Use of SDTM and ADaM in the RECOVERY trial

Dr Will Stevens, PhD

Senior Programmer/Analyst

Clinical Trial Service Unit and Epidemiological Studies Unit

University of Oxford

will.stevens@ndph.ox.ac.uk

My background

- Software development in industry and academia for 25 years.
- Clinical trials software development for 14 years.
- Software to support SDTM, ADaM, define.xml for the past 9 years.
- Large simple clinical trials – large number of participants, small amount of data per participant, hard outcomes.

RECOVERY trial

https://www.recoverytrial.net

- PIs: Prof. Martin Landray and Prof. Peter Horby.
- Randomised controlled trail of treatments for COVID-19.
- Hospitalised COVID-19 patients with consent are randomly allocated to one of a number of possible treatments. 20 treatments – multifactorial.
- First participant randomised on 19th March 2020.
- Over 47000 people randomised to date.
- RECOVERY showed that dexamethasone (June 2020), tocilizumab (Feb 2021), Regeneron's monoclonal antibody combination (June 2021), and baricitinib (March 2022) reduce mortality.
- Aspirin, Azithromycin, Colchicine, Convalescent Plasma, Hydroxychloroquine and Lopinavir-Ritonavir have no effect.

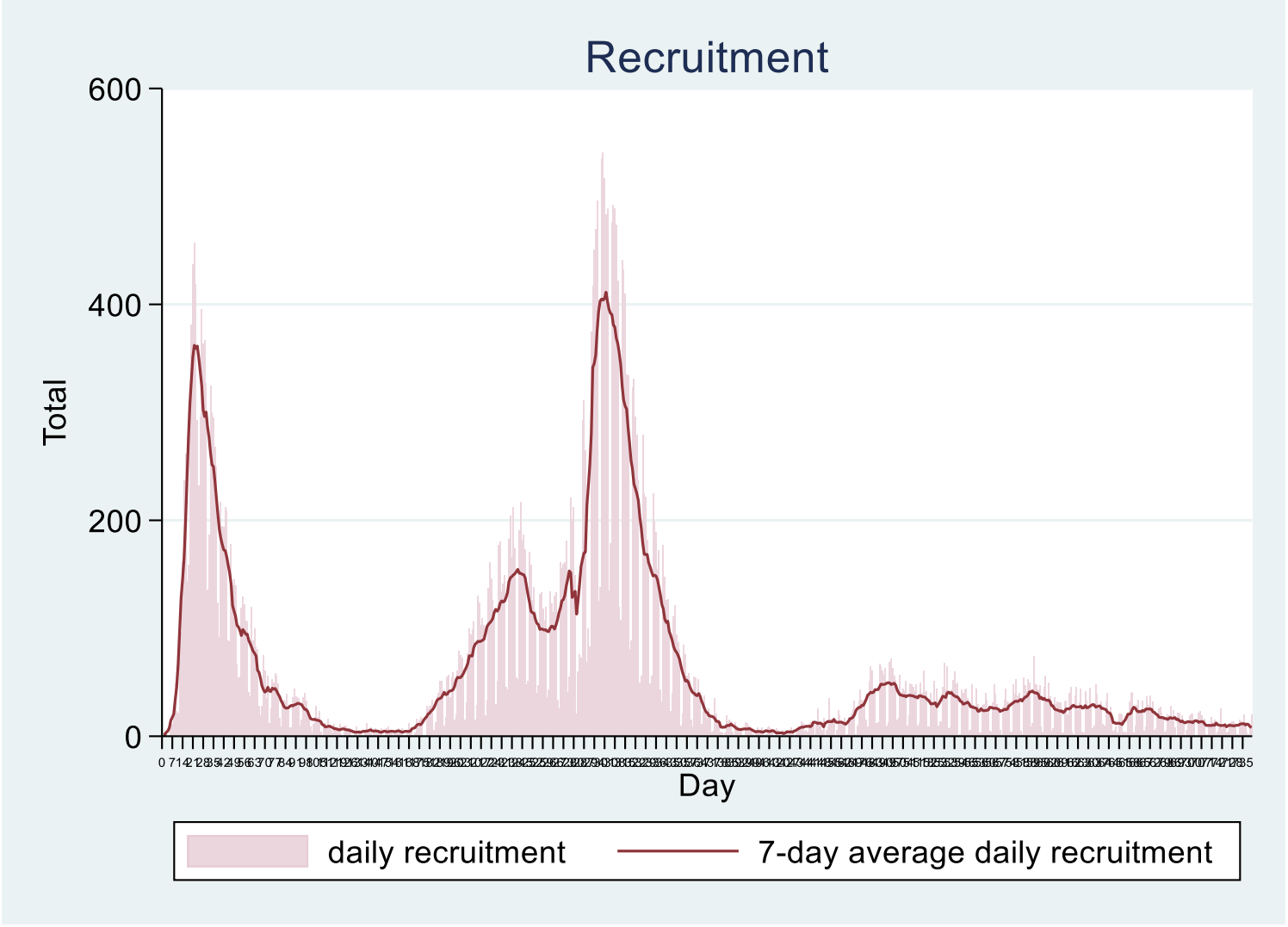
RECOVERY trial

https://www.recoverytrial.net

- Data mapped to SDTM and ADaM.
- Took part in task force for the Interim User Guide for COVID-19 published on 20 April 2020. RECOVERY CRFs and Protocol were shared with the team who produced this.

