murdoch children's research institute

LifeCourse

Describing data captured in the LifeCourse platform using a standardised terminology

Last updated: **25.08.2021**

Version: **1.0**

Acknowledgements

This work was supported by the Royal Children's Hospital Foundation grant #2018-984. LifeCourse acknowledges all collaborators, cohort representatives and participants. See https://lifecourse.melbournechildrens.com/contact/ for further details.

Suggested citation

O'Connor, M., Paiva, T., Duncan, A., on behalf of the LifeCourse Initiative. 2021. Describing data captured in the LifeCourse platform using a standardised terminology. Murdoch Children's Research Institute. Melbourne, Australia. https://doi.org/10.25374/MCRI.15236313.

Contents

Executive summary

The LifeCourse platform aims to enable researchers to capitalise on the wealth of longitudinal cohort data available at the Melbourne Children's Campus to advance understanding of key health issues facing children and young people. Making these valuable cohort data findable and accessible for researchers is integral to achieving this goal.

The LifeCourse website provides a bird's eye view of what data has been collected across our cohorts. Measures used within each study wave are described by a common set of terms that capture the constructs assessed, which are then broadly summarised into a smaller number of domains. This enables easy and intuitive browsing and searching.

Historically, LifeCourse has used an in-house developed set of terms and domain groupings to describe measures. This approach was flexible and allowed data custodians to describe data in a way that they preferred. However, with the scale that the platform has grown to, this approach is no longer tenable. Of particular concern is a lack of consistency, with the same measures described using different terms between studies.

Standardising how we describe and organise data captured across LifeCourse cohorts will allow us to achieve a more streamlined, consistent, and intuitive format for browsing and searching of cohort data. A vital first step is to standardise the LifeCourse terminology itself – the set of terms and groupings of these terms that are used to describe and organise the data. This is the focus of this report.

To standardise the LifeCourse terminology, we drew from pre-existing, internationally recognised ontologies. We prioritised the Systematized Nomenclature of Medicine (SNOMED), because it is a widely used structure for describing biomedical research data internationally and had the broadest coverage of constructs captured by LifeCourse cohorts. We proceeded term-by-term through the LifeCourse ad hoc developed system, identifying commensurate terms from SNOMED that met pre-defined criteria. Where an appropriate SNOMED term could not be identified, we searched in the Medical Subject Headings (MeSH) as another widely utilised system, before turning to an in-house solution reached through discussion and consensus in the LifeCourse team.

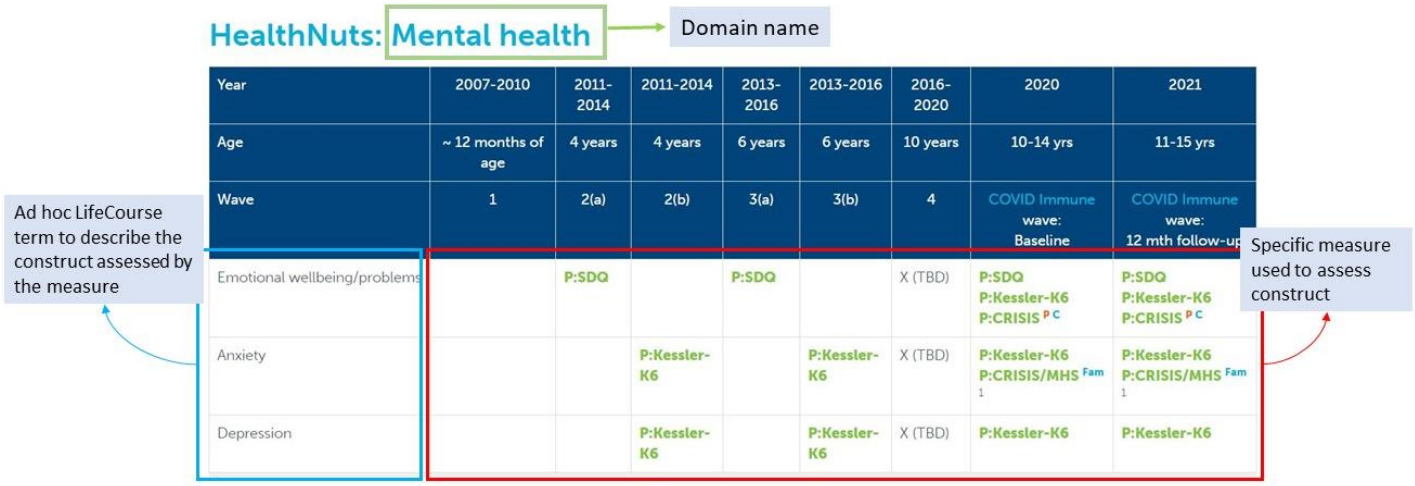
The LifeCourse standardised terminology now contains 629 unique terms, organised across 34 domains, which is a 15% reduction of terms in comparison to our previous ad hoc system. We found that 80% of concepts measured across LifeCourse cohorts could be described using SNOMED terminology, thereby achieving strong alignment to this international standard.

