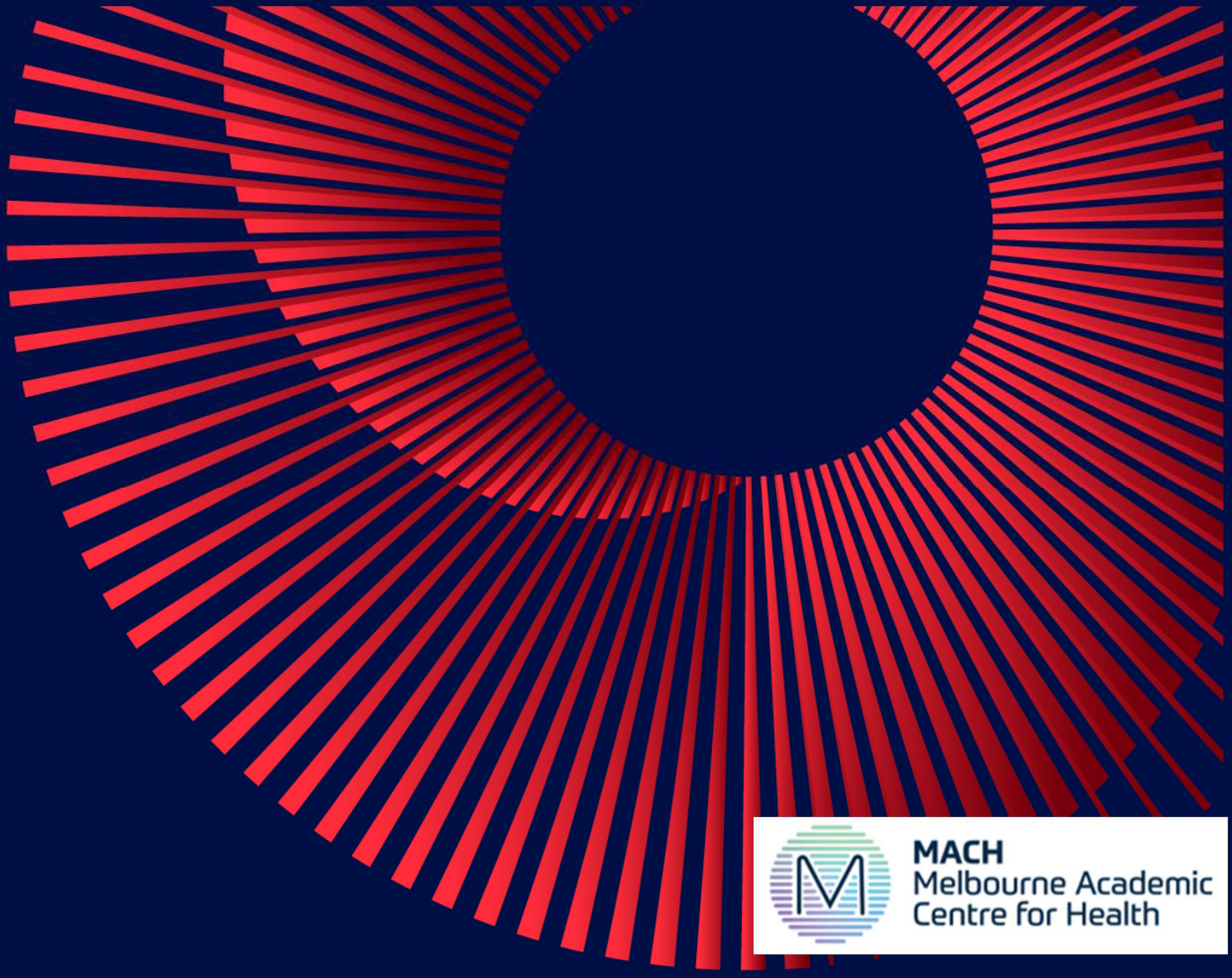
THE UNIVERSITY OF MELBOURNE

Faculty of Medicine, Dentistry and Health Sciences

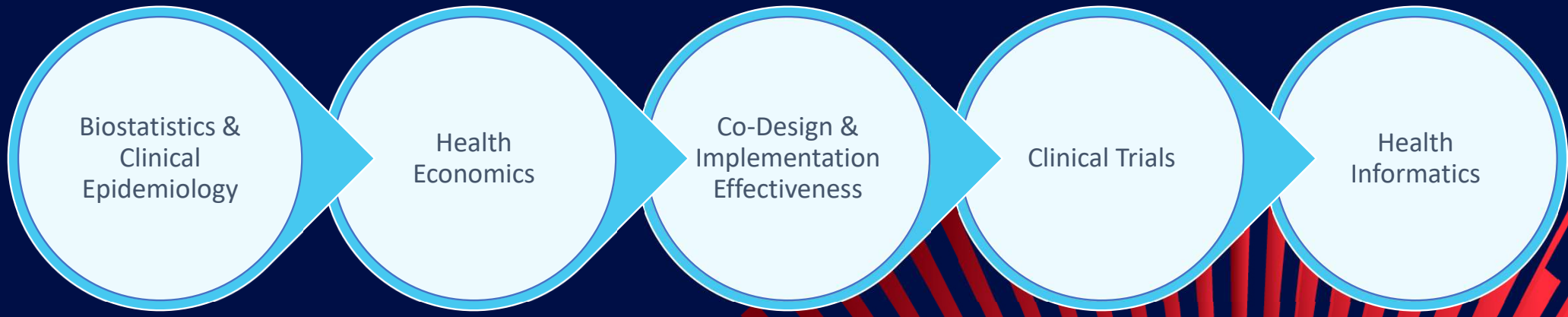
MISCH *Hub*

Methods and Implementation Support for Clinical and Health Research



MACH Melbourne Academic Centre for Health

MISCH Hub Structure



MISCH *Staff*


Clinical Trials Lead
Adam Deane
MISCH Manager
Shaie O'Brien
Co-Lead MISCH Biostatistics
Sabine Braat
Biostatistician
Ximena Camacho
Co-Design and Consumer Involvement
Jess Kirwan
Health Economist
Kim Dalziel
Co-Design and Implementation Effectiveness Co-Lead
Cathy Vaughan
Biostatistician
Emily Karahalios
Health Economist
Bianca Lau-Goodchild
MISCH Operations Manager
An Nguyen
Biostatistician
Lizzie Korevaar
Biostatistics Lead, Austin Hospital
Leonid Churilov
Biostatistician
Dominic Italiano
Biostatistician
Fiona McManus
Biostatistician
Yee-Foong Mok
Director of MISCH
Julie Simpson
Health Economist Lead
An Duy Tran
Biostatistician
Rob Mahar
Epidemiologist and Biostatistician
Digsu Koye
Biostatistician
Aistair McLean
Biostatistician
Peixuan Li
Biostatistician
Anurika De Silva
Co-Lead MISCH Biostatistics
Karen Lamb
Biostatistician
Sophie Zaloumis
Biostatistician
Phoebe Fitzpatrick
Clinical Trials Manager
Katie Ozdowska
Biostatistician
Diana Zannino
Health Informatician
David Ormiston-Smith
Biostatistician
Hannah Johns
Health Informatics Lead
Douglas Boyle
Health Economist
Peter Lee
Biostatistician
Nadine Phillips
Co-Design and Consumer Involvement
Mary Stathopoulos
Communications Coordinator
Jessie Lu-Lee

THE UNIVERSITY OF MELBOURNE

MACH EMCR Data Management Seminar

Making Data Work for You in Clinical Trials
Best Practices in Clinical Research Management

Associate Professor Sabine Braat
Co-Head of MISCH Biostatistics Hub

Data quality issues



Incorrect data

Imprecise data

Incomplete data

Duplicate data

Missing data

Inconsistent data

Unclear data

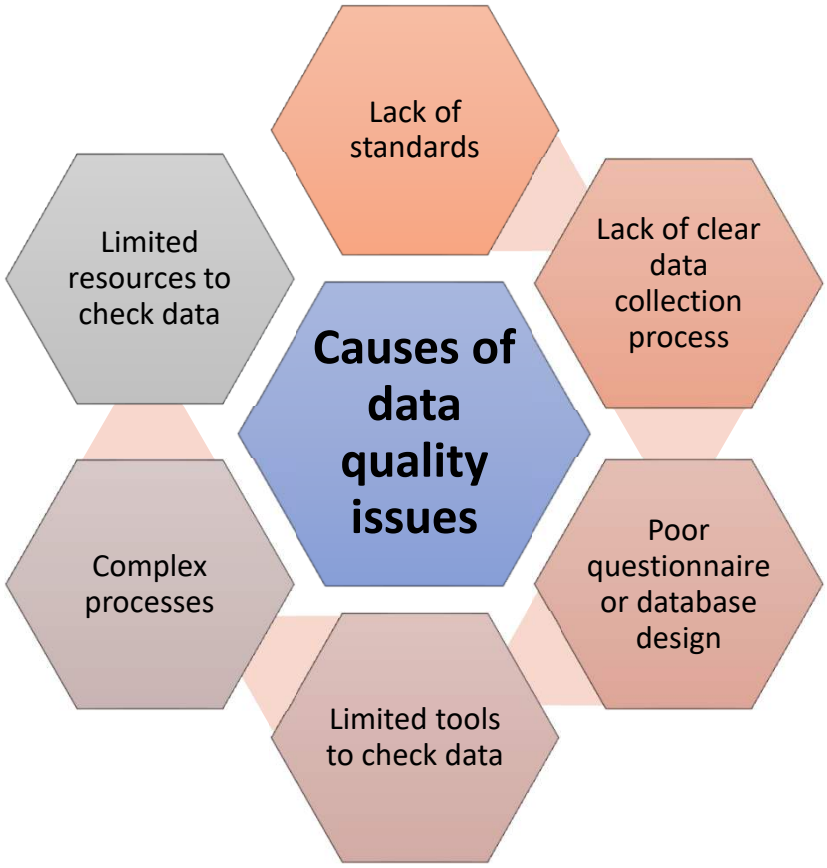
© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

What is the **biggest challenge** you face when ensuring data quality?