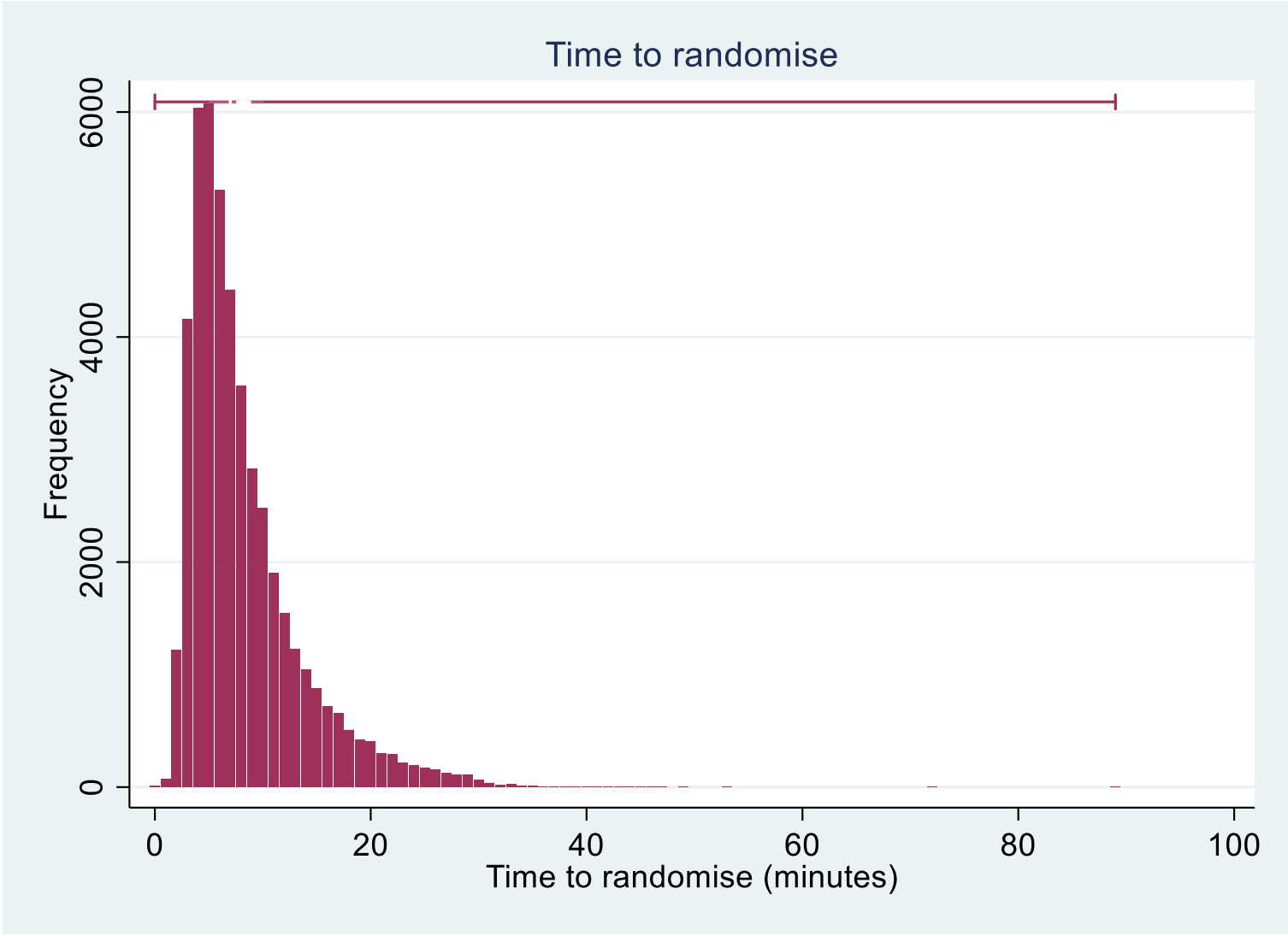
Recruitment over time

RECOVERY
Randomised Evaluation of COVID-19 Therapy



