Consensus Conference on Data Acquisition, Quality & Curation for Observational Research Designs (DAQCORD)
Sept. 18 – 19, 2018

Overview, Participants, and Meeting Notes

Overview: Sharing "big, complex data" has enormous potential to enhance and accelerate biomedical research and knowledge. However, it also comes with risks, as reflected in the still commonly heard phrase "Garbage in, garbage out". **Rather than assume that data is "good" or "bad", we propose to develop a practical self-assessment and reporting method for clinical research studies.** The goal is to capture key information about data acquisition, quality control measures, and curation in a tool that is linked to the dataset so that potential research collaborators can determine if the data meets their needs and expectations. While the impetus for the consensus conference came out of the International Traumatic Brain Injury Research (InTBIR) initiative, we believe that the DAQCORD reporting system will be relevant to many brain diseases and disorders.

Experts in clinical research design, bioinformatics, data management, biomarker and therapy development, and other relevant fields were invited to participate in a consensus conference. A modified Delphi process was used to reach agreement on the items and structure for the DAQCORD self-assessment and reporting tool. Prior to the meeting, participants received preparatory information by email and webinars to complete a pre-meeting survey (Round 1). During the meeting, participants reviewed, discussed and refined the survey in an iterative process (Rounds 2 and 3), and also developed a plan for implementation. Following the meeting, participants will be asked to beta-test the reporting tool on their own data, if available. In addition, participants will be invited to co-author a publication on the development of the DAQCORD tool.

Deliverables: Anticipated deliverables include:

- Consensus about the items and structure for the DAQCORD self-assessment and reporting tool.
- A peer reviewed publication describing the purpose, process and recommendations from the DAQCORD consensus conference;
- A strategy or plans for evaluation and implementation of the DAQCORD self-assessment and reporting tool in clinical research.

Participants: Steve Wisniewski, Univ. Pittsburgh, and Ari Ercole, Univ. Cambridge; Donald Marion, Defense and Veterans Brain Injury Center; Tony Fabio, Univ. Pittsburgh; Pradeep George, INCF; Mona Hicks, One Mind; Mike Jarrett, Quesgen; Matt McAuliffe, NIH/CIT; Lindsay Wilson, Stirling Univ.; Xinming An, Univ. North Carolina; Pat Bellgowan, NIH/NINDS; Guido Bertolini, Marion Negri Inst.; Vibeke Brinck, Quesgen; Doxa Chatzopoulou, UCLA; Ramon Diaz-Arrastia, Univ. Penn; Adam Ferguson, UCSF; Isabelle Gagnon, McGill Univ.; Joe Giacino, Harvard; Jeff Grethe,

UCSD; Robert Heinssen, NIH/NIMH; Ferath Kherif, Univ. Lausanne; Chris Lindsell, Vanderbilt; Louise Marshall, Wellcome Trust; Christine MacDonald, Univ. Washington; Carolina Mendoza-Puccini; NIH/NINDS; Erik Montes, One Mind; David Nelson, Karolinska Inst.; Tara Niendam, UC Davis; Pat Rinvelt, National Network of Depression Centers; Laurie Silfies, Univ. Pittsburgh; Stephen Strother, Univ. Toronto; Carol Taylor-Burds, NIH/NINDS; Theresa West, DOD

Sponsors: One Mind and NIH/NINDS

Brief Review of the Problem and Goals of the Meeting – Ari Ercole (Appendix A)

Round 1 Survey Results: Lindsay Wilson (Appendix B)

Discussion Questions and Responses from Small Group Breakout Sessions

- What is your overall impression of the items on the survey after completing Round 2? Are there important gaps or missing items that should be included in the tool?
    - The list of items was viewed as good overall, but the list either needs to be pared down or structured in a way that uses skip logic, algorithms, or tables to increase the relevance to an individual study and decrease the burden.
    - The focus is on best practices for managing "complex data". There may also be some "foundational items" that are relevant to "simpler" studies, too, but this is uncertain until we finalize it and beta-test it across multiple different types of studies.
    - The target audience includes principal investigators, data collectors, trainees, funding agencies, and journals.
    - eCRF should be removed from wording of questions where the item is relevant to studies that don't have an eCRF.
    - Clarity of the questions is still a concern.
    - Duplicates are still present and should be removed.
    - Most items look at random bias, may need to include more items related to systematic bias.
    - Gaps in assessing quality and how to manage data sharing were mentioned.