Join at
slido.com
#1925 378

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Data quality challenges



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Data start and finish



Usual Care

The NEW ENGLAND JOURNAL of MEDICINE

Mean % weight loss *MACH* diet: 20.2%

Mean % weight loss Usual Care: 13.7%

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey



Usual Care

The NEW ENGLAND JOURNAL of MEDICINE

Mean % weight loss *MACH* diet: 20.2%

Mean % weight loss Usual Care: 13.7%

Study design → **Protocol design** → **Data questionnaire design** → **Database design** → **Study conduct** → **Participant safety** → **Statistical analysis**

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Study design

Ensure quality via **study design**:

– Is *MACH* diet superior to Usual Care in reducing weight at Week 12?

Baseline **Week 6** **Week 12**



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Study design

Ensure quality via **study design**:

– Is *MACH* diet superior to Usual Care in reducing weight at Week 12?



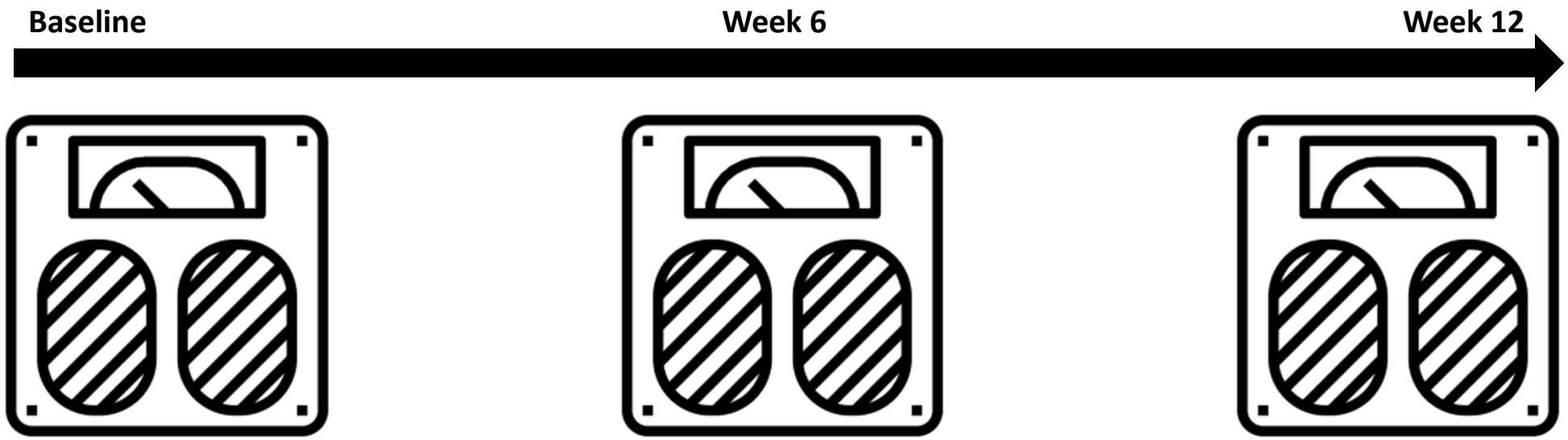
© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Study design

Ensure quality via **study design**:

– Is *MACH* diet superior to Usual Care in reducing weight at Week 12?



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

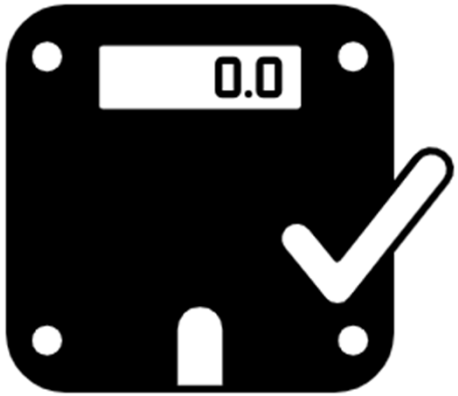
Protocol design

Ensure quality via **protocol**:

- How to collect data
- Who collects data
- Where to collect data
- When to collect data (visit schedule)
- What if a participant discontinues treatment, deviations visit schedule,...

We will follow the WHO STEPwise approach for physical measurements. A digital calibrated scale (Secca 876) with precision to the nearest 100 grams will be used to measure body weight (kilograms). Each measurement will be taken three times and verified by two researchers at the study site, while in fasting state, after removal of footwear, … at baseline, week 6, and week 12 at about the same time in the morning.





© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Protocol design

- Review **protocol**
- Check consistency of aspects related to statistics such as:
  - Objectives
  - Hypotheses
  - Sampling
  - Outcomes
  - Randomisation/blinding
  - Sample size/analyses
- Involve **data manager** and **statistician**

Required Elements		Yes	No	NA	If No, specify action taken or comment
1.	Title is clear and descriptive	☐	☐	☐	
2.	A trial schema is included and is clear and concise	☐	☐	☐	
3.	The schedule of assessments table contains all required baseline, treatment/intervention, safety and endpoint data (and nothing else)	☐	☐	☐	
4.	The hypotheses of the trial are clear and concise, if applicable	☐	☐	☐	
5.	The objectives of the trial are clear, concise and measurable	☐	☐	☐	
6.	The endpoints/outcomes of the trial are clear, appropriate and measurable (and a time point is specified)	☐	☐	☐	
7.	The eligibility criteria are internally consistent	☐	☐	☐	
8.	Recruitment strategy/sampling details are provided (e.g. number of study sites)	☐	☐	☐	
9.	Detail is provided on how informed consent will be obtained	☐	☐	☐	
10.	The number of study arms is clearly specified	☐	☐	☐	
11.	The treatment/intervention details (what and when) are clear for all arms of the trial (including the control arm)	☐	☐	☐	
12.	There is a clear description of the study design (e.g. parallel arm, crossover, etc.) and the study design is appropriate to address the objectives	☐	☐	☐	
13.	Randomisation details are clear and appropriate (allocation ratio, how sequence is generated, stratification, minimisation,	☐	☐	☐	

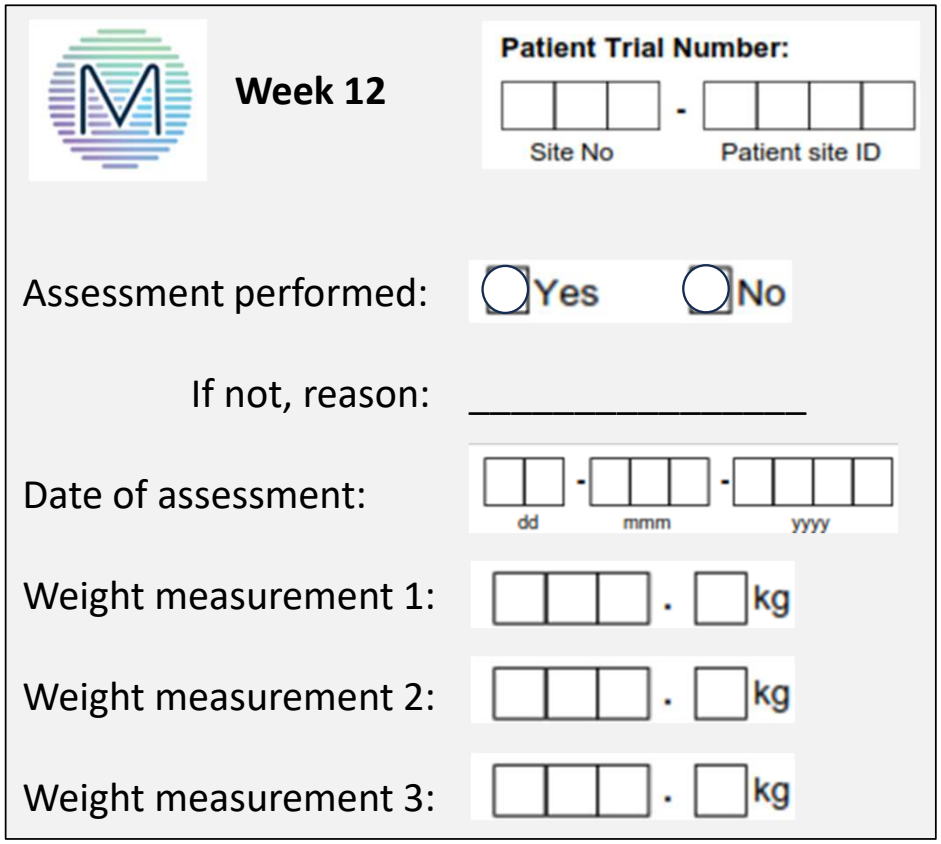
© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Data questionnaire design

Ensure quality via **questionnaire** tool:

- Align with protocol
- Avoid duplication and free text
- Standardise data formats and coding
- Pilot the questionnaire with patients and data collectors
- Provide training and data completion instructions for accuracy
- Ensure version control



Week 12

Patient Trial Number:

Site No - Patient site ID

Assessment performed: ◯ Yes ◯ No

If not, reason: _____

Date of assessment: dd - mmm - yyyy

Weight measurement 1: . kg

Weight measurement 2: . kg

Weight measurement 3: . kg

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Data questionnaire design

- Review **questionnaires**
- Involve **data manager** and **statistician**

Required Elements		Yes	No	NA	If No, specify action taken or comment
1.	Primary endpoint(s): Are all the data required for the primary endpoint being collected and in such a way that it can be analysed appropriately?	☐	☐	☐	
2.	Secondary endpoints: Are the data being collected for all the secondary outcomes detailed in the protocol in such a way that they can be analysed appropriately?	☐	☐	☐	
3.	Are all the data required to determine the analysis populations being collected e.g. identification of participants who are lost to follow-up (if applicable)?	☐	☐	☐	
4.	Are data on compliance being collected (where relevant)?	☐	☐	☐	
5.	Are all of the safety data detailed in the protocol being collected and in such a way that it can be analysed appropriately?	☐	☐	☐	
6.	Are the data needed to ensure the eligibility of the participant being collected?	☐	☐	☐	
7.	Are data captured in closed form wherever possible?	☐	☐	☐	
8.	Is data coding consistent? (e.g. if 1 is coded yes and 0 is coded no, this is used consistently)	☐	☐	☐	