Time to randomise

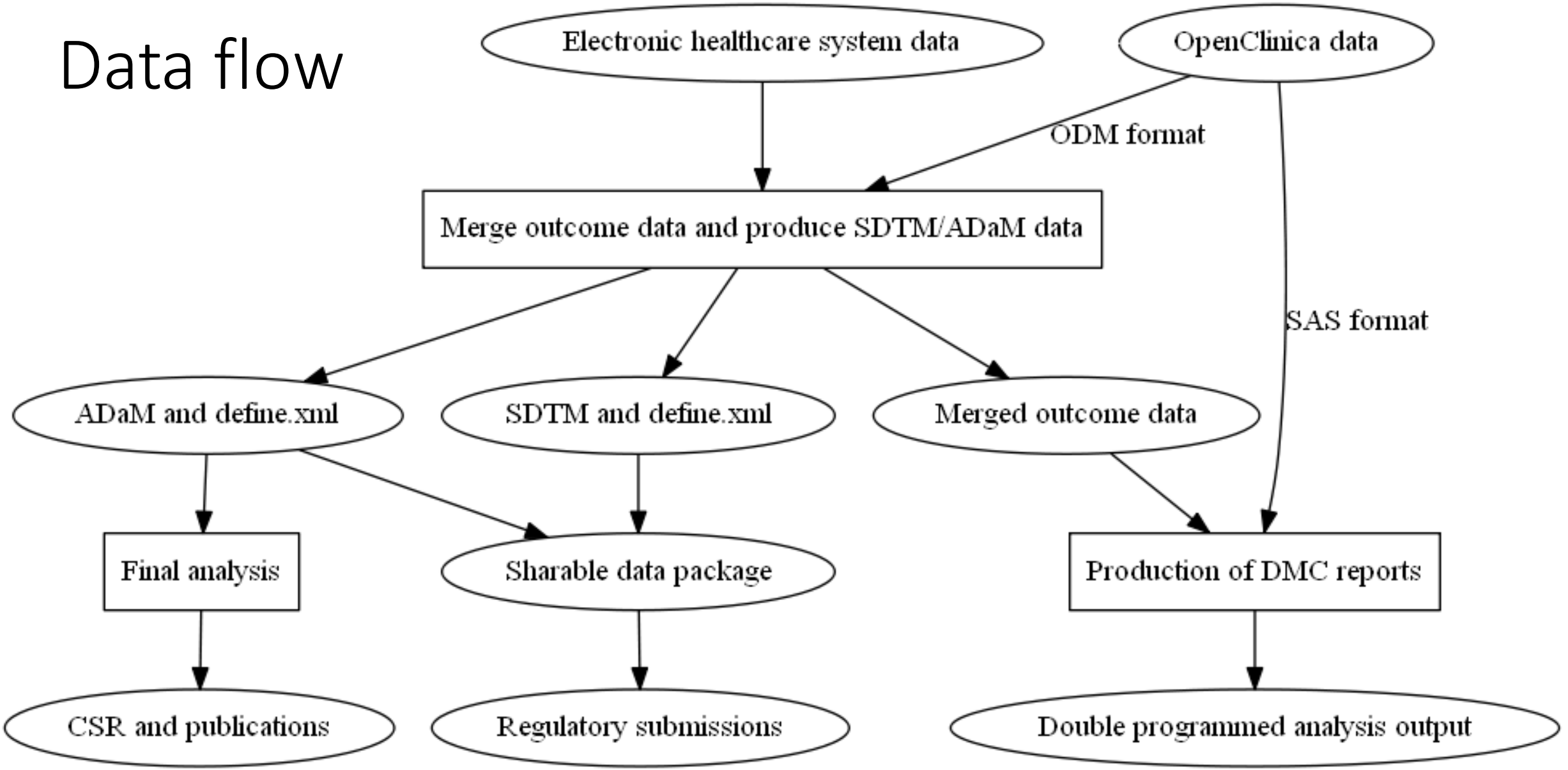
RECOVERY
Randomised Evaluation of COVID-19 Therapy