- What should the end product of a data quality reporting tool look like? E.g., number of items? Published or on-line or both? Checklist or scored items or combination of both? How to present the overall summary of data quality in terms of a score, graphics or what?
    - This should be a living web tool, where it can be modified based on users' experiences and advancements in technology, etc.

    - The final design of the tool will depend on which stages of the data life cycle are targeted: Project design, quality of curation, and long-term stewardship.
    - The tool should link to publications and to the data, wherever it is stored.
    - Publications and manuals are needed to provide justification and technical support for the tool.
    - A pre- and post-model may be useful, where the tool is used prospectively to design the data management process, and retrospectively to identify and explain the reason for changes in the design.
- What is the model or plan for validation and sustainability of the tool? How do we promote user uptake and/or penetration into biomedical research?
    - Validate the tool by retrospectively applying it to existing data sets, i.e. TRACK-TBI, CENTER-TBI, and other large, complex studies of other brain disorders, if possible. This will provide evidence demonstrating its value.
    - Validate the tool prospectively on studies currently under development to identify problems, etc. and refine the tool.
    - Use the tool in training investigators and show that it leads to change in practice.
    - Find an organization that is able to "own and support" the tool. Other partnering organizations or endorsers could also join later.
    - Webinars, publications, presentations, workshops and online training tools were all recommended.
    - Have the tool endorsed by one or more high impact journal.

Round 2 Survey Recap – Lindsay Wilson and Ari Ercole

(Appendix C)

Landscape/Needs Analysis – What other groups are working on or may be working on highly related projects?

Guidelines: CONSORT, STROBE, PRISM, PROSPERO, Cochrane Guidelines,
Industry: Quintiles, Westat, CROs, Informatica, COMET, Journals, e.g. Nature
Initiatives: Human Brain Project, DOD Joint Trauma Registry (JTS), Resource Identification Initiative
Organizations: AHRQ, NIH/NLM, European Commons?, INCF, WELLCOME, KAVLI, MILKEN, NIHR (UK), Moore Foundation, CDISC, WHO, Ontario Brain Institute, Canadian Institutes of Health and Information, Canadian Open Neurosciences, Gates Foundation, Allen Brain Institute, CAMARADES, Dept. of Energy, DARPA, ACRP, HHIS?, Chan-Zuckerberg Foundation, NSF, ELIXAR
Societies: Association of Medical Informaticists (AMIR); Association of Clinical Sciences; EQUATOR Network, Society for Clinical Research Coordinators, Society for Clinical Data Management, Society for Evidence-Based Health Care, Society for Neuroscience

Tools and Websites: ebCOG, ClinicalTrials.gov, NH Digital, Neurobank, ZingTree
Training: Clinical and Translation Science Awards (CTSA), Italy Certification of Data Managers
GAP: Relatively few groups affiliated with low- and middle-income countries (LMIC) were identified.

What are the short-term goals, deliverables and action plan?

1. Consensus and finalization of the DAQCORD items. Jan. 15
    a. The planning committee will compile the Round 2 results, including text comments and disseminate this information to the DAQCORD members. Estimated timeline is October 31.
    b. Disseminate Round 3 survey after removing duplicate items and incorporating quantitative and qualitative feedback from Round 2. ~ Nov. 15
    c. Round 3 surveys completed by members. ~ Dec. 1.
    d. The planning committee will review the Round 3 results and disseminate this information to the DAQCORD members. ~ Dec. 15.
    e. Finalize items or prepare and disseminate Round 4 survey. ~ Jan. 15.
2. DAQCORD publication #1: High level summary. April 2019.
    a. The planning committee will select a few members to take the lead on writing a brief and concise high-level summary about DAQCORD. Estimated timeline for first draft is March 15, 2019.
    b. A high impact journal is preferred. Suggestions from members are welcome.
    c. The lead authors will be highlighted, but all DAQCORD members will be included in a format that is dependent on the journal. A draft will be circulated to DAQCORD members for review prior to submission. Comments due March 31, 2019.
    d. Submit manuscript ~ April 15, 2019.
3. Establish a Steering Committee for DAQCORD. ~ Nov. 1
    a. The planning committee will assume this roll temporarily until interested members are identified and agree to join. ~ Nov. 1
    b. Initial tasks include creating a charter or governance plan and mission statement. Estimated timeline - TBD
    c. Future anticipated tasks include:
        i. Outreach to other groups to identify partners or an "owner" for DAQCORD;
        ii. Coordination of DAQCORD activities, e.g. by organizing an executive committee of DAQCORD workgroup or committee chairs and establishing regular communications.
        iii. Creating a collection of "slides" for presentations.
        iv. Needs assessment for workshops and training tools.
        v. Estimated timeline – TBD
4. Establish a workgroup to create web tools for managing complex data. ~ Nov. 1
    a. Include experts in "user-design" on the workgroup, if possible.