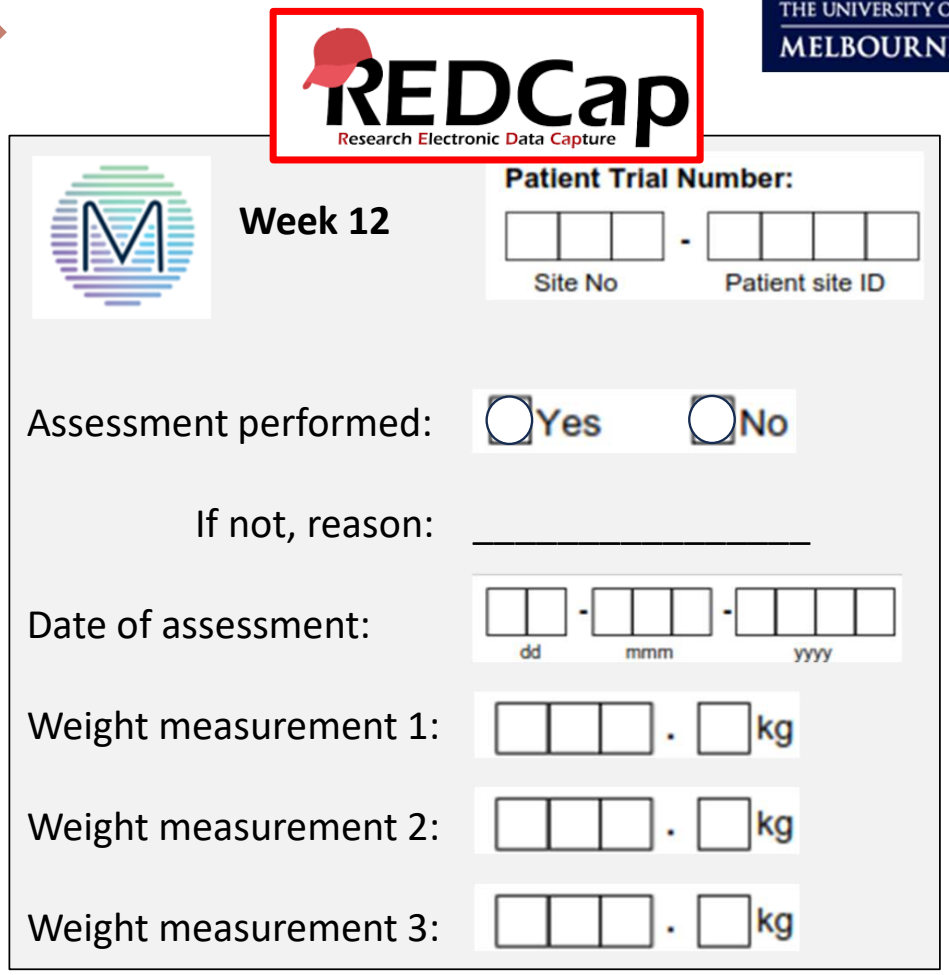
© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Database design

Ensure quality via **database**:

- Radio buttons versus tick boxes
- Numerical fields for numerical data
- Branching logic (if-then-else)
- Mandate/enforce if fields are not allowed to be incomplete/missing
- Range checks by data field
- Logic checks between fields
- Data audit trail



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Database design

Ensure quality via **data dictionary**:

- List of variables with:
  - Variable names
  - Variable label
  - Data type (e.g. text, date)
  - Format/definitions
  - Conditional dependencies
  - …
- Involve **data manager** and **statistician**

Name	Label	Type	Format
PID	Participant identifier	Integer	XXX
Visit_number	Visit number	Integer	XX
Visit_done	Assessment done	Boolean	1=Yes/2=No
Visit_date	Visit date	Date	dd/mmm/yyyy
Weight_kg_1	Weight (kg)	Number	XXX.X
Weight_kg_2	Weight (kg)	Number	XXX.X
Weight_kg_3	Weight (kg)	Number	XXX.X

Typically generated automatically from databases such as REDCap (Research Electronic Data Capture)

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Database design

- Review **database**
- Review **data dictionary**
- Test data entry
- Test data extraction
- Ensure correct user rights
- Check blinding
- Involve **data manager** and **statistician**

Required Elements		Yes	No	NA	If No, specify action taken or comment
1.	Data specified on case report forms match the database*	☐	☐	☐	
2.	Variables have logical variable labels	☐	☐	☐	
3.	Conditional formatting works as expected	☐	☐	☐	
4.	Variables that contain sensitive information (e.g., name, mobile number) are not included in the database or are flagged as identifier	☐	☐	☐	
5.	Variables that can lead to unblinding are flagged as identifier (can include more than the randomisation list, e.g., a laboratory outcome) (applies to a randomised trial only)	☐	☐	☐	
6.	Database has a data dictionary	☐	☐	☐	
7.	Data can be exported from the database by the statistician	☐	☐	☐	
8.	Exported data has variable labels and formats attached	☐	☐	☐	
9.	Exported data can be successfully imported in Stata or R	☐	☐	☐	
10.	Statistician does not have/will not have access to the randomisation module once the database is in production mode (applies to a randomised trial only)	☐	☐	☐	
11.	Statistician's user rights allow viewing de-identified data only	☐	☐	☐	
12.	Statistician's user rights do not allow viewing randomisation data (in a randomised trial, irrespective of blinding)	☐	☐	☐	
		☐	☐	☐	
		☐	☐	☐	

* This checklist is to be used in conjunction with the CRF checklist.

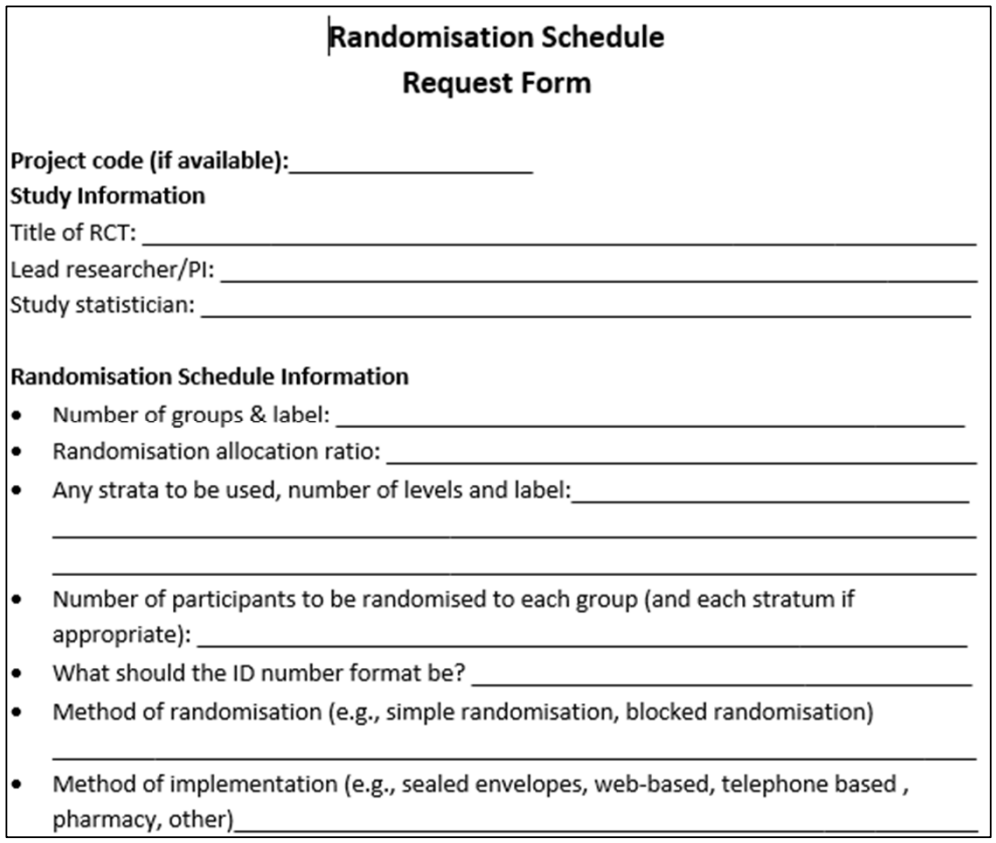
© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Database design

A **statistician** will:

- Design **randomisation** plan
- Generate randomisation list
- Work with drug manufacturer to discuss batches and packaging
- Work with **data manager** or other trial personnel to implement randomised allocation at site



Randomisation Schedule Request Form

Project code (if available):_____

Study Information

Title of RCT: _____

Lead researcher/PI: _____

Study statistician: _____

Randomisation Schedule Information

- Number of groups & label: _____
- Randomisation allocation ratio: _____
- Any strata to be used, number of levels and label:_____
  _____
  _____
- Number of participants to be randomised to each group (and each stratum if appropriate): _____
- What should the ID number format be? _____
- Method of randomisation (e.g., simple randomisation, blocked randomisation)
  _____
- Method of implementation (e.g., sealed envelopes, web-based, telephone based , pharmacy, other)_____

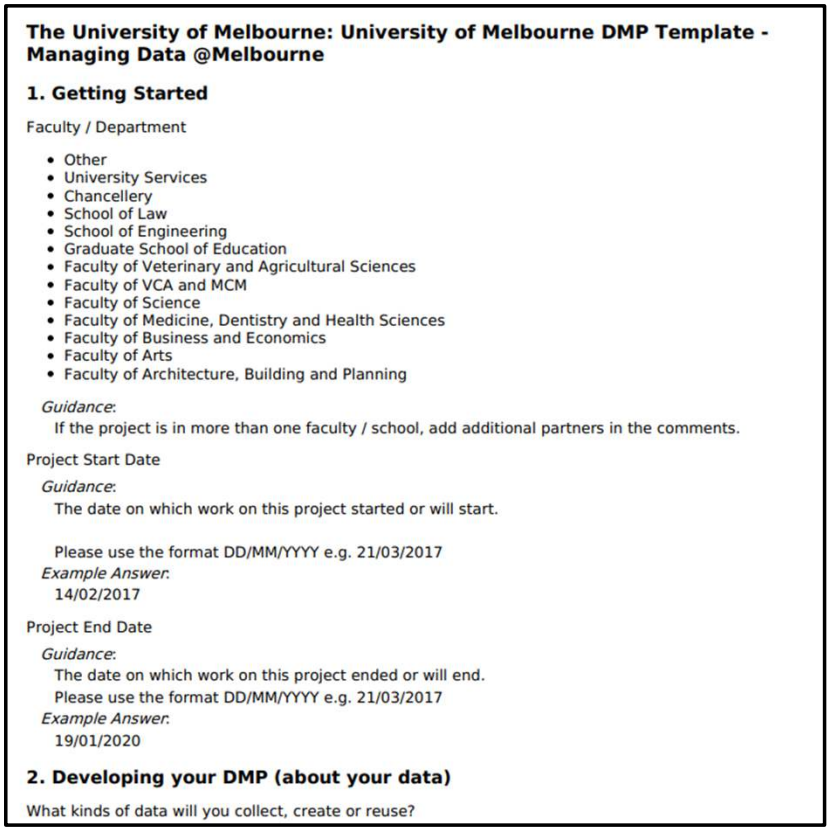
© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Database design