Data sources

- Data comes from web-based CRFs, NHS Digital, and other electronic sources.
- NHS Digital and other electronic sources are pre-processed by data engineer on receipt by RECOVERY to extract the relevant data – major piece of work.
- Data quality and completeness from each sources varies.
- Different pieces of information about the same outcomes from each source.
- RECOVERY main outcomes are death within 28 days of randomisation, time to discharge from hospital, use of ventilation, use of renal replacement therapy.

Data flow


Electronic healthcare system data
OpenClinica data
ODM format
Merge outcome data and produce SDTM/ADaM data
SAS format
ADaM and define.xml
SDTM and define.xml
Merged outcome data
Final analysis
Sharable data package
Production of DMC reports
CSR and publications
Regulatory submissions
Double programmed analysis output

Data sources - examples

- Personal Demographics Service – first indicator that somebody has died, deaths may appear in this dataset very soon after person dies. Death date may be corrected later on (quality indicator for this).
- Civil registration death data – data covering everybody who dies in England and Wales. Also includes cause of death (ICD10). Coverage is incomplete for those who died in the month prior to provision date.
- CRF – completed by investigator in hospital. Only captures in-hospital deaths and therefore misses about 6% of all deaths. Also includes high-level cause of death classification.
- 10 possible sources for death date in total.

Data sources – description and definitions

- High level document (available on the trial website) describes all of the data sources used, and informally defines how outcomes are defined using the data sources.
- More detailed definition document (an appendix to SDRG and ADRG) describes how data from different sources maps to SDTM, and how outcomes are defined in terms of the SDTM data.
- Annotated CRF.

Example – death data

6.1 All-cause mortality

The primary outcome is all-cause mortality at 28 days after randomisation. All-cause mortality will also be assessed at 6 months and other later time points.

6.1.1 Sources

Information on death may come from the following sources:

- FU eCRF (for deaths within first 28 days after randomisation)
- PDS (for participants in England)
- PDS Wales ((or participants in Wales)
- SUSAPC (for participants in England)
- SMR01 (for participants in Scotland)
- PEDW (for participants in Wales)
- ONS mortality data (for participants in England and Wales)
- NRS mortality data (for participants in Scotland)

Example – death data

Outcomes		
	SAP(2.0) 5.1.1	Mortality (all-cause) within 28 days after randomisation (i.e. if randomisation = day 1, th
	SAP(2.0) 6.1.1.1	Cause-specific mortality
Relevant SDTM and ADaM datasets		
	DD (death details), DM (demograpics)	
	ADCNDTH (all-cause mortality censoring data), ADDTHTTE (all-cause mortality time-to-event)	
	ADCNDCS (cause-specific mortality censoring data), ADDCSTTE (cause-specific mortality time-to-event)	
	Censoring rules are described in the define.xml documentation for the censoring datasets	
	Time-to-event tables are structured according to the ADaM TTE guidance document	

Data sources and SDTM mapping notes

A row is created in DD from each of the following data sources. DDSPID identifies the data source.