b. A “living tool” that can be refined over time is recommended.
c. Specific tasks and tools will be determined by the workgroup.
d. Suggested or potential tools include a:
    i. “Quick Start Guide”,
    ii. Comprehensive tool with algorithms or skip-logic based on the type of study and types of data/domains to decrease the burden and increase the relevance
    iii. An output tool that builds a customized guideline of DAQCORD items and also creates a reporting tool based on input about the study type and data/domains.
    iv. Pilot test the tools.
    v. Develop a training manual to accompany the tools.
    vi. Convene workshops and webinars for training as needed.
    vii. Disseminate information to relevant scientific societies, guidelines groups, funding agencies, etc.
    viii. Estimated timeline – TBD

5. Establish a workgroup to coordinate and assist with DAQCORD publications ~ Nov. 1
    a. Publication #1 (see above)
    b. Suggestions for future publications include:
        i. Detailed reflections on lessons learned and how they were addressed from retrospective use cases as an argument for prospective implementation of best practices
        ii. Detailed publication about the DAQCORD process, including the modified Delphi method, target users, validation of the guidelines, etc.
        iii. Domain specific best practices
        iv. Estimated timeline – TBD

Appendix A - Brief Review of the Problem and Goals of the Meeting – Ari Ercole

DAQCORD

01000100 01000001 01010001 01000011 01001111 01100010 01000100

Data Access, Quality & Curation for Observational Research Designs

Background and scope

Ari Ercole MD PhD
University of Cambridge, UK
On behalf of the DAQCORD team

InTBIR International Initiative for Traumatic Brain Injury Research

NIH National Institute of Neurological Disorders and Stroke

ONE MIND

Data curation

- Based on experience of data curation
  - CENTER-TBI
  - CC-HIC (ICU data-sharing project of U.K. multicentre "MIMIC III")
  - Discussions with InTBIR partners

We found…

Things we could fix but it was hard

Things we could make better but never perfect

Things that were OK, but made the data harder to use than necessary for the non-expert researcher

Things we couldn't fix and would remain limitations for ever

•So much could have been avoided at design-time

- Data logic
- Data specification
- Structure specification
- Database design

And the result?

- We have some AWESOME data, but…
- A lot of effort has been expended
  - And more will be required
- Data will remain difficult to use
  - Investigator effort otherwise potentially introduce analysis errors
- Data limitations complex
  - Hard to document robustly
  - Will limit re-use
    - How will the non-specialist cope?
    - Hard to federate

"What has been seen, cannot be unseen"

- We want to avoid others making these mistakes…
- BUT now that I know this…
  - I will not trust a study without similar attention to data quality
  - But how would we assess this?

Huge (unique) experience!

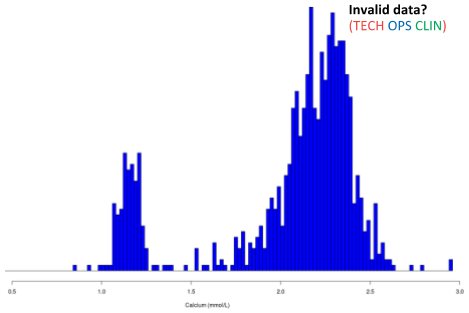
- InTBIR datasets really rather complicated
- Multiple centres, multiple countries, multiple languages C.F. Framingham
- >5000 patients (not counting registry)
  - Heterogeneous patients: ICU, ED, ward
  - Heterogeneous data: Clinical, outcomes, interventions, imaging, lab… C.F. '100,000 genomes'
- Individually quite 'big' but together unparalleled

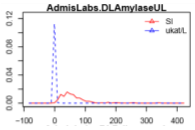
What is data quality?

Missing

- If **missing** is a data quality problem then are these data quality problems too?
  - How much is missing?
    - Hard to search (needs database programming) (TECH)
  - Missing
    - Because unknown
      - Clinically unknown / logistically indeterminate (CLIN)
    - By accident
      - Forgotten (OPS)
    - By study design (stratum-specific) DESIGN
    - By collusion with CRF design
      - Because not mandated but is this possible/desirable always (TECH CLIN)
  - How do I correct before too late?
    - Process for ongoing run-time Q.A. (TECH OPS CLIN)

If **missing data** is a data quality problem, what is the scope of the solution?





Needs to be design-time consideration
(on-the-fly hard)
- Filter?
- Validate?

