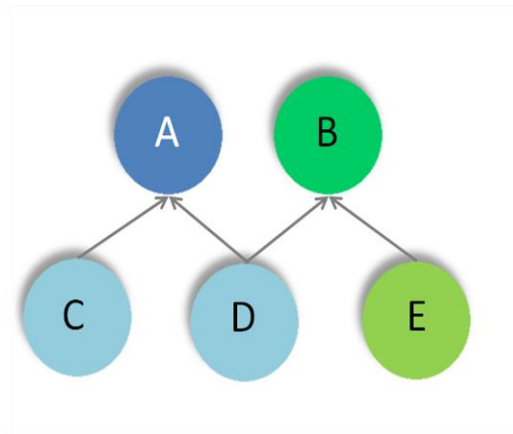
Both types of files resemble the form and formats of standard SNOMED CT-AU release files, but at this time they have been provided in *.txt format to facilitate review and feedback. These can be converted to RF2 format if required.

This approach hides all the inherent complexity from end users, providing them with what we believe are the most straightforward means of managing SNOMED CT encoded patient data for statistical purposes. While this appears to be manageable from a production and release perspective, their own operational demands and routine data analyses may demand different formats. Further consultation with data custodians will help to specify their expectations.

The aggregation RefSet is a single product, and using this requires only a single and simple programming task to accumulate (subsume) all patient case counts as shown in this recursive code fragment (example only):

```
public int calculateFreq (Concept c)
{
    int total = 0;
    if (originalFreq.containsKey(c))
    {
        total = originalFreq.get(c);
    }
    for (Concept child : c.getChildren())
    {
        total = total + calculateFreq(child);
    }
    return total;
}
```

Others have suggested that an approach that resembles ‘classification aligned ranking’ of concepts might be another way to manage the polyhierarchical nature of SNOMED CT. The calculation of frequency might function as “if(containsKey(c) && not already counted)” condition. This would overcome the inherent double counting due to multi-parenting.



However, if we consider this case and use

$\text{calculateFreq}(A) = \text{freq}(A) + \text{freq}(C) + \text{freq}(D)$

$\text{calculateFreq}(B) = \text{freq}(B) + \text{freq}(E)$

this would work well enough since when iterating over the children of B, D would already have been counted.

The problem with this approach is that either we need a stable ordering of children (and what is the “right” order), or the order is random and thus the same frequency data will

produce different aggregation counts (i.e. in case B’s value is computed before A’s).

We reviewed the outcomes achieved against the criteria previously outlined. Table 6 summarises the features of the aggregation RefSet provided.

Table 6: Aggregation technique and outcomes assessed against evaluation criteria

Criteria	Assessment method	Results achieved	
Small number of mutually exclusive top level categories (disjoint)	Less than 50 SNOMED CT high level concepts that act as report category labels and ancestors of the content of the Problem Diagnosis RefSet	✓	34 top categories for current specified APNMDS use case; less than 50 additional non-APNMDS categories
Single primary inheritance path for aggregation purposes, no dual parenting	The number of relationships in the aggregation RefSet should approximate the number of concepts in the Problem Diagnosis RefSet (ie: nodes and edges should be roughly equal in number). If the number of edges exceeds the number of nodes in the aggregation RefSet then dual or multi-parenting exists (and the aggregation methodology will have failed).	✓	No dual parenting; nodes and edges equal
Minimise orphan concepts	A residual ("other specified") to aggregate concepts that may or may not be relevant to the admitted patient sector but that can be used when required, recognised in exchange scenarios and interrogated or retrieved separately	✓	No orphans, every concept has a place Exhaustive of all Problem Diagnosis RefSet members, whether these are directly applicable to (current) APNMDS reports or not
Multiple levels of granularity	Between 2 and 23 levels based on native IS-A steps between concepts.	✓	Levels of granularity vary from 2 to 18. Drill down and roll up is possible
Maximise automation of aggregation RefSet development	Maximal automated techniques need to be developed using native SNOMED CT description logic, analytics and transitive closure reductions to provide ongoing development, update and release agility. Reliance on hand-crafting techniques should be less than 10% of RefSet volume.	✓	Minimal map reliance, maximum algorithmic development and maintenance methods
Inclusion of all patient cases (sensitivity)	The total number of patient cases attributed to concepts in the aggregation RefSet must equal the number of patient cases in the original (real world, relevant) data collection.	✓	This is yet to be tested with prospective real world data. Synthetic data tests indicate the criteria has been met.