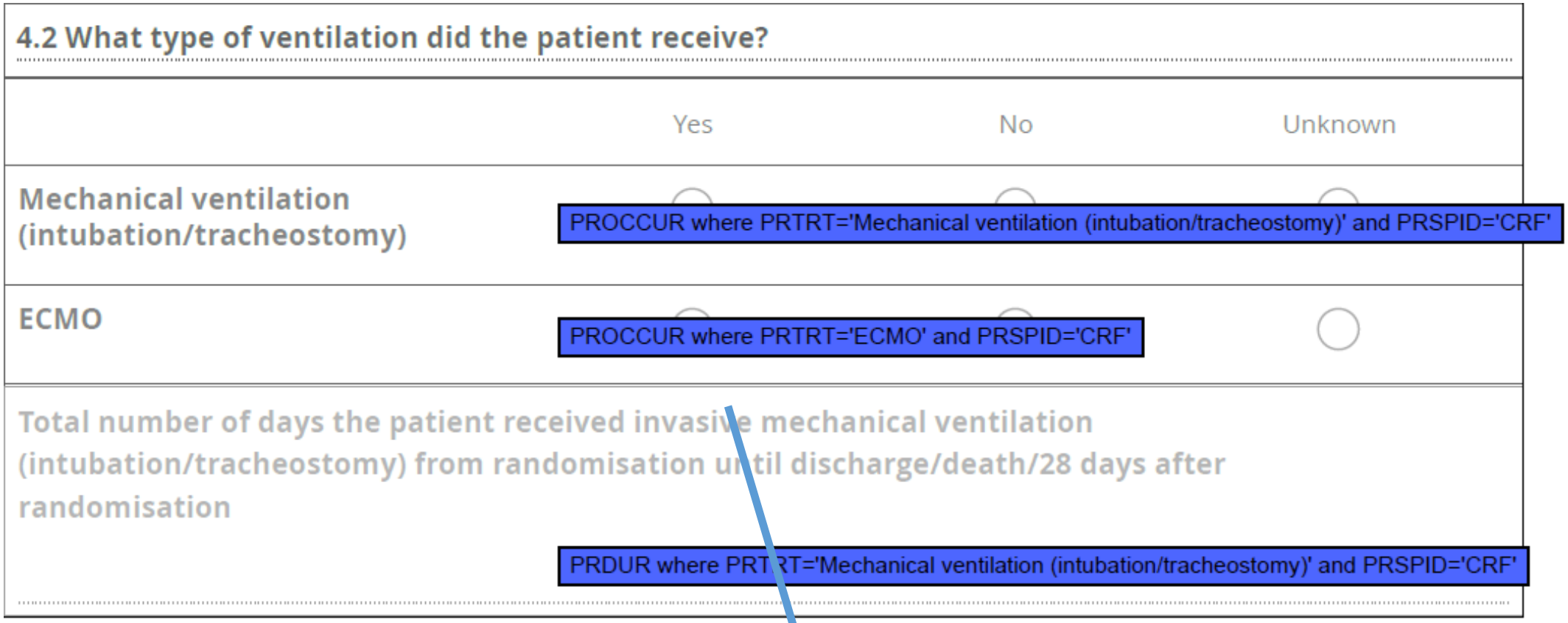
INT04 Civil registration	Date of death: DDTESTCD='DTDTH', DDORRES = date of death. ICD10 code cause of death
INT03 PDS formal	Date of death: DDTESTCD='DTDTH', DDORRES = date of death, DDSPID='INT03', DDRESCA
INT12 NHSCR PDS	Date of death: DDTESTCD='DTDTH', DDORRES = date of death. ICD10 code cause of death
INT10 SMR01	Date of death: DDTESTCD='DTDTH', DDORRES = date of death, DDSPID='INT10'
INT14 PEDW	Date of death: DDTESTCD='DTDTH', DDORRES = date of death, DDSPID='INT14'
INT01 SUS+	Date of death: DDTESTCD='DTDTH', DDORRES = date of death, DDSPID='INT01'
CRF	Date of death: DDTESTCD='DTDTH', DDORRES = date of death. Cause of death: DDTESTCD
INT03 PDS informal	Date of death: DDTESTCD='DTDTH', DDORRES = date of death, DDSPID='INT03', DDRESCA
INT19 Welsh Demographic Service	Date of death: DDTESTCD='DTDTH', DDORRES = date of death. DDSPID='INT19'
INT13 NHSCR death registry	Date of death: DDTESTCD='DTDTH', DDORRES = date of death. ICD10 code cause of death

Example – identifying and handling disagreement

- Some sources are extremely reliable and considered the gold standard for the data we use - e.g. civil registration death data.
- But have a time lag, so more up-to-date data is needed: CRF, personal demographics service.
- All data sources can be assessed for accuracy against the gold standard source.
- Ranked in order of accuracy, and used preferentially in that order.
- Even the least reliable sources are better than no data.