(TECH OPS CLIN)

```
library(RMySQL)
library(psych)

#Create MySQL connector
concleaned <- dbConnect(MySQL(),
   dbname = "CLEAN_centertbi",
   port = 3306,
   host = "127.0.0.1",
   user = "******",
   password = "******")

#Run query
res <- dbGetQuery(
  concleaned,
  "SELECT DLCalciummmolL
     from Labs
     where SubjectID > 1000")

# Draw histogram
hist(res$DLCalciummmolL, breaks=100)
```

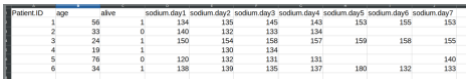
If **invalid data** is a data quality problem, what is the scope of the solution?

registration.institutionType
tion meta-data: "The nature of the institution completing this eCRF. Etc etc"
Type: text
Permissible values: "academic teaching hospital"; "non-teaching hospital"; etc...
Mandatory: yes
Mutually exclusive: yes
...
registration.catchmentArea
escription meta-data: "The approximate catchment area of the institution..."
Type: text
Permissible values: "<100.000"; "100.000-250.000"; ...
Mandatory: yes
Mutually exclusive: yes
...
registration.medicalStaffing
...

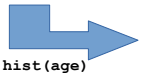
patientElements.age
Description meta-data: "Patient age on admission"
Type: integer
Units: years
Permissible values: >0 AND <150
Mandatory: yes
...
patientElements.medicalHx.cardiovascular.congenitalYND
Description meta-data: "Past medical history of congenital heart disease"
Type: text
Permissible values: "yes"; "no"; "unknown"
Mutually exclusive: yes
Mandatory: yes
...

Failure to structure data

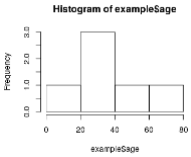
If **poor data structure/ specification** is a data quality problem, what is the scope of the solution?



Describe age?



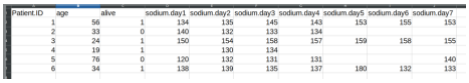
```
hist(age)
mean(age)
stdev(age)
```



Age and survival?

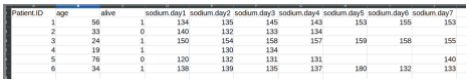
```
glm(alive~age, family=binomial)
```

EASY!



Sodium trajectories? How?

```
plot(what against what?)
```



Errors in sodium?

How?


```
hist(what?)
```

If **poor database design** contributes to a data quality problem, what is the scope of the solution?

Huge unique experience!

- No prior art
  - We (neurosciences) probably have the greatest experience in large, heterogeneous observational datasets
- We need to share our experience of this for the future
  - To improve data quality, research and data sharing
- How has data quality been assured in other domains?

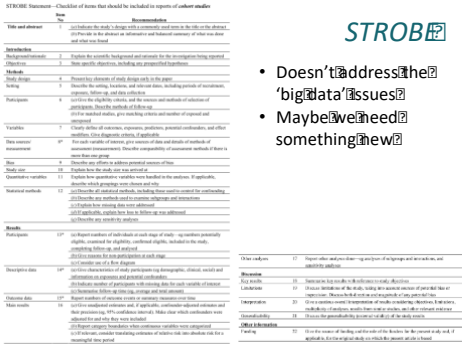
PRISMA
TRANSPARENT REPORTING OF SYSTEMATIC REVIEWS AND META-ANALYSES

STROBE Statement
Strengthening the reporting of observational studies in epidemiology

CONSORT TRANSPARENT REPORTING of TRIALS

SPIRIT

Et al...


STROBE Statement—Checklist of items that should be included in reports of *cohort studies*