The LifeCourse standard terminology is a living system and is expected to change and develop over time. For example, constructs assessed by new measures may not be represented in the current terminology. We outline a systematic process that meets the needs of responding to new measures, inaccuracies, or other concerns that may arise, while ensuring fidelity to a standardised approach so that consistency is retained.

The next step in this process involves mapping of measures to the terms defined. Our goal is to develop a measures-to-terms map which will be used to autopopulate the LifeCourse website. The latter will be made possible by work currently underway to implement a website content management system.

Background

A set of terms are used to describe and organise the constructs captured across the LifeCourse studies on the LifeCourse website. These terms are further summarised into a smaller number of domains. These descriptions and domain groupings allow data users to browse and search the available data more easily. See *Figure 1* below for an example of how this information is currently presented on the LifeCourse website.



Year	2007-2010	2011-2014	2011-2014	2013-2016	2013-2016	2016-2020	2020	2021
Age	~ 12 months of age	4 years	4 years	6 years	6 years	10 years	10-14 yrs	11-15 yrs
Wave	1	2(a)	2(b)	3(a)	3(b)	4	COVID Immune wave: Baseline	COVID Immune wave: 12 mth follow-up
Emotional wellbeing/problems		P:SDQ		P:SDQ		X (TBD)	P:SDQ P:Kessler-K6 P:CRISIS P C	P:SDQ P:Kessler-K6 P:CRISIS P C
Anxiety			P:Kessler-K6		P:Kessler-K6	X (TBD)	P:Kessler-K6 P:CRISIS/MHS Fam 1	P:Kessler-K6 P:CRISIS/MHS Fam 1
Depression			P:Kessler-K6		P:Kessler-K6	X (TBD)	P:Kessler-K6	P:Kessler-K6

Figure 1. Labelled example of domains and terminology on the LifeCourse website

Historically, LifeCourse has used an ad hoc set of terms and domain groupings which have developed over time. This included 732 terms across 43 domains, developed by the LifeCourse team and data custodians in-house. While providing a flexible solution early in the development of LifeCourse, this ad hoc approach now requires standardisation to improve consistency and optimise browsing and searching functions. This document outlines the benefits of a standardised terminology, the process used to develop a standardised LifeCourse terminology, and future directions in this space.

Why is a standardised LifeCourse terminology important?

Standardising the terminology used to describe LifeCourse data is important for creating a more consistent and streamlined format for browsing and searching of cohort metadata, in line with our commitment to FAIR (Findable, Accessible, Interoperable, Reusable) data principles. The previous ad hoc approach resulted in the same data being described with varying terms across cohorts. For example, one study could assign the term ‘Depression’ to a measure, while another listed the same measure under ‘Depressive symptoms’. This creates unnecessary confusion when browsing cohort metadata, and limits the capacity to identify comparable data across cohorts.

A standardised approach will also help to improve efficiency when adding new studies or waves of data to the LifeCourse website. The process for assigning terms to the measures within a new LifeCourse study

will be more efficient when choosing from a clearly defined, standard set of terms, and can be automatically assigned for commonly used measures.

In addition, a standardised system is essential for enabling other key LifeCourse priorities and potential enhancements in future, most notably the implementation of a back-end database for the LifeCourse website, which can automatically organise study data.

Available pre-existing terminologies

To ensure alignment with existing standards, we canvassed the availability of pre-existing and widely used terminology systems. These systems are often referred to as ontologies because they contain not only standardised terms, but also relationships between terms.