Example – ventilation data

Outcomes		
	SAP(2.0) 5.1.2.2 (component of)	Invasive mechanical ventilation within the 28 days on and after randomisation
	SAP(2.0) 5.1.3.1 (component of)	Use of ventilation (overall and by type)
	SAP(2.0) 5.1.3.2 (component of)	Duration of invasive mechanical ventilation (time to successful cessation of invasive mechanical ventilation)
Relevant SDTM and ADaM datasets		
	PR (procedures)	
	ADPE (endpoints)	
	ADVENT (ventilation occurrence and duration outcomes)	
	ADCNCES (censoring dates for cessation of IMV), ADCESTTE (cessation of IMV time-to-event)	
	Censoring rules are described in the define.xml documentation for the censoring datasets	
	Time-to-event tables are structured according to the ADaM TTE guidance document	



From OpenClinica CRF

usubjid	prdecod	proccur	prspid	prstrtpt	prsttpdy	prenrtpt	prentpdy	prdur
12345678	EXTRACORPOREAL MEMBRANE OXYGENATION	N	CRF	BEFORE OR COINCIDENT	7			
12345678	INVASIVE MECHANICAL VENTILATION	Y	CRF	BEFORE OR COINCIDENT	7			P3D
12345678	INVASIVE MECHANICAL VENTILATION OR ECMO		INT09	COINCIDENT	-1	BEFORE O	7	P4D
12345678	INVASIVE MECHANICAL VENTILATION OR ECMO		INT17	COINCIDENT	-1	BEFORE O	7	P4D

From ICU admission electronic datasets

Example – ventilation data

Outcome definitions

From CRF, occurrence of IMV is based on: mech_vent='1' or ecmo='1' or imv_days > 0. If imv_days > 0 then this is used as the duration. Reports of IMV from CR

From SUS+, ventilation = 1, with a start date within the 28 day period counts as IMV. Duration of IMV not available from SUS+

From SICSAG, if part of the admission interval overlaps the 28 day period, this counts as IMV, and the number of days overlap is the duration.

From INT10 SMR01 and INT14 PEDW, if the start date of the event falls in the 28 day period, this counts as IMV, number of days not available from these sourc

Consider rows in PR that represents an IMV episode from INT09, INT17 or INT24:

PRSTRTPT and PRENRTPT indicate whether the days in the episode should all be counted at the beginning of the episode or

The estimate of where abouts in the admission the days on IMV should count is modified by the date of death (from DM.D

The estimate of where abouts in the admission the days on IMV should count is also modified by whether the participant i

For each participant

For each date from day 1 to day 28

If the date is one of the days in an episode in PR, then set a flag indicating this

If the date is in an episode in PR, but not one of the days (i.e. the days are all

