RESEARCH · RESEARCH DESIGN

Multi-Site Research Coordination and Standardization for Comparable Knowledge

How to coordinate a study run at several sites so that the results can be compared and combined, covering governance, protocols, what to standardise and what to adapt, translation and measurement equivalence, training, data management, monitoring, analysis of site effects and the use of site differences as findings.

ModQu research article · 2,873 words · 26 references (APA 7th edition)

Many important studies cannot be carried out in one place. A clinical trial may need patients from thirty hospitals to reach its sample size. A cross-national survey compares attitudes in twenty countries. A retailer tests a new service model in stores across several regions. A group of laboratories repeats the same experiment to see whether a published finding holds. All of these are **multi-site studies**: one research design carried out by different teams, in different places, often at the same time.

Multi-site designs bring large benefits. They recruit participants faster, produce larger samples and show whether findings hold across settings rather than in one favourable location. They also bring a specific risk. Every site has its own staff, procedures, participants and habits. Unless these are coordinated, the study becomes a set of loosely related projects whose results cannot be compared or combined. This article explains how multi-site research is organised, what must be standardised, what may be adapted, and how differences between sites should be monitored, analysed and reported.

Why run a study at several sites?

The first reason is size. Many questions need more participants than one site can supply in a reasonable time, and adding sites is often the only practical way to reach the sample required. The second is generalisability. A finding from one hospital, one school district or one market may depend on local conditions. Repeating the study across sites tests whether the result travels. Weinberger et al. (2001) noted that this is particularly important in health services research, where the way care is organised differs widely between institutions.

A third reason is to examine variation itself. When the same intervention produces different results in different places, the pattern can reveal the conditions under which it works. Large coordinated replication projects in psychology, such as the effort to repeat 100 published studies using shared protocols across many teams, have used this design to estimate how far findings depend on the laboratory that produced them (Open Science Collaboration, 2015).

These benefits come at a cost. Kraemer (2000) described the pitfalls of multi-site clinical trials: differences between sites in the populations recruited, in how treatments are delivered and in how outcomes are measured add variability, can reduce statistical power and, if they are not anticipated, can make the overall result hard to interpret. Most of the work of coordination is aimed at controlling these differences where they are unwanted and measuring them where they are of interest.

Where site differences come from

It helps to separate the sources of variation, because each needs a different response. Weiss et al. (2014) proposed a framework for explaining why the effect of a programme varies across sites. They distinguished differences in the **treatment contrast** (what participants in each group actually receive), in **implementation**, in the **characteristics of participants** and in the **context** in which the programme operates. To these, a researcher planning a multi-site study must add two more sources that have nothing to do with the intervention: differences in **measurement**, such as interviewers who ask questions in different ways, and differences in **data handling**, such as sites coding the same answer in different ways.

Some of this variation is real and informative. Participants genuinely differ between a city clinic and a rural one. Other variation is an artefact of the research process and should be removed. The aim of coordination is to remove the artefacts so that the real differences can be seen.

Governance and roles

Multi-site studies need a clear structure of responsibility. Typical arrangements include:

- a **coordinating centre**, which manages the protocol, training, communication, data and timelines;
- a **steering committee**, with representatives from the sites and the coordinating centre, which makes decisions on design and changes;
- a **site lead** at each location, responsible for local recruitment, delivery and data quality;
- a **data management** function, often within the coordinating centre, which builds the data collection system and checks incoming data;
- in clinical trials, an independent **data monitoring committee** that reviews safety and interim results.

Weinberger et al. (2001) emphasised that operational issues, such as communication between sites, agreement on responsibilities and the time needed to obtain local approvals, take more effort than investigators usually expect. A written agreement covering authorship, data ownership, publication and costs avoids disputes later.

Start-up, approvals and communication

Sites rarely start at the same time. Each may need its own ethics or privacy approval, its own contracts and its own staff recruitment, and delays at one site hold back the whole study. A realistic plan staggers site start-up, begins with one or two sites that can serve as a vanguard, and uses their experience to refine the manual before the others begin. This is similar in spirit to an internal pilot, and the lessons from the first sites are often the most valuable of the whole study.

Communication must be regular and predictable. Most coordinating centres hold scheduled calls with site leads, circulate short written updates, and keep a shared log of questions and answers so that a ruling given to one site is applied at all of them. Without such a log, sites answer the same practical question in different ways, and small differences in procedure accumulate. A single named contact at the coordinating centre for each site also reduces confusion when problems arise.