	Item No	Recommendation
Title and abstract	1	(a) Indicate the study's design with a commonly used term in the title or the abstract
		(b) Provide in the abstract an informative and balanced summary of what was done and what was found
Introduction		
Background/rationale	2	Explain the scientific background and rationale for the investigation being reported
Objectives	3	State specific objectives, including any prespecified hypotheses
Methods		
Study design	4	Present key elements of study design early in the paper
Setting	5	Describe the setting, locations, and relevant dates, including periods of recruitment, exposure, follow-up, and data collection
Participants	6	(a) Give the eligibility criteria, and the sources and methods of selection of participants. Describe methods of follow-up
		(b) For matched studies, give matching criteria and number of exposed and unexposed
Variables	7	Clearly define all outcomes, exposures, predictors, potential confounders, and effect modifiers. Give diagnostic criteria, if applicable
Data sources/ measurement	8*	For each variable of interest, give sources of data and details of methods of assessment (measurement). Describe comparability of assessment methods if there is more than one group
Bias	9	Describe any efforts to address potential sources of bias
Study size	10	Explain how the study size was arrived at
Quantitative variables	11	Explain how quantitative variables were handled in the analyses. If applicable, describe which groupings were chosen and why
Statistical methods	12	(a) Describe all statistical methods, including those used to control for confounding
		(b) Describe any methods used to examine subgroups and interactions
		(c) Explain how missing data were addressed
		(d) If applicable, explain how loss to follow-up was addressed
		(e) Describe any sensitivity analyses
Results		
Participants	13*	(a) Report numbers of individuals at each stage of study—eg numbers potentially eligible, examined for eligibility, confirmed eligible, included in the study, completing follow-up, and analyzed
		(b) Give reasons for non-participation at each stage
		(c) Consider use of a flow diagram
Descriptive data	14*	(a) Give characteristics of study participants (eg demographic, clinical, social) and information on exposures and potential confounders
		(b) Indicate number of participants with missing data for each variable of interest
		(c) Summarize follow-up time (eg, average and total amount)
Outcome data	15*	Report numbers of outcome events or summary measures over time
Main results	16	(a) Give unadjusted estimates and, if applicable, confounder-adjusted estimates and their precision (eg, 95% confidence interval). Make clear which confounders were adjusted for and why they were included
		(b) Report category boundaries when continuous variables were categorized
		(c) If relevant, consider translating estimates of relative risk into absolute risk for a meaningful time period
Other analyses	17	Report other analyses done—eg analyses of subgroups and interactions, and sensitivity analyses
Discussion		
Key results	18	Summarize key results with reference to study objectives
Limitations	19	Discuss limitations of the study, taking into account sources of potential bias or imprecision. Discuss both direction and magnitude of any potential bias
Interpretation	20	Give a cautious overall interpretation of results considering objectives, limitations, multiplicity of analyses, results from similar studies, and other relevant evidence
Generalizability	21	Discuss the generalizability (external validity) of the study results
Other Information		
Funding	22	Give the sources of funding and the role of the funders for the present study and, if applicable, for the original study on which the present article is based

STROBE!

- Doesn't address the 'big data' issues
- Maybe we need something new

So what have we done about it?

- We came up with DAQCORD
  - Data Access, Quality and Curation for Observational Research Designs
- We want to systematically develop metrics that are useful for designing in data quality

DAQCORD

01000100 01000001 01010001 01000011 01001111 01100010 01000100

DAQCORD: Why?

- To promote **high quality data**
  - To improve primary research
    - Less missingness
    - Better data integrity
    - Simplified access for researchers
  - To improve data sharing for the future
    - Improve consistency
    - Encourage data standards and formats
- Concepts are generic & go beyond neuroscience
- We have important experience to share

What are we aiming for?

DAQCORD will provide:

- Criteria against which to demonstrate study design quality
  - Like STROBE, PRISMA, CONSORT...
- A tool to assist in good study design prospectively
- A tool to assist in designing-in quality data collection/QA
- A tool to improve data sharing through design and documentation

Not just a 'check list'..

Where are we now?

- Started in June 2017
- Virtual meetings so far
  - Teleconferences (>30hours to date!), emails, webinars...
- We have
  - Literature review (there isn't any...!)
  - Decided on a prototype list of quality indicators
  - Organised, agreed and themed these indicators
    - Conducted an initial first round Delphi-like process
      - Developed a platform for this
- **We still need**
  - Consensus expand / refine by completing the Delphi process
  - Validate our tool (e.g. inter-rater reproducibility)
    - Retrospective application to existing projects
  - Discussion, expertise and consensus of ways forward
    - Model for best implementation

Any questions?

InTBIR International Initiative for Traumatic Brain Injury Research

ONE MIND

Appendix B - Round 1 Survey Results: Lindsay Wilson

01/10/2018

DAQCORD

DAQCORD: Delphi

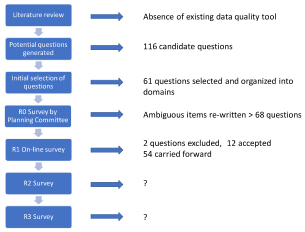
Lindsay Wilson
For DAQCORD Team

InTBIR International Initiative for Traumatic Brain Injury Research

QuesGen SYSTEMS, INC.