Set CONTRFL to 'Y' for any date that has INT09FL='Y', or has (INT17FL='Y' or I

Set CONTRFL to 'N' for any date where there has been an earlier changes fro

All dates with CONTRFL='Y' contributes to duration of IMV for this participant

Completeness of follow-up data

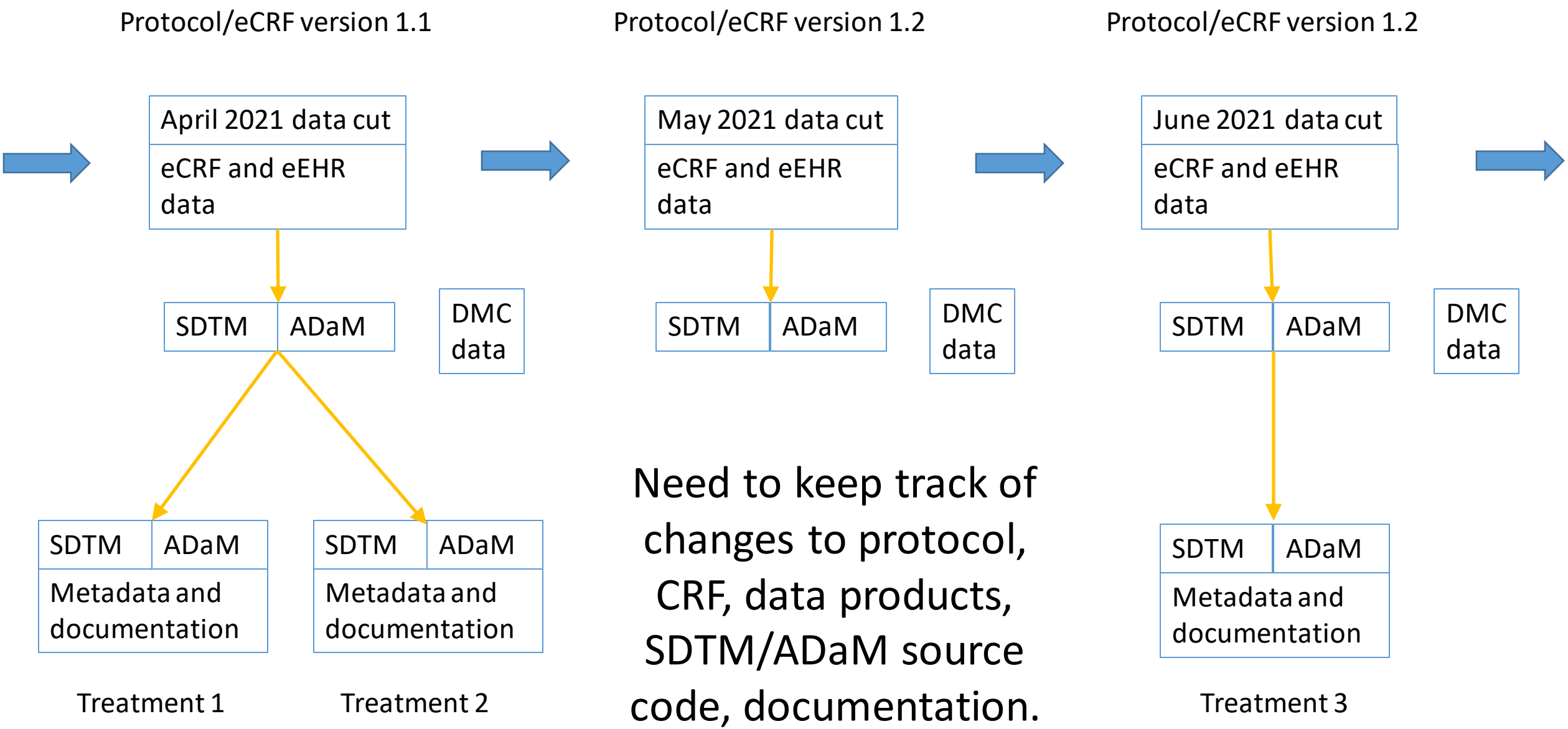
- Calculated from CRFs and from electronic healthcare data sources.
- Completeness for primary and secondary outcomes > 99%
- Having both CRF and electronic healthcare data sources enables us to repeat analyses using only CRF data, or using only electronic healthcare data. Good agreement between the two.

Completeness of follow-up data

- Death : because there is a national death registry then reasonable expectation that if somebody dies, data about death will be available.
- Use of ventilation within 28 days after randomisation : because all randomised participants are inpatients in an NHS hospital, we can be reasonably sure that if we get data about their admission from an NHS data source, we would expect to see any ventilation use.
- More problematic for longer-term outcomes after discharge: if participant discharged, how do we know whether they have moved outside UK, or between nations in the UK? How do we know whether they received private care?