8.6 Efficiency and reproducibility gains

What this means is that significant efficiencies are achievable, and the cost of comprehensive mapping is avoided. It addresses scope and gap issues. It also acknowledges the asynchronous nature of both ICD and SNOMED CT.

Aggregation techniques are performed algorithmically, so can easily be re-run and reproduced each six months to account for content variations across SNOMED CT releases, using computational approaches.

The stability of ICD (released every 2 years) means that few, if any revisions are likely to Chapters across editions, so the burden of keeping a comprehensive mapset aligned over time is also avoided.

8.7 Implications

The results obtained here reveal a number of factors that continue to worry implementers and users of SNOMED CT products.

Currently the end-to-end use of health information has been motivated by secondary data users who specify and mandate what 'their end' of the data flow requires, at the expense of clinical utility. That is, under most peak committees and standards development initiatives, the reporting requirements - the outputs - have been the drivers of what health data is and should be collected – the inputs. We see this in NHISSC and in ABF initiatives. Clearly, this is a "cart before the horse" scenario.

The usual concerns and arguments we hear from stakeholders continue to focus on re-using SNOMED CT, and not re-purposing SNOMED CT content, as we have demonstrated here.

This is a fundamental misunderstanding about the model of meaning (as encapsulated in a clinical terminology) and the model of use (as specified by secondary data management protocols).

Purist terminologists will advocate that the definitions and descriptions provided by the logic and content of SNOMED CT should dictate how the terminology is used in practice. For example: “Don’t use a Disorder concept in a Family history context”.

Purist secondary data users advocate that the data element definitions should determine which terms are valid and what they mean in the context of the information model. For example: “Only concepts related to diseases are valid concepts for a Diagnosis data element”.

Unfortunately, real clinical users and their patients are caught between these two purist approaches.

But when we look to real world practice and real data collections, we find that the clinical community has, does and will likely continue to document patient cases in a way that suits clinical care; they consider downstream data collection protocols as “office use only” slots on forms, not applicable to care delivery.

Empiricism shows that:

- users capture concepts that represent reason for encounter, procedures, events, observations and situation concepts in diagnosis, problem, complaint, presenting problem, data elements
examples: Intentionally harming self
 Suspected child abuse
- existing (and classic) health classification like ICD and ICPC also acknowledge that non-diagnosis concepts have a real place in real practice and can be considered valid entries documenting patient conditions in a diagnosis or problem field
examples: Pregnancy
 Normal delivery
- despite their protestation, secondary data analysts already allow the collection and counting of non-diagnosis concepts in an otherwise diagnosis-focussed reporting protocol for the APNMDS use case under consideration here
examples: Chemotherapy
 Dialysis

These sort of demarcations between meaning and use are unhelpful, and if the artificial partitions between RefSets and RefSet content were relaxed, a greater degree of interoperability and exchange-ability might be achieved between care sectors along the end-to-end health information pipeline.

9 Conclusion

The investigative work undertaken here shows promising results and further opportunities could be pursued to provide more definitive guidance about how secondary users might re-purpose SNOMED CT encoded patient data.

Comparative and repeated measures are needed to ensure that these techniques are suitable and/or can be refined, thus providing objective evidence that stable and reliable outputs for secondary data users can be produced.

Recent and ongoing development in SNOMED CT query language and data analytic approaches have informed some aspects of these techniques, and as the work being undertaken by the IHTSDO progresses further options are sure to emerge.

Appendix A

The following files can be obtained by requesting through help@nehta.gov.au

- APNMDS category aggregation nodes
- APNMDS category aggregation edges

- Other specified nodes
- Other specified edges

- Other nodes
- Other edges

- Aggregation concepts nodes
- Aggregation links edges

CONTACT US

t 1300 363 400
+61 3 9545 2176
e enquiries@csiro.au
w www.csiro.au

YOUR CSIRO

Australia is founding its future on science and innovation. Its national science agency, CSIRO, is a powerhouse of ideas, technologies and skills for building prosperity, growth, health and sustainability. It serves governments, industries, business and communities across the nation.

FOR FURTHER INFORMATION

eHealth Program
Digital Productivity Flagship
Dr David Hansen
t +61 7 3253 3610
e david.hansen@csiro.au

eHealth Program
Digital Productivity Flagship
Dr Michael Lawley
t +61 7 3253 3609
e michael.lawley@csiro.au