ONE MIND

DAQCORD: Modified Delphi flow diagram



Round 1 Rating Scales

Clarity, Expertise, Validity, Feasibility, Discriminability, Actionability

Rating scale: 5= Strongly agree, 4= Agree, 3= Neither agree nor disagree, 2= Disagree, 1= Strongly Disagree

Median score:
- ≥4 = Good agreement rating
- ≤3 = Neutral or poor rating

Interquartile range (IQR):
- IQR 0, 1 = Very good consensus
- IQR 2 = Good consensus
- IQR >2 = No consensus

Clarity: "This is clearly written and understandable"

Question	Short title	Median	IQR
23	Defining_data_representations	2	2
1	Duplicate_data	2.5	4
66	Variables_hierarchy	3.5	2

Items with lowest overall ratings on clarity.
Scale: 5= Strongly agree to 1= Strongly Disagree

Clarity

- Three items with low scores (median<4), a further 15 with lack of consensus (IQR ≥2)

- Reviewed written feedback for all questions
- Reworded to clarify concepts
- Added examples to many of the questions

Expertise

Question	Short Title	Percent
57	Biospecimen_audit_trail	35
25	Database_specification	35
35	Neuroimaging_acquisition_and_processing	36
	.....	
9	Data_dictionary_end_users	78
65	Expertise_data_collectors	79
15	Permissible_value_encoding	83

Examples of items with low and high ratings.
Percent who agree / strongly agree with the statement "This is my area of expertise"

Round 1 Survey rating scales

- Clarity
- Expertise
- Validity
- Feasibility
- Discriminability
- Actionability

Selection criteria

'This metric is likely to reflect the quality of the data.'

'This is something that can be measured or assessed, and is quantifiable.'

'There is likely to be variation in practice on this between different study designs.'

'Improving this metric could be used in practice to make changes to a study that improve data quality.'

Selection rules

Criteria

Validity, Feasibility, Discriminability, Actionability

Example rule for acceptance
Median score of ≥4 on all criteria -> Overall agreement
An interquartile range IQR ≤ 1 on validity and IQR ≤ 2 on the other criteria -> Overall consensus across criteria

Outcomes

(1) Accept
Consensus + good score -> Accept

(2) Reject
Consensus + bad score -> Exclusion

(3) Carry forward to next round
Consensus + bad score + new definition/comments -> next round
No consensus -> next round

From Lingsma & Huijben

Round 1 – main outcomes

- Rejected - 2 questions
- Accepted - 12 questions
- Carried forward to next round - 54 questions

Rejected questions

Q	Short title	Validity		Feasibility		Discrimin-ability		Action-ability	
		Med.	IQR	Med.	IQR	Med.	IQR	Med.	IQR
11	Dataset_scope	3	1.5	4	2	4	2	4	2
60	Anonymisation_strategy_validation	3	1.5	4	2	4	2	4.5	2

Question Text

11. The scope and content of the data are aligned with specified research question(s) to ensure that data collection is necessary and sufficient.

60. The strategy for anonymisation is validated and includes an estimated risk for re-identification

All ratings with median agreement <4

Q	Short title	Validity		Feasibility		Discrimin-ability		Action-ability	
		Med.	IQR	Med.	IQR	Med.	IQR	Med.	IQR
3	CRF_design_stakeholders			3	2				
~~11~~	~~Database_specification~~	3	1.5						
14	Data_encoding_logic	3	3	3	2				
25	Database_specification	3	2						
30	Data_portal_design	3.5	2						
57	Biospecimen_audit_trail					3.5	3		
59	De_identification_by_condition	3.5	3			3	3		
~~60~~	~~Data_encoding_logic~~	3	1.5						
63	Ongoing_data_hosting	3	2						
64	Unique_identifier					3	4		
66	Variables_hierarchy	3	2						

All other ratings of the 68 questions had a median of 4 or 5 –> Overall high ratings of agreement across the 4 criteria

Questions with low validity rating

14. Variables are named and encoded in a user-friendly way

25. A clear rationale for the choice of database architecture is given.

30. The data portal has appropriate measures in place for security and access.

59. Information that may be personally identifiable is removed from the database as appropriate to the pre-specified level of anonymisation.