Challenges : Platform trials

- Platform trials : some comparisons ongoing while others are being analysed and published. Protocol continually changing. Version control is important.
- CDISC standards in platform trials : separate data package for each comparison, containing e.g. different Trial Design datasets for each comparison, slightly different versions of SDRG and ADRG. Annotated CRFs change over time as protocol changes.
- Data sharing – limitations on what we are allowed to share from NHS digital. E.g. no absolute dates – how will regulators handle this?


Protocol/eCRF version 1.1
Protocol/eCRF version 1.2
Protocol/eCRF version 1.2
April 2021 data cut
eCRF and eEHR data
May 2021 data cut
eCRF and eEHR data
June 2021 data cut
eCRF and eEHR data
SDTM
ADaM
DMC data
SDTM
ADaM
DMC data
SDTM
ADaM
DMC data
SDTM
ADaM
Metadata and documentation
SDTM
ADaM
Metadata and documentation
SDTM
ADaM
Metadata and documentation
Treatment 1
Treatment 2
Treatment 3
Need to keep track of changes to protocol, CRF, data products, SDTM/ADaM source code, documentation.

Challenges : Data sharing

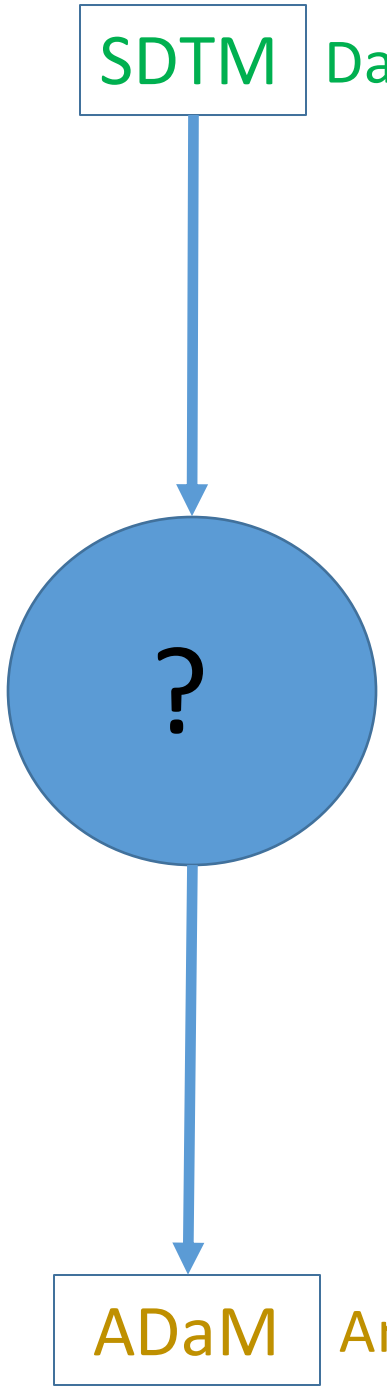
- Data sharing – limitations on what we are allowed to share from NHS digital. E.g. no absolute dates.
- Use SDTM –DY variable rather than date variable when exporting data.
- Not every SDTM date variable has a –DY variable, some non-standard variables need to be added.
- Many validation rules depend on dates. Two-step validation process used.

Current areas of work

- Template for standard description of data source attributes: what attributes do we need to know about? Geographical coverage, temporal coverage, expected max lag time between event occurrence and appearance in dataset. Summary of data entry methodology (+ reference to full data entry manual)
- Template for describing origin of data fields and SDTM mapping – annotated CRF equivalent.
- Need to understand how to describe data sources in a way that will be acceptable to regulators

Future areas of work

- What properties could healthcare data sources have to make them better suited to clinical trials?
- Immutability – always able to access data as it was at some point in time in the past.
- Standardisation



Where to represent “best estimate of the truth”?:

SDTM was developed in the context of CRF-based trials, where the CRF data is regarded as the final, clean version of what happened to the participant. When there are multiple sources, which could disagree with each other, how should disagreement and resolution be represented?

An extra layer is needed : derived from the multi-source data in SDTM, but which models what happened to the participant. Could be another instance of SDTM domains, but representing consolidated, best-estimate data.

Alternatively, the multi-source data could be represented using another data standard, prior to SDTM mapping, and included in data packages and regulatory submissions in addition to SDTM and ADaM.