Key existing ontologies and resources for browsing terms identified include:

- The BioPortal recommender function (https://bioportal.bioontology.org/recommender), which provides recommendations for the most relevant ontologies for a given term of interest
- Systematized Nomenclature of Medicine (SNOMED, http://bioportal.bioontology.org/ontologies/SNOMEDCT?p=classes), commonly used in the medical research field
    - Australian-maintained SNOMED version (SNOMED CT-AU), which contains additional terminology specific to the Australian context. It can be browsed using the Shrimp terminology browser at http://ontoserver.csiro.au/shrimp/
- Medical Subject Headings (MeSH, https://bioportal.bioontology.org/ontologies/MESH/?p=classes), developed by the National Library of Medicine and widely used in medical databases (e.g. MEDLINE/PubMed)

Of these, the most relevant ontology identified was SNOMED. SNOMED had the most comprehensive coverage of the constructs captured across LifeCourse data and has also been widely utilised in biomedical research. Some gaps in coverage were noted however, particularly for concepts relating to education and childcare, bioanalyses, and omics.

Nevertheless, drawing on and adapting from already established terminologies where possible is important to align the LifeCourse approach to international conventions, and to enable comparisons of data availability across research groups and institutes. Hence, we prioritised SNOMED as the foundation for the LifeCourse terminology, recognising that additional terms would be needed to fill some gaps in coverage.

Approach to standardising the LifeCourse terminology

In developing our approach to standardising the LifeCourse terminology, our goals were to produce a solution that:

- Enabled data to be consistently described on the LifeCourse website
- Linked to internationally adopted systems (predominantly SNOMED) where available; and
- Was feasible to undertake within current resourcing and time constraints.

Our process was informed by our experience with the Comprehensive Monitoring Project (CMP), which defined a psychosocial data classification system to structure a mental health population surveillance tool. To refine our strategy, we also consulted with the MCRI Ontology Working Group, which includes members from the MCRI Data Working Group and Generation Victoria (GenV) team.

In summary, key steps to standardising the LifeCourse terminology included:

1. Identifying standard terms: browsing pre-existing ontologies (using SNOMED as a foundation) to find parallel terms to those in the ad hoc system, and evaluate their appropriateness based on a set of pre-determined criteria.
2. Organising terms into domains: reviewing and streamlining current domain names, while assessing the location of the terms located within each domain.
3. Implementation: this system can now be applied to the website retroactively and prospectively (work to still be undertaken).

The first two steps were undertaken from September 2020 to February 2021 and are outlined below. After completing those steps, the standardised LifeCourse terminology generated contained 629 unique terms, summarised in 34 domains (Table 1). Contact lifecourse@mcri.edu.au for the full list of domains and terms.

Table 1. Domains and example terms in the standardised LifeCourse terminology.

Domain name	Description	No. of terms	Example term
Allergies	Potential allergies or allergy treatments	27	Food allergy
Anthropometrics	Measurements of body size	25	Height
Bioanalyses and omics	Analytical methods and substances within samples being observed	58	Genotyping
Biosamples	Samples collected	34	Buccal
Cardiovascular health	Cardiac history, interventions and assessments	18	Cardiorespiratory fitness

Community environment	Characteristics of the community and an individual's engagement with the community	8	Social support
Demographics	Basic descriptive information	31	Socioeconomic status
Education and childcare	Characteristics of the educational setting and an individual's engagement with education	21	Current education
Environmental exposures	Allergens or toxins and exposure to these	17	Pet exposure
Family environment	Characteristics of the family environment	22	Parenting behaviour
Health services	Engagement with and perceptions of the health care system	50	Quality of health care
Hearing	Hearing loss and associated interventions	6	Hearing aid
Imaging	Imaging technology and scans performed or reported on	10	Magnetic resonance imaging
Medications and supplements	Prescribed and non-prescribed medications and supplements	13	Medications
Mental health and behaviour problems	Mental health problems and externalising behaviours	42	Antisocial behaviour
Methodology	Characteristics of study design and participation	18	Eligibility for trial
Neurocognitive development	Neurocognitive functioning	23	Executive cognitive functions
Nutrition	Diet and nutritional intake	9	Dietary intake
Other health information	Health outcomes and information not listed elsewhere	63	Cause of illness or injury
Peer relationships	Quality and characteristics of peer relationships	3	Peer relations
Physical activity	Levels of physical activity and fitness	7	Sports activity
Physical appearance	Descriptions of physical attributes	5	Hair colour
Pregnancy and birth	Characteristics of pregnancy and birth	68	Due date
Psychosocial wellbeing	Psychosocial assets that promote wellbeing	20	Self-esteem
Puberty	Pubertal milestones and outcomes	5	Age at menarche
Respiratory health	Respiratory health conditions and assessments	21	Respiratory tract infection