63. There should be a clear and detailed sustainability plan to ensure data availability after the study has ended.

66. Variables are described in a hierarchy that categorizes and organizes the data.

Accepted questions

Q	Short title	Validity		Feasibility		Discrimin-ability		Action-ability	
		Med.	IQR	Med.	IQR	Med.	IQR	Med.	IQR
2	Data_Collection_Procedure	4	1.0	4	1.5	4	1.3	4.5	1.0
6	Validation_of_data_collection_teams	5	0.3	4	1.0	5	1.0	5	1.0
9	Data_dictionary_end_users	5	1.0	4	1.0	4	1.0	4.5	1.0
10	Data_submission_timeliness	4	1.0	5	1.0	4	1.0	5	1.0
13	Inter_rater_reliability	5	1.0	5	1.0	4	1.5	4	1.0
20	eCRF_skip_logic_design_and_documentation	4	1.0	4	1.0	4	2.0	5	1.0
27	Significant_figure_handling	4	1.0	5	1.0	4	1.3	4	1.0
28	Audit_trail	5	1.0	5	1.0	4	1.3	4	1.0
33	Bioassay_QC_and_QA	4	1.0	5	1.0	4	1.0	4	1.0
38	ETV_software_design	4	1.0	5	1.0	4	1.0	5	1.0
44	Feedback_to_data_collection_team	5	1.0	5	1.0	4	1.0	5	1.0
46	Univariate_validation	5	1.0	5	1.0	4.5	1.3	5	1.0

Median score of ≥4 on all criteria -> Overall agreement
An interquartile range IQR ≤ 1 on validity and IQR ≤ 2 on the other criteria -> Overall consensus across criteria

Consensus across criteria

IQR	Validity	Feasibility	Discrimin-ability	Action-ability
0,1	49	50	13	61
2	15	18	41	6
>2	4	0	14	1

Consensus on ratings of 68 items in Round 1. IQR 0, 1 = Very good consensus; IQR 2 = Good consensus; IQR >2 = No consensus

Q	Short title	Validity		Feasibility		Discrimin-ability		Action-ability	
		Med.	IQR	Med.	IQR	Med.	IQR	Med.	IQR
8	Source_validation_of_inclusion_criteria	4.5	1	5	1	4	3	5	1
16	Mutual_exclusivity_design	5	1	5	1	4	3	5	1
21	Measurement standardisation	5	1	5	1	4	3	5	1
48	Pipeline_documentation	5	1	5	1	4	3	5	1
54	Scoring	5	0.8	5	1	4.5	3	5	1
62	Lock_down_and_version	4.5	1.8	5	1	4	3	5	1

Examples of items with lack of consensus on discriminability.

Lack of consensus on discriminability: Text of examples

8. Data collection includes fields for documenting that participants meet inclusion/ exclusion criteria.

16. Mutual exclusivity or otherwise is defined: There should be rules to ensure that incompatible choices are excluded both within a data element (e.g. cannot be both male and female) and between related elements (e.g. male cannot be pregnant).

21. For physiological data the methods of measurement, units, and (where relevant) datums are defined for all sites.

48. Automated variable transformations are documented and tested before implementation and if modified.

54. Scoring of tests is checked. Scoring is performed automatically where possible.

62. The version lock-down of the database for data entry is clearly specified.

Round 1 – main outcomes

- Rejected - 2 questions
- Accepted - 12 questions
- Carried forward to next round - 54 questions

Progress in next round

- Further questions rejected - primarily consensus on low validity
- Further questions accepted – primarily good consensus across criteria, particularly discriminability

DAQCORD Round 2 Delphi

- 54 Items carried forward from Round 1
- 32 People completed the questionnaire

Clarity

Round 1:
- Three items with low agreement scores (median<4),
- 15 with lack of consensus (IQR ≥2)

Round 2:
- No items with low agreement scores
- One with lack of consensus (IQR=3):

"36. Inter-site sample differences are described in meta-data."

Validity: *'This metric is likely to reflect the quality of the data.'*

- Agreement scores:
  - 48 items with median 4/5
  - 6 items with median <4
- IQR
  - 52 items with range 0-2
  - 2 items with range >2

Round 1: Validity

Q	Short title	Validity	
		Med.	IQR
~~11~~	~~Database_specification~~	3	1.5
14	Data_encoding_logic	3	3
25	Database_specification	3	2
30	Data_portal_design	3.5	2
59	De_identification_by_condition	3.5	3
~~60~~	~~Data_encoding_logic~~	3	1.5
63	Ongoing_data_hosting	3	2
66	Variables_hierarchy	3	2

Ratings with median agreement <4

Round 2: Validity

Q	Short title	Validity	
		Med.	IQR
23	Defining_data_representations	3	2
25	Database_specification	3.5	1
30	Data_portal_design	3	1.5
59	De_identification_by_condition	3	3
63	Ongoing_data_hosting	3	1
66	Variables_hierarchy	3	1.5

Ratings with median agreement <4

Questions with low validity rating

23. Appropriate data output formats are available.

25. A clear rationale for the choice of database architecture is given.

30. The data portal has appropriate measures in place for security and access.

59. Information that may be personally identifiable is removed from the database as appropriate to the pre-specified level of anonymisation.