Ensure quality via **data management plan**:

- Description of data collected
- Ethics and legal aspects
- Organising, storing, backing-up
- Sharing, retention, destruction
- Ownership
- Involve **data manager**, with input from the **statistician**



The University of Melbourne: University of Melbourne DMP Template - Managing Data @Melbourne

1. Getting Started

Faculty / Department

- Other
- University Services
- Chancellery
- School of Law
- School of Engineering
- Graduate School of Education
- Faculty of Veterinary and Agricultural Sciences
- Faculty of VCA and MCM
- Faculty of Science
- Faculty of Medicine, Dentistry and Health Sciences
- Faculty of Business and Economics
- Faculty of Arts
- Faculty of Architecture, Building and Planning

Guidance:
If the project is in more than one faculty / school, add additional partners in the comments.

Project Start Date

Guidance:
The date on which work on this project started or will start.

Please use the format DD/MM/YYYY e.g. 21/03/2017

Example Answer:
14/02/2017

Project End Date

Guidance:
The date on which work on this project ended or will end.
Please use the format DD/MM/YYYY e.g. 21/03/2017

Example Answer:
19/01/2020

2. Developing your DMP (about your data)

What kinds of data will you collect, create or reuse?

Source: https://dmp.research.unimelb.edu.au/

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey



- **Source data verification** - typically by staff on site:
  - Random sample of participants
  - Checking data from source to data entered
- **Site monitoring** - typically by staff on site:
  - In-person or remote interaction with site
  - Focus on protocol adherence, process monitoring, and good clinical practice.
- **Central monitoring** - typically by staff in an oversight role:
  - Central review of all data
  - Find outliers or trends in poor performing sites (e.g., visits outside window)

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

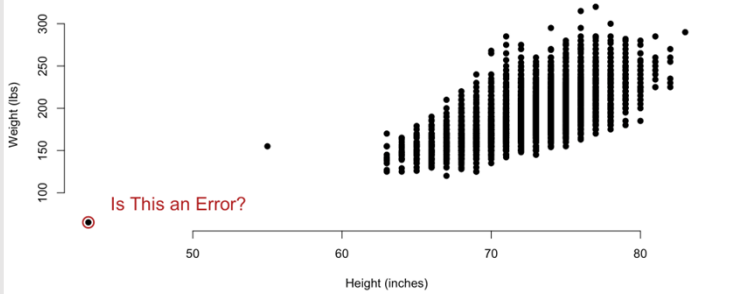
Study conduct

Ensure quality via **data validation plan**:

- Data checks, including missing data, range checks, checks on duplicates, external data checks, logical inconsistencies, and protocol deviations
- Manual and automatic checks
- Involve **data manager**, with input from the **statistician**

Data checks:

- Weight not missing
- Weight >30kg and <200kg
- Scatter plot of weight versus height
- Body mass index >15 and <40kg/m^2
- Absolute monthly weight change <20kg
- % monthly weight change <10%
- Outlying values based on plots



Source: p-062-customized-graphics.knit

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey



- **Medical review** - typically by trial physicians:
  - Targeted review of specific data, safety and efficacy, using medical knowledge
  - Consistency of individual data from a medical perspective
  - Can be supported by summary tables or listings generated by **statistician**

- **Safety monitoring** - typically by Data and Safety Monitoring Board:
  - Targeted review of accumulating summary data using medical knowledge
  - Monitor participant safety risks while trial is still ongoing
  - Reports created by chief investigator, trial manager, **data manager**, **statistician**

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Statistical analysis

Ensure quality via **statistical review**:

– Programming of (blinded) output outlined in **statistical analysis plan**

– Find problematic data impacting validity and interpretability of planned analyses

– Including:
  - Derivation of variables required for analysis
  - Checking underlying model assumptions
  - Test-running statistical models
  - ‘Dry run’ (blinded tables review)

– Typically conducted by **statistician**

Table 1: Baseline characteristics

Characteristic	(N=XX)	Usual care (N=XX)
Age, mean (SD)	XX.X (XX.X)	XX.X (XX.X)
Weight (kg), Mean (SD)	XX.X (XX.X)	XX.X (XX.X)
Height (cm), Mean (SD)	XX.X (XX.X)	XX.X (XX.X)
BMI (kg/m²), Mean (SD)	XX.X (XX.X)	XX.X (XX.X)
Ethnicity, n (%)		
Aboriginal/Torres Strait Islander	XX (XX.X)	XX (XX.X)
Indian/Pakistani	XX (XX.X)	XX (XX.X)
Asian	XX (XX.X)	XX (XX.X)
Māori/Polynesian	XX (XX.X)	XX (XX.X)
African	XX (XX.X)	XX (XX.X)
Caucasian	XX (XX.X)	XX (XX.X)
Other	XX (XX.X)	XX (XX.X)

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey

Statistical analysis

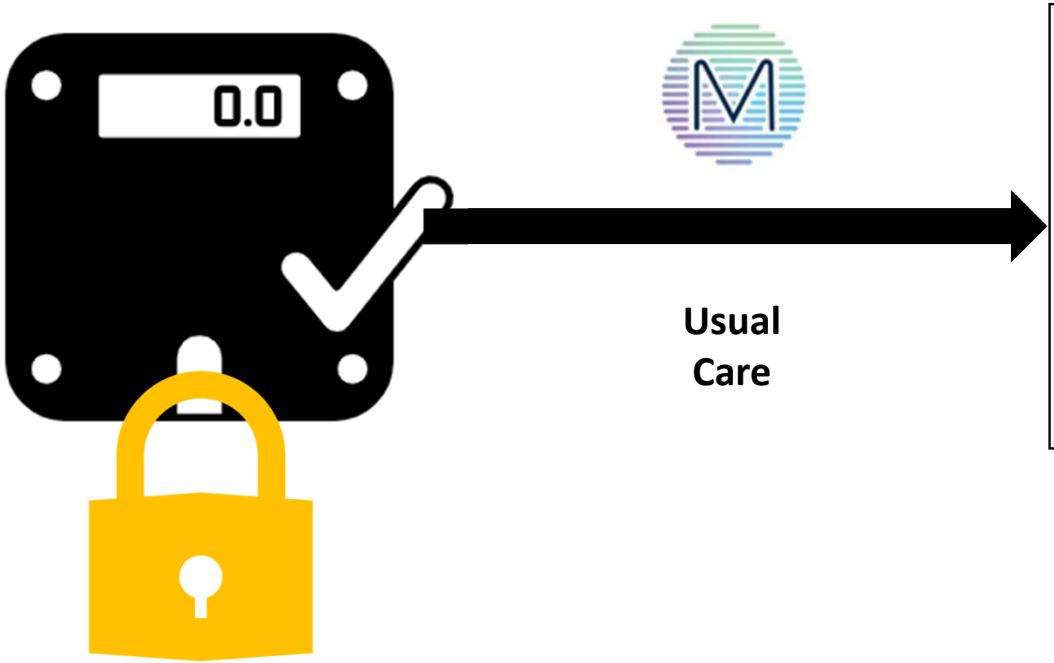
Ensure quality via **blinded data review**:

- All (blinded) data available
- All safety data complete and coded
- All site activities completed
- No outstanding data queries
- Confirm analysis populations
- Statistical analysis plan final
- Typically led by **statistician**



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

The data journey



Usual Care

The NEW ENGLAND JOURNAL of MEDICINE

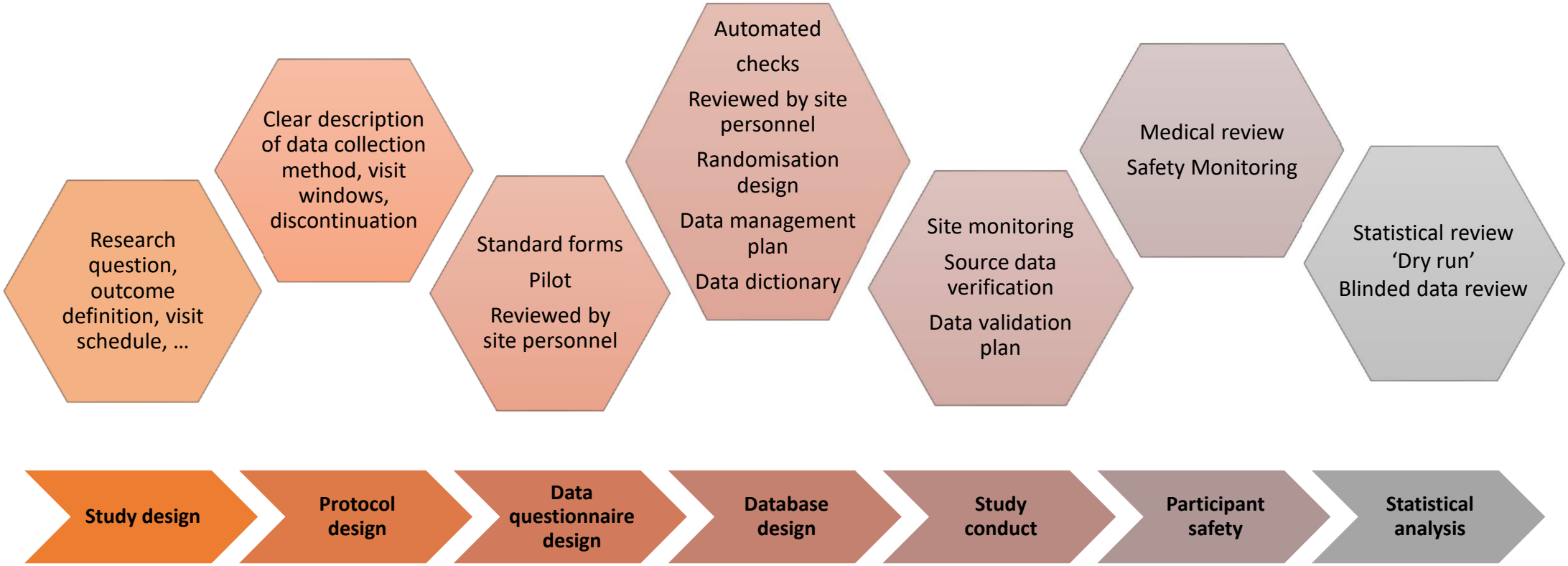
Mean % weight loss *MACH* diet: 20.2%

Mean % weight loss Usual Care: 13.7%

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Data journey from start to finish