The protocol and the manual of procedures

Every multi-site study needs a single written protocol that all sites follow. The SPIRIT statement sets out what a trial protocol should contain, including objectives, eligibility criteria, interventions, outcomes, sample size, recruitment, data collection, data management and analysis (Chan et al., 2013). Although written for clinical trials, the checklist is a useful template for any multi-site study.

The protocol states what is to be done. A **manual of procedures** explains how, in enough detail that a new member of staff at any site could follow it. It covers recruitment scripts, consent procedures, the order and wording of measures, how to handle refusals and missing data, how to record deviations and whom to contact when something goes wrong. The manual is where most of the practical standardisation happens, and it should be updated, with version control, as questions arise during the study.

What to standardise and what to adapt

Complete uniformity is neither possible nor always desirable. Sites differ in language, staffing, regulations and the populations they serve, and forcing identical procedures on all of them can make the study unworkable or reduce its relevance. The task is to decide which elements are core, and must be identical, and which may be adapted to local conditions, with every adaptation recorded.

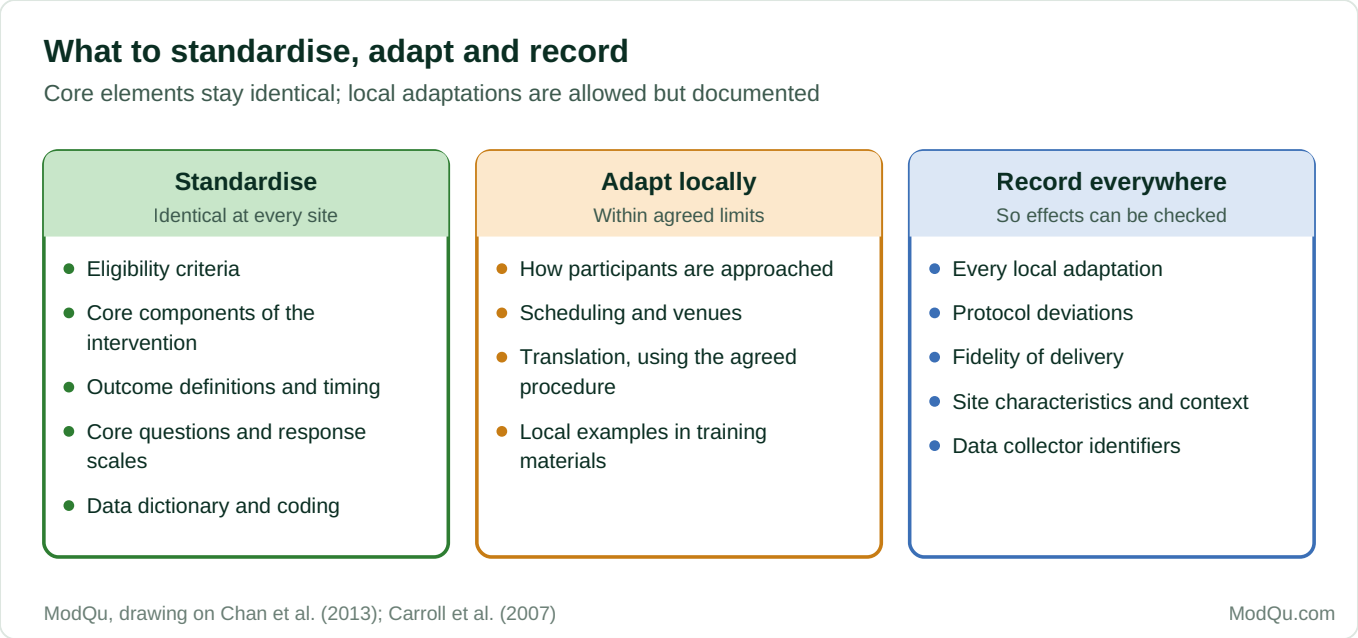


Figure 1. Deciding what a multi-site study holds constant, what each site may adapt, and what must be recorded. Diagram by ModQu, drawing on Chan et al. (2013) and Carroll et al. (2007).

As a rule, anything that defines what is being measured or compared should be standardised: eligibility, the intervention's essential components, outcome definitions, the core questions and their response scales, the timing of measurement, and the data format. Elements that affect only how the study is carried out locally, such as how participants are first approached or when sessions are scheduled, can usually be adapted. The essential step is to document every adaptation so that its possible effect can be checked in the analysis.