Romantic relationships	Quality and characteristics of intimate relationships	4	Intimate partner relations
Screen and technology use	Use of screens, social media, and other devices	3	Social media
Sexual health and activity	Sexual relationships and health	4	Sexual activity
Sleep	Sleep behaviour or problems	6	Quality of sleep
Speech and language	Speech and language development and problems	8	Stuttering
Substance use	Use of any substance that alters behaviour or cognitive function	16	Alcohol consumption
Temperament and personality	Aspects of temperament and personality	5	Curiosity
Miscellaneous	All other terms not covered by above domains	9	Ability to drive

Identifying standard terms

Approach to selecting standard terms

We used our list of ad hoc LifeCourse terms as the basis for identifying commensurate terms in existing ontologies. Members of the LifeCourse project team (Anna Duncan, Tehani Paiva) who were familiar with the ad hoc LifeCourse terms and corresponding data proceeded through the list of LifeCourse terms in a term-by-term fashion. Team members cross-coded one domain initially to ensure consistency of approach and identify any remaining process issues. From there, team members reviewed domains based on their area of content expertise.

During a preliminary check, duplicate terms were removed, and some terms were split if they described multiple constructs that did not cohere into a single higher-order factor (e.g., where '/' had been used to combine multiple disparate constructs). If an unknown LifeCourse term was encountered, its description and current use of the term was reviewed to understand the construct being described.

Following this preliminary check, a prioritisation approach was used to select standardised terms:

1. Using the ad hoc set of LifeCourse terms as a basis, preliminary work was conducted to list alike SNOMED and MeSH terms
2. Due to the widely used nature of the SNOMED medical ontology, terms from SNOMED were prioritised. That is, where an appropriate SNOMED term was identified, this was selected as the LifeCourse standard term if it fulfilled the below criteria (Box 1)
3. If the SNOMED term was found to be inappropriate according to the below criteria (Box 1), the MeSH term was adopted

4. If the MeSH term was also found to be inappropriate according to the below criteria (Box 1), the LifeCourse team member reviewing that domain suggested an alternate term. These LifeCourse-developed terms were discussed within the LifeCourse team until a consensus was reached.
5. If the LifeCourse team were unable to reach consensus and agree upon a term, an expert consultant was contacted. This step was performed sparingly.

Box 1. Criteria for standard LifeCourse terms

The following criteria was used to define appropriate LifeCourse standard terms:

1. A standard term should describe the construct measured, so that data users can intuitively search for and identify potentially relevant data
2. While a standard term provides an indication of the overarching construct, it is not intended as a comprehensive description of the data (data users still need to review the specific item/measures used)
3. The term describes the core construct captured, but not other characteristics of the data such as the measure name, reference period, mode of data acquisition, or who the measure is about
4. Terms should describe one construct only, avoiding use of '/' to combine multiple constructs
5. Standard terms should not be redundant with one another; duplicates should be removed
6. Given the breadth of data captured across LifeCourse, we accepted that there would be some variation in the level of specificity across terms
7. Terms were considered acceptable that capture the appropriate construct, regardless of the directionality implied (e.g. 'peer relationship problems' would be acceptable, though specific measures may not necessarily focus on the deficit end of this continuum)
8. For mental health problems, terms were considered acceptable that capture the appropriate construct, regardless of whether a diagnostic or dimensional/symptoms inventory approach was implied (e.g., depressive disorder or depressive symptoms would both be considered acceptable)
9. For biosamples and bioanalyses, a general term was considered acceptable when specific tests are not known (e.g. Hormone measurement). When specific tests are known, these terms were also included (e.g. Testosterone, Oestrogen, FSH, LH, GH, etc.)