Key players: Chief investigator, Trial manager, Site personnel, **Data manager**, **Statistician**



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Data quality issues



Incorrect data

Imprecise data

Incomplete data

Duplicate data

Missing data

Inconsistent data

Unclear data

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

To ensure data quality

- Start early at study design stage
- Have a well-documented data management process to prevent missing data, unclear data, imprecise data, incomplete data, ….
- Collaborate between chief investigator, trial manager, **data manager**, and **statistician**

It is vital to ensure data quality to obtain a high-quality and clean final dataset for statistical analysis and reporting.

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

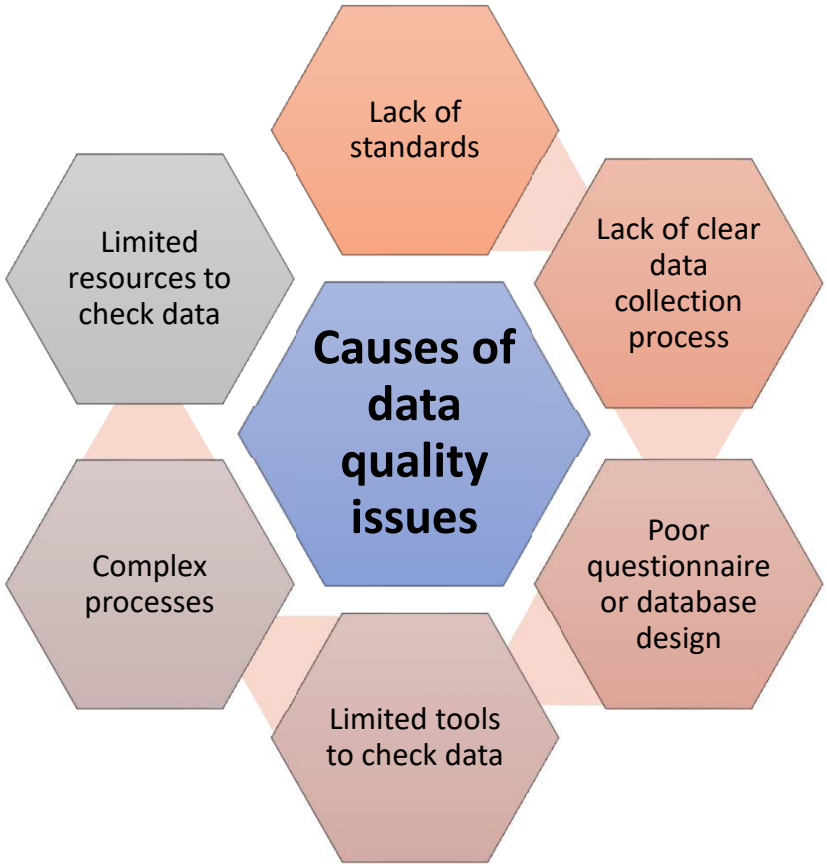
THE UNIVERSITY OF MELBOURNE

MACH EMCR Data Management Seminar

Making Data Work for You Utilising Routinely Collected Electronic Medical record Data
Best Practices in Clinical Research Management

Professor Dougie Boyle
Informatics Head, MISCH Biostatistics Hub
Dept General Practice & Primary Care

Data quality challenges



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Common challenges with routinely collected EMR data

- Biggest Challenge: Missing or incomplete data
- Patients move between care providers like GP surgeries and hospitals depending on circumstance. Despite the promise of MyHealthRecord, data is fragmented
- Collection challenges:
  - Consultation time windows are short – the problem you present with is the problem that will be dealt with – your whole medical history will not be captured or reviewed
  - Time during the consultation is limited – the medical practitioner will concentrate on you – not on clinically coding the entirety of the conversation. Up to 40% of diagnoses are recorded in free text in GP consultations
  - Even simple metrics for example Body Mass Index will not routinely be captured. Unless BMI is a specific issue, why would you measure it in a person who is obviously of an ideal weight? If a patient is morbidly obese and they know it, why would you upset them further by measuring it every time?
  - Some conversations are hard – mental health, early dementia or domestic violence are under reported. Note that a medical record is a legal document. If a patient has recorded dementia, they must notify their car insurance company and DVLA – they may lose their driving license

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Common challenges with routinely collected EMR data

- Challenge: Erroneous data recording
  - Coded data is not always correct. In GP records, the GP will see the free text they entered – they may not immediately see if a wrong code is recorded. Example: Coronary Heart Disease ‘CHD’ being coded by mistake as ‘CVD’ – Cardiovascular Disease
  - In hospitals, clinical coders are employed. Erroneous interpretation of notes can happen. ‘Upcoding’ where a particular code is chosen over another due to a greater re-imbursement for the hospital. More commonly, only a limited number of secondary codes are entered leading to missing data.
  - Free text entry contains huge numbers of spelling mistakes. Interpreting the data and coding it for research is a time-consuming and error-prone enterprise. Double-entry and cross validation of results is often necessary
  - Dates of historic events can be widely variable and dependent on memory e.g. ‘When were you first diagnosed with high blood pressure’?

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Common challenges with routinely collected EMR data

- Other Challenges:
  - Routine EMR data can be HUGE. Patron contains several **Billion** rows of data – normal stats packages and databases cannot handle such analysis
  - Data collection is for a clinical purpose first. Coding systems are not necessarily employed, and coding standards can vary e.g. SNOMED-CT, ICD-10-AM, ACHI
  - Diagnostic thresholds vary over time. When looking at longitudinal data, this can cause issues. Example: Gestational Diabetes Thresholds, 2014 and 2025 modifications
  - Laboratory reference ranges can vary and this can be a challenge with retrospective data analysis over many years and for analyses using results from different pathology labs. The NRL works with labs to calibrate their testing.
  - Researchers and IT providers introduce error when preparing datasets – cleaning data from EMR's is challenging. A typical hospital EMR holds its data across hundreds or thousands of data tables. Can lead to missing data or the wrong interpretation of data in a field or table.
  - Representativeness – is the data you are using representative of the population you are researching? Geographical area, socio-economic status etc.
  - If data is record-linked across multiple databases, false positive record linkage is common (around 1-5 erroneous linkages per 1,000).

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Common challenges with routinely collected EMR data

- Derived data and information may seem consistent – but what assumptions have been made?
  - What decisions have been made with regard to establishing denominators?
  - Data can have complex rules about how simple concepts like a clinical consultation is defined
  - Multiple rows of data may represent a single consultation or none at-all
  - Data may be flagged as deleted but when the data is extracted for use has the deleted flag been retained?
  - What factors at the time of data entry may affect accuracy and data completeness?
  - How can we be sure two repositories are comparable or if two data audit tools produce different answers, who is correct (if anybody)

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

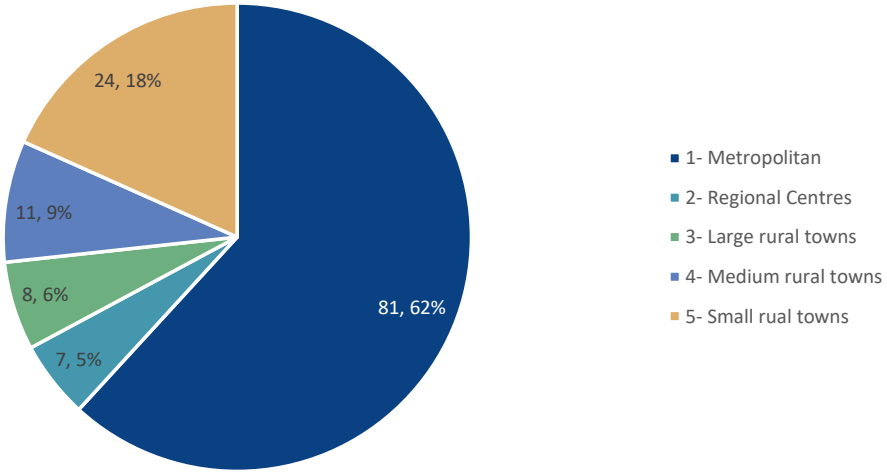
Data for Decisions and the Patron program of research

Creating knowledge from primary care data through research

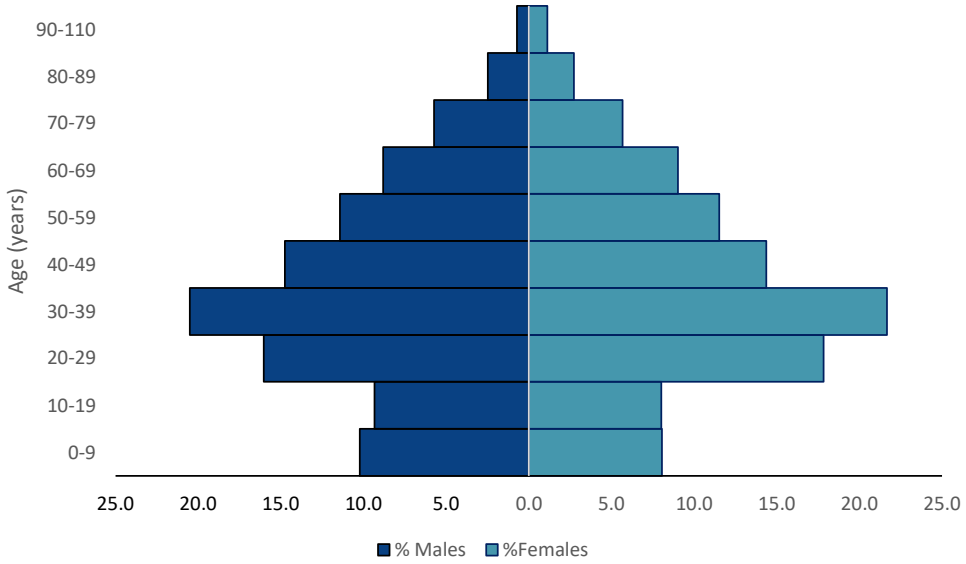
Total patient count: 5,042,008 Active patient count: 1,876,892