63. There should be a clear and detailed sustainability plan to ensure data availability after the study has ended.

66. Variables are described in a hierarchy that categorizes and organizes the data.

Feasibility & Actionability

Feasibility: *'This is something that can be measured or assessed, and is quantifiable.'*
- Median agreement scores 4 or 5 for all items
- IQR range 0 to 2

Actionability: *'Improving this metric could be used in practice to make changes to a study that improve data quality.'*
- Median agreement scores 4 or 5 for all items
- IQR range 1 to 2

Discriminability: *'There is likely to be variation in practice on this between different study designs.'*

- Agreement scores:
  - 51 items median 4/5
  - 3 items with median <4 (Questions 1, 8, 15)
- IQR
  - 44 items with range 0-2
  - 10 items with range >2

Discriminability: Items with IQR = 3/4

8. Data collection includes fields for documenting that participants meet inclusion/ exclusion criteria.

18. Missingness is defined and is distinguished from 'not available', 'not applicable', 'not collected' or 'unknown.' For optional data, 'not entered' is differentiated from 'not clinically available' depending on research context.

19. Data-types are specified for each variable.

31. Assessors are blinded to treatment allocation where appropriate and such blinding is explicitly recorded.

64. Each individual has a unique identifier.

65. Data collection that requires specific content expertise is carried out by trained and/or certified investigators.

Importance: *"Overall, this question is important"*

- Agreement scores:
  - All items median 4/5
- IQR
  - Range 0-2

Importance

	Short Title	Mean	Median	IQR
7	eCRF_data_range_and_logic_checks	4.8	5	0
4	eCRF_documentation	4.7	5	0
41	Agreement_with_Source_Data	4.7	5	1
5	Training_of_data_collectors	4.6	5	1
54	Scoring	4.6	5	1
64	Unique_identifier	4.6	5	1
34	Biospecimen_collection_and_processing	4.6	5	1
50	Data_Consistency	4.6	5	1
51	Identification_of_site_issues	4.6	5	1
18	Optional_data_point_encoding	4.6	5	1

Top 10 items on importance rating.

Appendix C - Round 2 Survey Recap – Lindsay Wilson and Ari Ercole

DAQCORD

01000100 01000001 01010001 01000011 01001111 01100010 01000100

Data Access, Quality & Curation for Observational Research Designs

Round 2 initial thoughts

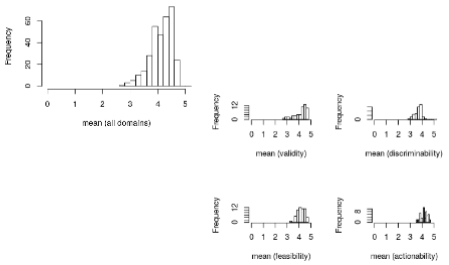
Ari Ercole MD PhD
University of Cambridge, UK
On behalf of the DAQCORD team

InTBIR

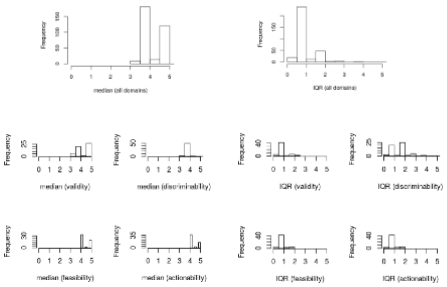
NIH National Institute of Neurological Disorders and Stroke

ONE MIND

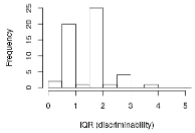
Mean scores



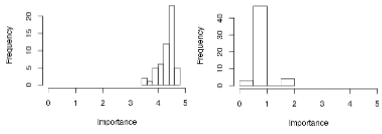
Median scores / IQRs

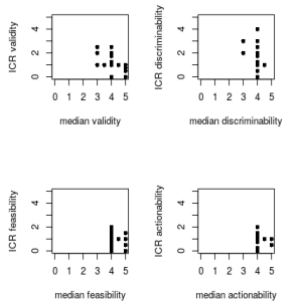


Discriminability IQR spread



Importance ratings





DAQCORD

01000100 01000001 01010001 01000011 01001111 01100010 01000100

***D**ata **A**ccess, **Q**uality & **C**uration for **O**bservational **R**esearch **D**esigns*

Better data, better research

InTBIR

NIH National Institute of Neurological Disorders and Stroke

ONE MIND