In undertaking this process, we found good alignment with SNOMED, with almost 80% of standardised terms being drawn from SNOMED (Table 2). The remaining 20% of terms were comprised of a combination of MeSH and LifeCourse-derived terms. Additionally, many terms from our ad hoc system (N=268) were merged into other constructs or removed completely. The total number of terms was reduced by 15% (from N=732 to N=629).

Table 2. Identification of LifeCourse standardised terms.

Outcome	Number of terms	% of terms
SNOMED term adopted	499	79.4
MeSH term adopted	54	8.6
LifeCourse-derived term adopted	76	12.1
Term merged	221	n/a
Term removed	47	n/a

Presentation of standard terms on the LifeCourse website

Once a standard term was identified, the presentation of this term on the LifeCourse website was also considered. Some alteration to the website presentation was made for almost 30% of terms (Table 3).

Table 3. Terms requiring alteration to their presentation on the LifeCourse website.

Outcome	Number of terms	% of terms	Example
Presented as full term (no change)	422	67.7%	No change from confirmed term
Presentation changed for brevity	180	28.8%	Removal of “finding of -”
Presentation changed for clarity	17	2.7%	“FH:Obesity” clarified as “Family history of obesity”
Presentation changed for spelling or grammar	5	0.8%	“Randomization” changed to “Randomisation” (Aust. spelling not available)

Over one quarter of terms required a change for brevity (Table 3). SNOMED terms can be lengthy and often include words or phrases that are not essential to the meaning of the term in this context. Common examples include: ‘finding’, ‘finding related to’, ‘finding of’, ‘finding of level of’, ‘observable entity’ or ‘observation’. If the meaning of the term would not be lost or changed, a briefer version of the term removing these phrases was recorded for display on the website. SNOMED terms also sometimes included abbreviations as well as the full unabbreviated term (e.g. MRI – Magnetic Resonance Imaging); we similarly removed abbreviations from the website display if the meaning would not be altered.

MeSH terms did not always provide Australian spelling alternative labels. When a MeSH term was used, and the alternative labels did not provide an option with Australian spelling, the presentation on the

LifeCourse website was altered to adapt the term to standard Australian spelling conventions (for example, ‘utilization’ presented as ‘utilisation’).

While not the intended audience, it is possible for a cohort participant to arrive at the website. Content therefore needs to be sensitive to potential participant concerns. Terms were reviewed to ensure that no descriptors were overtly stigmatising or offensive. That is, language outside of usual academic norms which might result in detrimental consequences for a cohort, such as participants un-enrolling from the study. No such changes were deemed necessary based on this review.

Organising terms within domains

After confirming the list of standard terms, domains were reviewed, including both their groupings and the domain name assigned to describe each grouping. Once the list of domains was finalised, the standard terms sitting within each domain were re-reviewed to ensure sensible placement in one or more domains.

Determining a domain

To determine domain groupings and names, we began with the list of 43 ad hoc LifeCourse domain names and took the following iterative approach:

1. The existing domain groupings were evaluated for their appropriateness according to the pre-defined criteria below (Box 2), with possible outcomes of:
    - Retaining the current domain grouping
    - Merge two or more domains, or remove a domain and re-allocate terms to more appropriate domains
    - Split a domain into two or more separate domains
2. Once groupings were established, domain names were searched in both SNOMED and MeSH to determine whether there were available terms to describe the domain level construct
3. If SNOMED and MeSH terms were found to be inappropriate according to the below criteria (Box 2), the LifeCourse team member suggested an alternate domain name. These LifeCourse-developed domain names were discussed within the LifeCourse team until a consensus was reached
4. After finalising the domain names, the terms within each domain were re-reviewed to ensure they still aligned and that all relevant terms were captured

Box 2. Criteria for domain groupings and names

The following criteria were used to define the requirements for domain groupings and names, whereby domains should:

1. Describe one high level construct only; if two or more high level constructs are represented within one domain (e.g., by using 'and' to combine disparate constructs), these should be split into separate domains
2. Be easily understood to allow for intuitive browsing (e.g., by using commonly used language rather than bespoke or theory-specific groupings and/or names)
3. Domain name should be brief (ideally no more than four words for ease of website presentation), and contain no symbols (e.g. avoid forward slash '/')
4. Groupings should be substantial enough to capture a range of terms, and niche categories with a small number of terms should only be used where essential (e.g., if terms do not fit elsewhere and the few relevant terms are widely utilised)
5. The same term could sit in more than one domain if there is a clear and meaningful rationale for both
6. Domain names should describe the core construct captured by the domain, but not other features of the data such as the mode of data acquisition; except in cases where this is the central concept for a group of terms (e.g., biosamples, imaging, and methodology)
7. Terms should only be included in a domain if they are directly relevant (e.g., in the Mental health domain, all mental health conditions are listed but risk factors for mental health problems are not included). That is, erring on the side of being under- rather than over-inclusive with placement of a term within a domain

Organisation into domains

After carrying out the process outlined above, we condensed and re-organised the original 43 ad hoc domains into 34 domains (26% reduction in the number of domains). The final domain groupings ranged from including 3 to 68 terms, as shown previously in Table 1.

10 of the domain names aligned with a SNOMED or MeSH term (SNOMED = 6, MeSH = 4). While many of the other concepts did align to SNOMED or MeSH terms, these terms were often not sufficiently intuitive and browser-friendly. In these cases, the domain names are LifeCourse-derived.

Future directions

Integration into LifeCourse website

The next step in this process involves mapping of measures to the terms defined. Our goal is to develop a measures-to-terms map which will be used to autopopulate the LifeCourse website. The latter will be made possible by concurrent work to implement a website content management system.

In undertaking this work, consideration should be given to how terms are shown on the website. Currently, no guidance about what terms represent (e.g., via a column heading) is provided. This can raise concerns when the term is not fully comprehensive of a data element or not the preferred term of a cohort custodian.

One option for addressing this concern is to provide further orienting text for visitors to the 'Explore our cohorts' page, that outlines what terms represent. Another easily implemented solution is to include a heading for the list of terms on individual cohort pages, such as 'Relevant LifeCourse concept/s'.

Ongoing development and improvement

The LifeCourse standard terminology is a living system. It is expected to develop further as new measures are encountered for which no relevant terms are available, or when issues in the application of terms to a specific measure arise. Issues may also arise in the presentation of standard terms on the website.

Where this occurs for standard terms:

1. The term of concern should be flagged by a LifeCourse team member and a rationale provided as to why the existing term or its website presentation is inappropriate, or no existing term is available, using the previously defined criteria
2. The LifeCourse team will discuss and reach a consensus on whether any action is required (e.g. adding a new term, modifying existing term). A conservative approach will be taken, requiring clear justification that current terms are inappropriate or insufficient
3. If no consensus is reached, expert input will be sought
4. Changes will be recorded by LifeCourse team members in a change log and will be implemented onto the website as required.

For domains, a similar process will be used:

1. Concerns with the domain grouping or name should be flagged by a LifeCourse team member
2. The team member flagging this concern will provide a rationale as to why the existing domain name is inappropriate or insufficient, or why the location(s) of a term within a domain is inappropriate, according to the domain criteria
3. The LifeCourse team will discuss and reach a consensus on whether any action is required (e.g. adding a new domain, modifying existing domain name, altering the location(s) of a term within a

domain). A conservative approach will be taken, requiring clear justification that current domain grouping or name is inappropriate or insufficient

4. If no consensus is reached, expert input will be sought
5. Changes will be recorded by LifeCourse team members in a change log, and will be implemented onto the website as required

Ensuring alignment to campus conventions

The MCRI Ontology Working Group will provide additional opportunities for collaboration and alignment of our approach with other MCRI groups in future. Other campus-level work also provides opportunities for embedding and implementing the use of standard terms. For example, standard terms can be allocated during a survey's development in REDCap and integrated within a pre-defined REDCap measurement library.

Engaging with SNOMED

SNOMED also has the advantage of providing the option for collaboration by extending upon the existing set of terms. It allows external groups to propose new terms which may be integrated into their official ontology. Feeding back the results of this process to SNOMED could be an opportunity to further develop and align to international standards in future.

Contact details

Melbourne Children's LifeCourse Research
E: lifecourse@mcri.edu.au

Murdoch Children's Research Institute

The Royal Children's Hospital
50 Flemington Road
Parkville, Victoria, 3052 Australia

www.mcri.edu.au