Adults with type 2 diabetes or pathology consistent with type 2 diabetes: 42,086

Number of practices in Patron by MMM



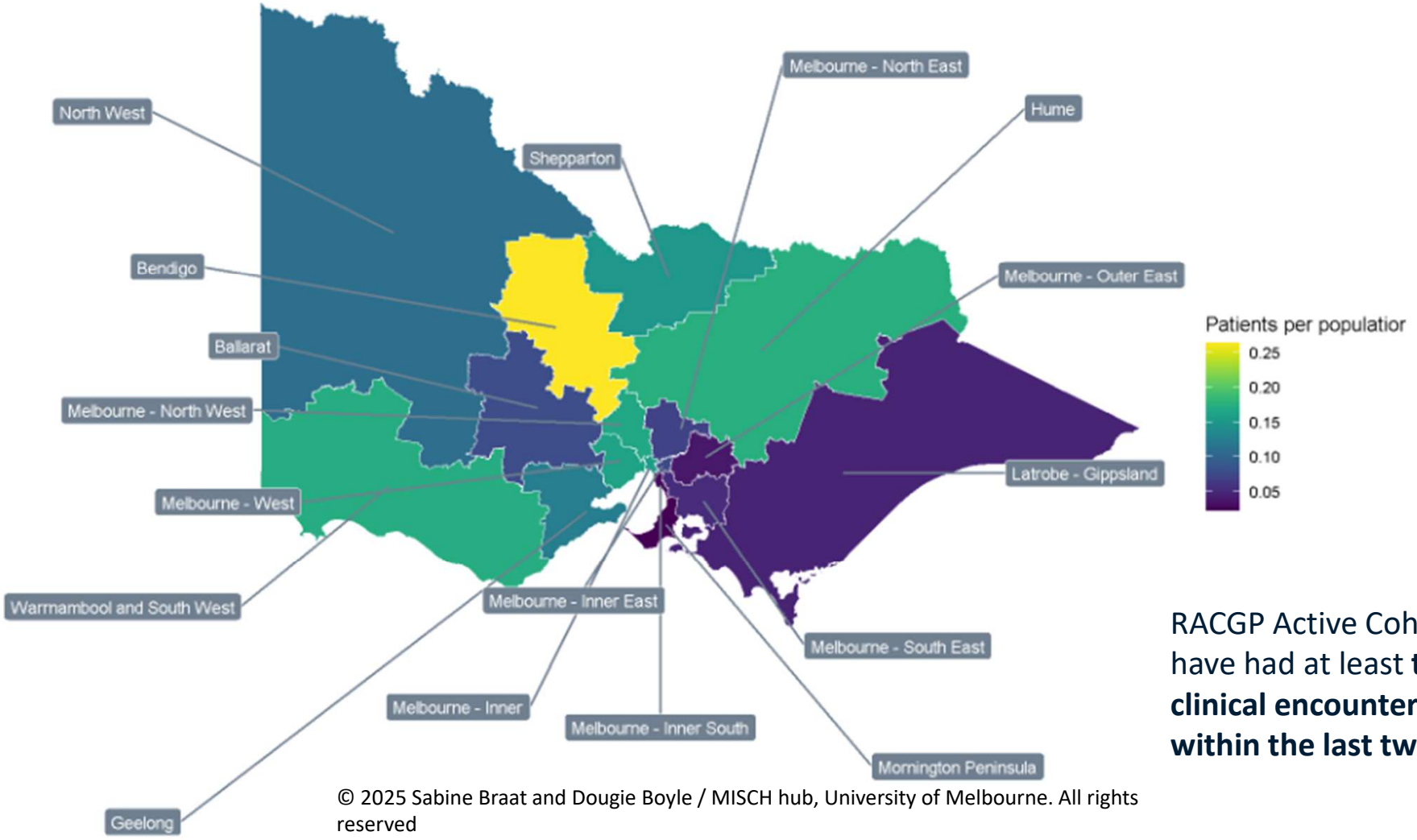
Sex and age distribution



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

https://medicine.unimelb.edu.au/school-structure/general-practice-and-primary-care/research/research-groups/data-for-decisions

RACGP active patients per population (SA4)

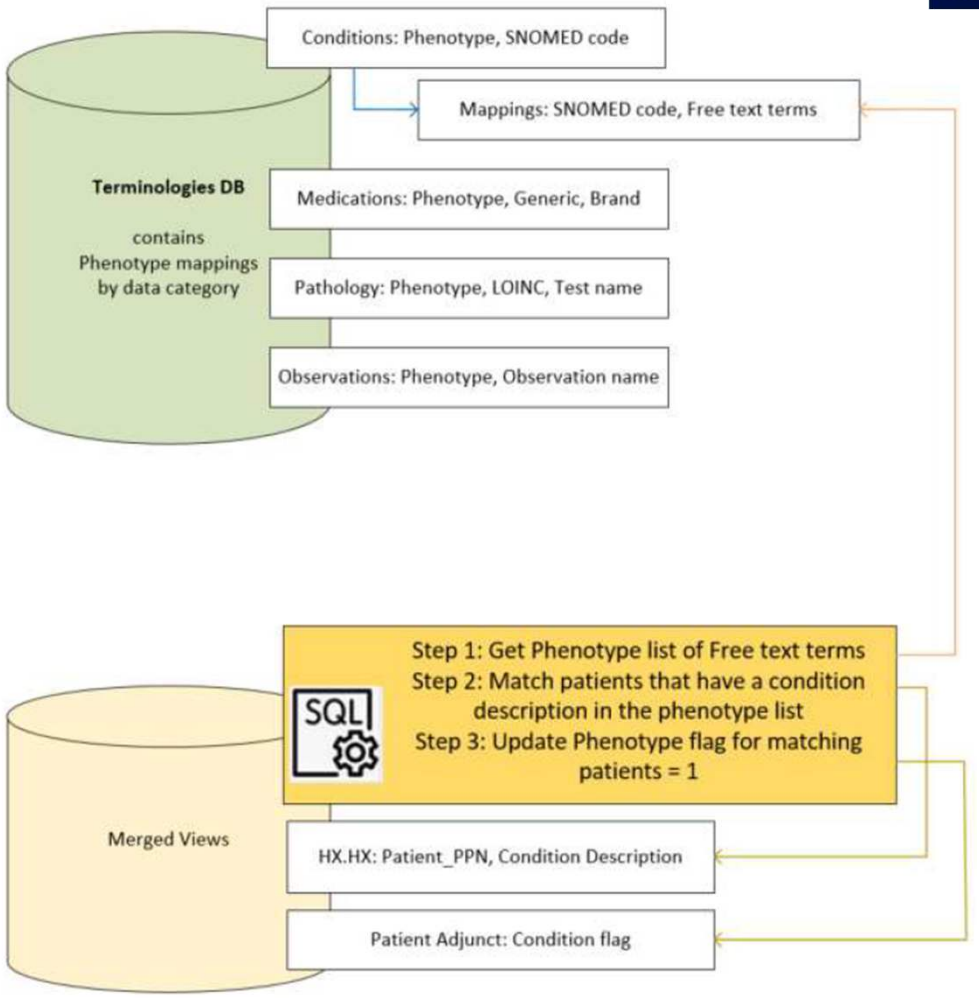


© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

RACGP Active Cohort: Patients who have had at least **three visits or clinical encounters at a practice within the last two years**

Accurate GP Data Phenotyping Essential

- **Medications**
- **Pathology**
- **Observations**



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Example validated, pre-summarised phenotypes

Conditions

Hepatitis_G	Hepatitis G
Hepatitis_ISC_Acute	Hepatitis Ischeamic - Acute
Hepatitis_Mumps	Hepatitis Mumps
Hepatitis_Resolved	Hepatitis - Resolved
Hepatitis_UN	Hepatitis Unkown
Hepatitis_UN_Acute	Hepatitis Unknown - Acute
Hepatitis_UN_Chronic	Hepatitis Unknown - Chronic
HF	Heart Failure
HT	Hypertension
IFG	Impaired Fasting Glucose
IGT	Impaired Glucose Tolerance
LADA	Latent Autoimmune Diabetes in Adults
Liver_Cancer	Malignant tumour of prostate
Lung_Cancer	Primary malignant neoplasm of respiratory tract
MI	Myocardial Infarction
MODY	Maturity-Onset Diabetes of the Young
Obesity	Obesity
Oesophageal_varices	Oesophageal varices
Ovarian_Cancer	Malignant tumor of ovary
PBC	Primary biliary cholangitis
PCI	Percutaneous Coronary Intervention
PCOS	Polycystic ovary syndrome
PIH	Pregnancy-induced Hypertension
Prediabetes	Prediabetes
Prostate_Cancer	Malignant tumour of prostate
Renal_Dialysis	Renal Dialysis
Renal_Transplant	Renal Transplant
Stroke_HEM	Haemorrhagic Stroke

Medications

GC_Inhaled	Glucocorticoids - Inhaled
GC_Oral	Glucocorticoids - Oral
GC_Topical	Glucocorticoids - Topical
GLP1	Glucagon-Like Peptide-1 Receptor Agonist
Glucocorticoids	Glucocorticoids
Heparin	Heparin and related
HF_specific	Heart Failure Specific
Insulin	Insulin
Loop_Diuretic	Loop Diuretic
MAOIs	Monoamine oxidase inhibitors (Antidepressants)
MCRA	Potassium Sparring Diuretic - Mineralocorticoid Receptor Antagonist
Metformin	Metformin
Mineralocorticoid	Mineralocorticoid
Mood_Stabiliser	Mood Stabiliser [Lithium]
Potassium_sparring	Potassium Sparring Diuretics
SGLT2	SGLT2 Inhibitor
SNRIs	Serotonin and norepinephrine reuptake inhibitors (Antidepressants)
SSRIs	Selective serotonin reuptake inhibitors (Antidepressants)
Statin	Statin (lipid-lowering)
Sulfonylurea	Sulfonylurea
TCAs	Tricyclic antidepressants (Antidepressants)
Thiazides	Thiazide Diuretics
Thrombolytic	Thrombolytic drugs
TZD	Thiazolidinedione (Glitazones)
Vasodilator	Vasodilator
Warfarin	Warfarin

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Undercounts in standard reporting where coded GP database entries only are relied-on

Table 1 Variance in counts between coded only and free-text plus coded diagnoses in general practice medical records, *N* = 456,125 'active' patients*

	A	B	C	D	E	F
Disease	**Patients with a coded diagnosis**	**Patients with a clinically validated free-text diagnosis (with or without a code)**	**Patients with a clinically validated free-text diagnosis but no coded diagnosis. (B-A = undercount)**	**% Undercount**	**Z value**	***P* value**
Asthma	46,853	74,038	27,185	36.72	156.18	< 0.001
Chronic kidney disease	8303	10,721	2418	22.55	17.72	< 0.001
Chronic obstructive pulmonary disease	7774	9573	1799	18.79	13.79	< 0.001
Dementia	2153	2525	372	14.73	5.45	< 0.001
Type 1 diabetes	2033	2112	79	3.74	1.23	0.219
Type 2 diabetes	23,264	23,877	613	2.57	2.90	0.004
All diagnosis types **	90,380	122,846	32,466	26.4	80.32	< 0.001

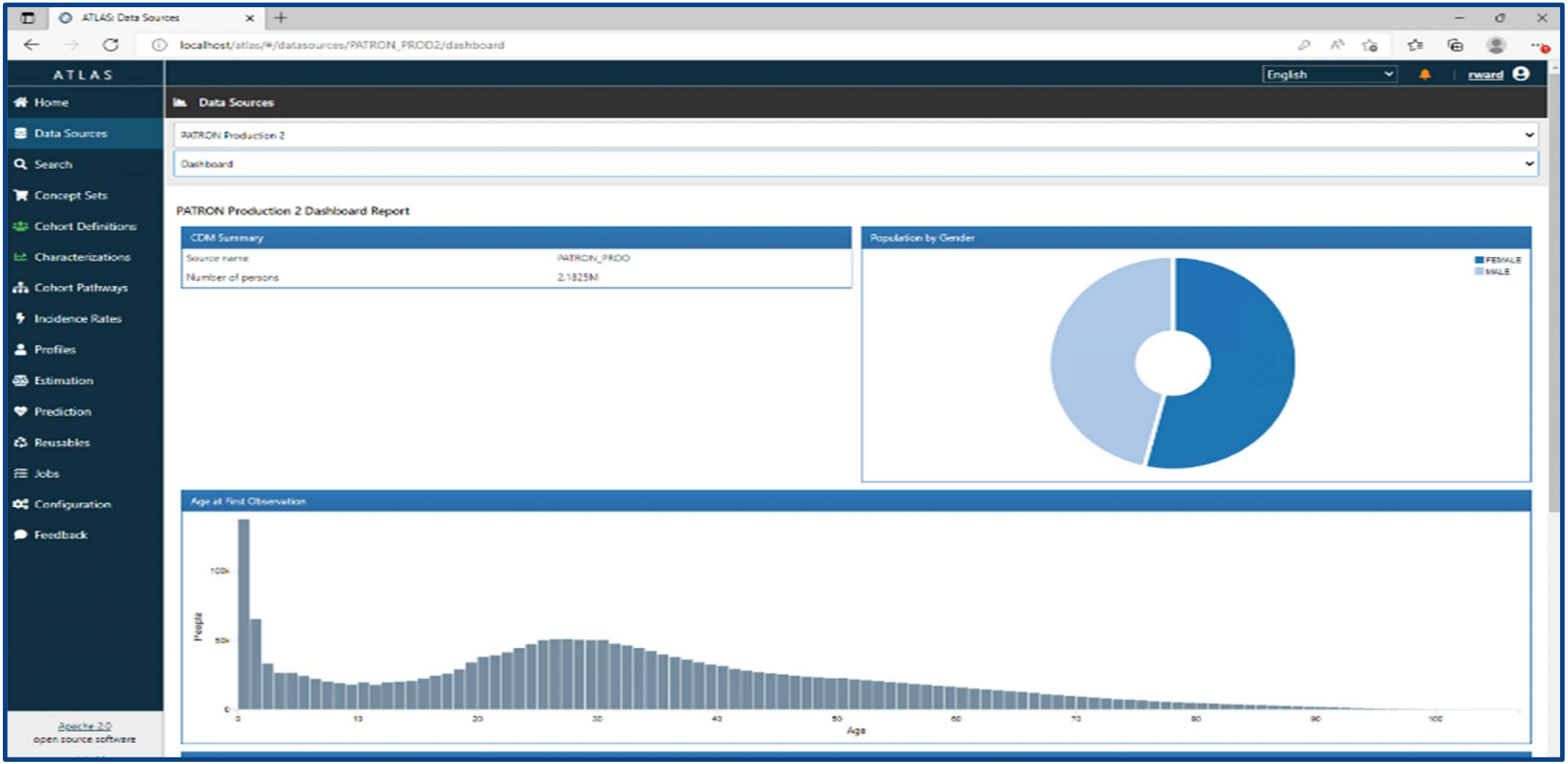
* Active as per the RACGP definition, i.e. a patient who has had a general practice consultation three or more times in the past two years [25], which in this instance was between 1 January 2021 and 31 December 2022

** Patients are counted only once per disease, but the total for 'All diagnosis types' will include patients multiple times if they have more than one disease

Canaway, R., Chidgey, C., Hallinan, C.M. *et al.* Undercounting diagnoses in Australian general practice: a data quality study with implications for population health reporting. *BMC Med Inform Decis Mak* **24**, 155 (2024). https://doi.org/10.1186/s12911-024-02560-w

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

OMOP via ATLAS – Snapshot of Patron Data – June 2022



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Patron OMOP via ATLAS – Data Snapshot– June 2022

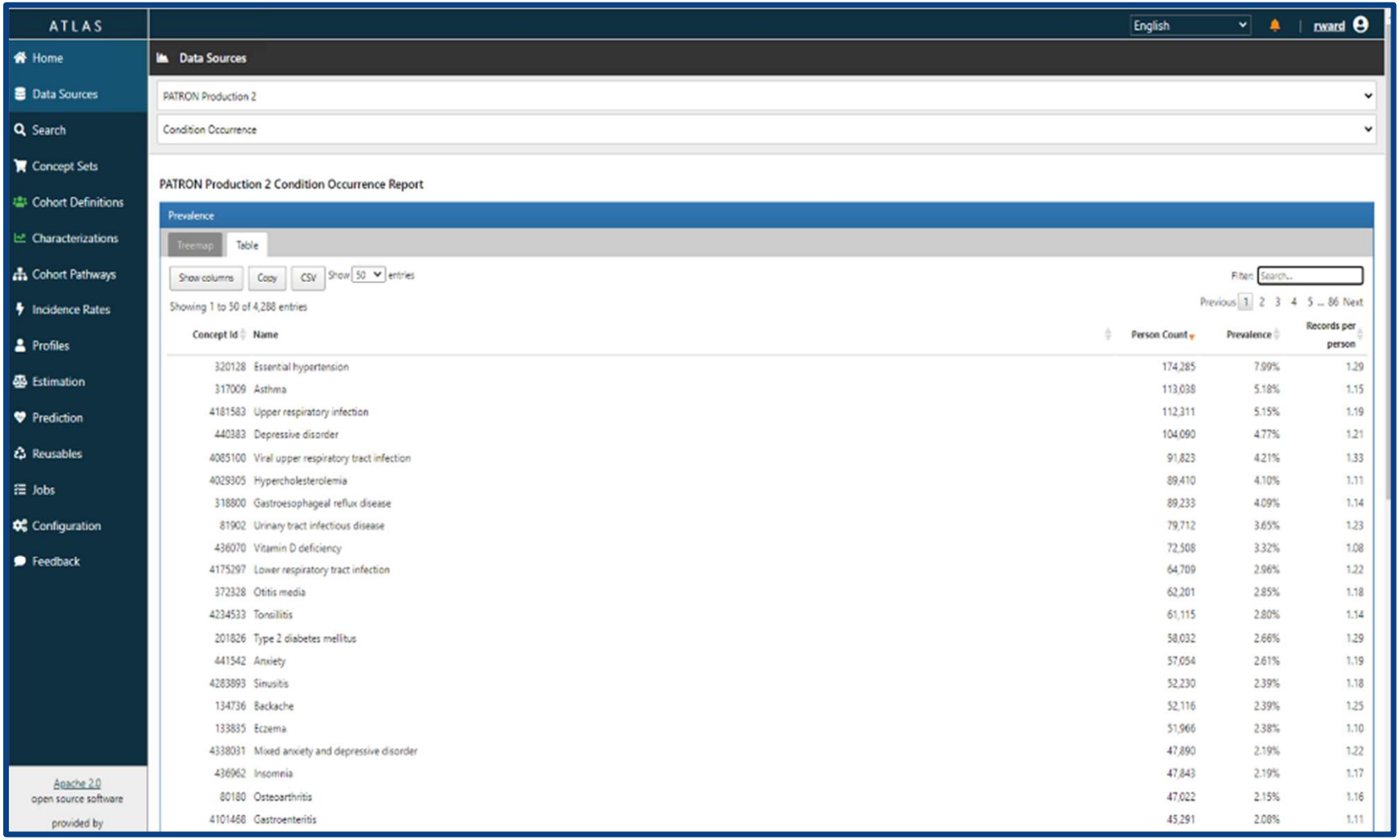
Highlighted in green on left-hand navigation

Cohort Definitions
Construct a cohort based on inclusion and exclusion criteria

Characterisations
Summarise characteristics of cohort/s

Output
Provides a condition occurrence report for the whole of Patron OMOP database

Top 6 conditions
Hypertension, asthma, Upper respiratory tract infections (URTI), depressive disorder, viral upper respiratory tract infection, hypercholestremia



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Patron OMOP via ATLAS – Data Snapshot – June 2022

Highlighted in green on left-hand navigation

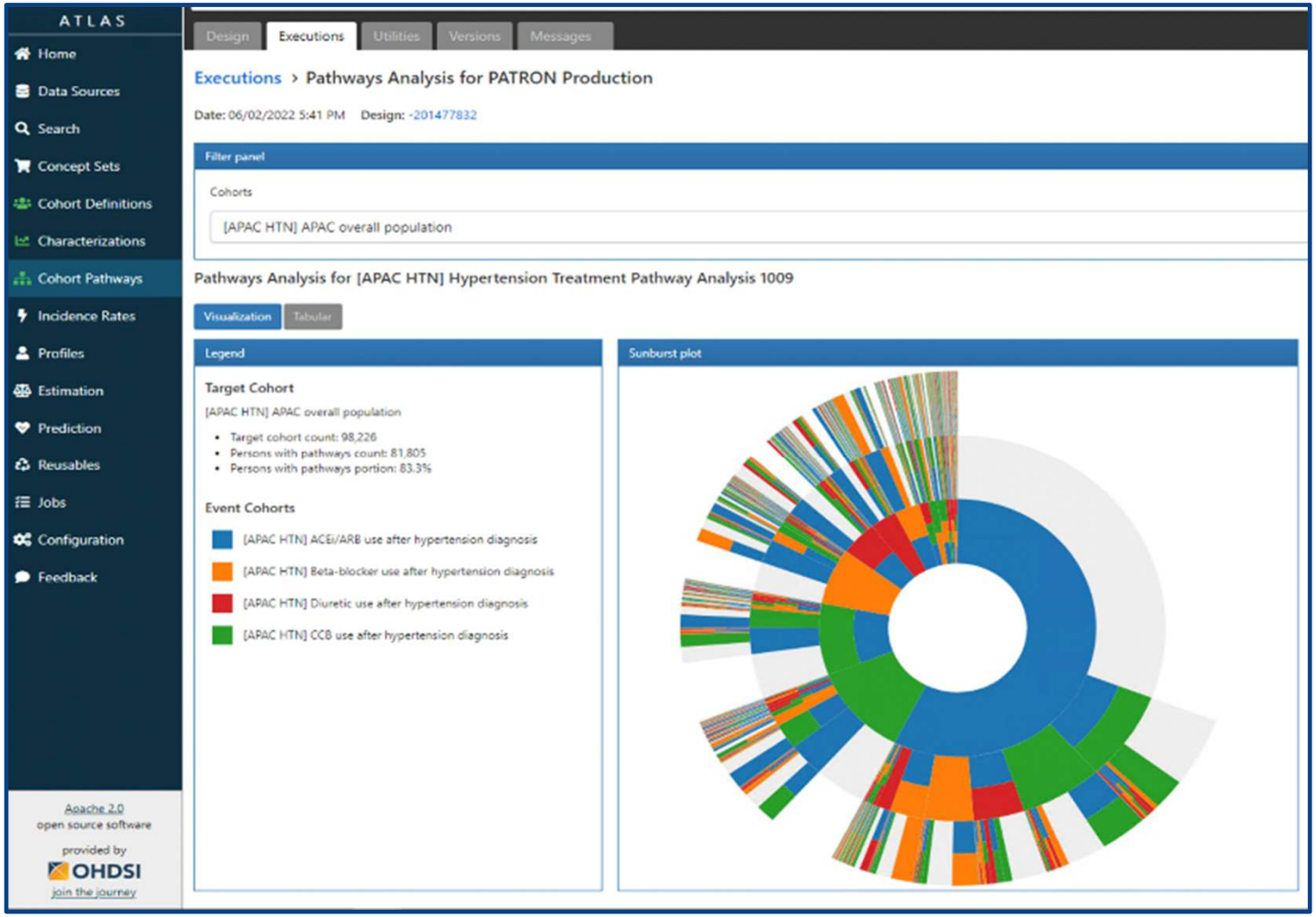
Cohort Definitions
Asia Pacific Community (Australia, China, Japan, Singapore, South Korea, Taiwan)

Characterisations
Hypertension

Cohort Pathways
Hypertension ACE/ARB, Beta-blocker, Diuretic, and Calcium Channel-blocker treatment pathway

Output
Provides cohort treatment pathways across time.

Top treatment pathway
Most common first line therapy for the whole cohort is ACE/ARBs
These plots can be split to demonstrate treatment pathways for each of the different countries in the study.



© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

eGEMs
Generating Evidence & Methods to improve patient outcomes

A Harmonized Data Quality Assessment Terminology and Framework for the Secondary Use of Electronic Health Record Data

Michael G. Kahn, MD, PhD;[i] Tiffany J. Callahan, MPH;[i] Juliana Barnard, MA;[i] Alan E. Bauck;[ii] Jeff Brown, PhD;[iii] Bruce N. Davidson, PhD;[iv] Hossein Estiri, PhD;[v] Carsten Goerg, PhD;[i] Erin Holve, PhD, MPH, MPP;[vi] Steven G. Johnson, MS;[vii] Siaw-Teng Liaw, MBBS, PhD, FRACGP, FACHI;[viii] Marianne Hamilton-Lopez, PhD, MPA;[ix] Daniella Meeker, PhD;[x] Toan C. Ong, PhD;[xi] Patrick Ryan, PhD;[xii] Ning Shang, PhD;[xiii] Nicole G. Weiskopf, PhD;[xiv] Chunhua Weng, PhD, FACMI;[xiii] Meredith N. Zozus, PhD;[xv] Lisa Schilling, MD[xi]

ABSTRACT

http://repository.edm-forum.org/egems/vol4/iss1/18

Objective: Harmonized data quality (DQ) assessment terms, methods, and reporting practices can establish a common understanding of the strengths and limitations of electronic health record (EHR) data for operational analytics, quality improvement, and research. Existing published DQ terms were harmonized to a comprehensive unified terminology with definitions and examples and organized into a conceptual framework to support a common approach to defining whether EHR data is 'fit' for specific uses.

THE UNIVERSITY OF MELBOURNE

Table 2. Crosswalk Between Harmonization Terminology, Categories, and Subcategories Versus Pre-Existing Categories and Frameworks (Cont'd)

PROPOSED CATEGORIES	KAHN 2012 [44f]	NAHM 2012 [57g]	MCGILVRAY 2008 [64h]	EPPLER 2006 [65i]	WANG 1996 [42j]
CONFORMANCE					
Value	• Attribute Domain Constraints • Historical Data Rules • State-Dependent Object Rules	• Granularity • Precision	• Data Integrity Fundamentals		• Representational Consistency
Relational	• Relational Integrity Rules	• Attribution	• Data Specifications		
Computational	• Attribute Dependency Rules				
COMPLETENESS					
Completeness	• Attribute Domain Constraints	• Completeness	• Data Integrity Fundamentals		• Completeness
PLAUSIBILITY					

Data quality example - Diagnosis

- **Qualitative assessment**: GPs can enter the diagnosis from a drop down list with search. However, if they cannot find a suitable term they can enter short strings of free text.
- **Harmonised framework assessment**:

	Context	
Dimension	**Verification**	**Validation**
Value Conformance	% of Diagnoses using internal codes	% of Diagnoses using SNOMED CT
Relational Conformance	N/A	N/A
Computational Conformance	N/A	N/A
Completeness	% of Diagnoses that are not null/missing	Number of diagnoses per episode concordant with international reference
Uniqueness Plausibility	N/A	N/A
Atemporal Plausibility	% of sex-specific diagnoses with concordant sex	Distribution of chronic diagnoses follows expected distribution for similar populations
Temporal Plausibility	Date of 1st diagnosis is => date of first encounter	Diagnosis = URI counts follow expected seasonal patterns

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Data quality example – Reason for prescription

- **Qualitative assessment**: Reason for prescription (RfP) is not a mandatory field for all prescriptions. GPs can also 'pull' the reason for visit and re-use it as a reason for prescription.
- **Harmonised framework assessment**:

	Context	
Dimension	**Verification**	**Validation**
Value Conformance	% of Diagnoses using internal codes/structured data	% of reasons for prescription using SNOMED CT
Relational Conformance	N/A	N/A
Computational Conformance	N/A	N/A
Completeness	% of RfP that are not null/missing	% of RfP that are null/missing is comparable to expected values
Uniqueness Plausibility	N/A	N/A
Atemporal Plausibility	% of prescriptions for sex-specific conditions that are concordant	Frequency of
Temporal Plausibility	Prescriptions for 'Diabetes' do not occur before diagnoses for Diabetes	Prescriptions for 'Upper respiratory infection' follow a seasonal pattern

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

OHDSI Data Quality Dashboard

Kahn – Compatible DQ Metrics



The OHDSI data quality dashboard (DQD) produces metrics on data Plausibility, Conformance and Completeness. It performs over 3,000 checks on the data.
The DQD tool goes table by table and field by field to quantify the number of records in a CDM that do not conform to the given specifications. In all, over 3,570 checks are performed, each one organized into the Kahn framework (Kahn, Callahan et al. 2016). For each check the result is compared to a threshold whereby a FAIL is any percentage of violating rows falling above that value. The results on the primary care OMOP database are presented below (Table 6).

	Verification				Validation				Total			
	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass	Pass	Fail	Total	% Pass
Plausibility	1982	53	2035	97%	285	2	287	99%	2267	55	2322	98%
Conformance	746	30	776	96%	157	0	157	100%	903	30	933	97%
Completeness	289	14	303	95%	6	6	12	50%	295	20	315	94%
Total	3017	97	3114	97%	448	8	456	98%	3465	105	3570	**97%**

© 2025 Sabine Braat and Dougie Boyle / MISCH hub, University of Melbourne. All rights reserved

Thank you

- :-https://clinicalresearch.mdhs.unimelb.edu.au/
- :- misch-info@unimelb.edu.au
- :- misch-hub