Implementation research offers a useful vocabulary for this. Carroll et al. (2007) defined **implementation fidelity** as the degree to which an intervention is delivered as intended, and proposed measuring adherence to its content, coverage, frequency and duration. Durlak and DuPre (2008) reviewed nearly 500 studies and found that the level of implementation affected the outcomes obtained, and that perfect fidelity was rare. In a multi-site study, measuring fidelity at each site is the only way to know whether differences in results reflect differences in the intervention as delivered.

Comparable measurement across languages and cultures

When sites are in different countries or serve different language groups, comparable measurement becomes a central concern. A question translated literally may not mean the same thing, and a response scale may be used differently in different cultures. Brislin (1970) described **back-translation**, in which one translator converts the questionnaire into the target language and another translates it back, so that discrepancies can be spotted. Back-translation is still used, but survey methodologists now generally prefer a team approach in which several translators produce versions, reviewers discuss them, an adjudicator decides, and the result is pretested and documented (Harkness et al., 2003).

The European Social Survey, which has run in more than 30 countries, illustrates the full approach. A central team specifies the questionnaire, sampling standards, response rate targets, translation procedures and fieldwork rules, and national teams carry them out, with each step documented and checked (Jowell et al., 2007). The wider field of comparative survey research, sometimes called **3MC** for multinational, multiregional and multicultural contexts, has developed detailed guidance on design, translation, sampling and quality assurance for such studies (Harkness et al., 2010; Johnson et al., 2019).

Even with careful translation, it should not be assumed that a scale measures the same thing in every group. **Measurement invariance** testing checks whether the relationships between items and the underlying concept are the same across groups, so that differences in scores can be interpreted as real differences rather than differences in how the scale works (Vandenberg & Lance, 2000). Multi-site studies with multi-item scales should plan these tests in advance.

Training and calibration

People collecting data are a major source of site differences. Kish (1962) showed that interviewers can add substantial variance to survey estimates, because each interviewer influences answers in slightly different ways, and that this variance reduces the effective precision of the survey. Fowler and Mangione (1990) set out the principles of **standardised interviewing**: reading questions exactly as worded, probing in neutral ways, recording answers without interpretation and maintaining a consistent manner with every respondent. Groves et al. (2009) describe how training, supervision and monitoring reduce interviewer effects.

In multi-site studies, training should be centralised or delivered from common materials, and every data collector should be certified before starting. Where measurement involves judgement, such as coding open answers, rating observed behaviour or assessing patients, sites should calibrate against each other. Raters at all sites score the same set of example cases, and agreement is measured, for example with the **intraclass correlation** (Shrout & Fleiss, 1979). Training is repeated if agreement falls below an agreed level, and calibration is checked again at intervals during the study, because practices drift over time.

Common data systems

Data must be collected in the same format at every site. That requires a single **data dictionary**: a list of every variable, its name, its definition, its allowed values and its coding. Web-based data capture systems allow one dictionary to define the forms used at every site, with built-in validation rules that reject impossible values at the point of entry. REDCap, developed for this purpose, is now used by thousands of institutions (Harris et al., 2009).

Common systems also make it easier to combine and share data later. The FAIR principles state that research data should be findable, accessible, interoperable and reusable, which in practice means consistent identifiers, standard formats and clear documentation (Wilkinson et al., 2016). A multi-site study that adopts these practices from the start avoids the slow and error-prone work of reconciling different site datasets at the end.

Monitoring quality during the study

Problems at one site are much cheaper to fix early than late. Quality monitoring traditionally relied on site visits, in which monitors checked records against source documents. Baigent et al. (2008) argued that this approach is costly and often finds only minor errors, and that **central statistical monitoring**, which uses the incoming data to detect unusual patterns, can be more effective at finding serious problems. Venet et al. (2012) described a method that compares each site's data distributions with those of all other sites, flagging sites whose data differ more than chance would explain, such as unusually low variability, implausible dates or too many identical values.

In practice, most coordinating centres combine the two. A regular report shows each site's recruitment, retention, missing data rate, protocol deviations and key distributions. Sites that fall outside agreed limits receive a call or visit. These reports also give site teams a clear view of their own performance relative to others, which tends to improve standards across the study.

Analysing data from several sites

Participants at the same site tend to be more alike than participants at different sites, and their outcomes are often correlated. Ignoring this **clustering** can produce standard errors that are too small and conclusions that are too confident. The strength of clustering is measured by the intraclass correlation: the share of total variation that lies between sites rather than within them. Even small values matter when each site contributes many participants, because the **design effect** grows with cluster size (Kish, 1965).

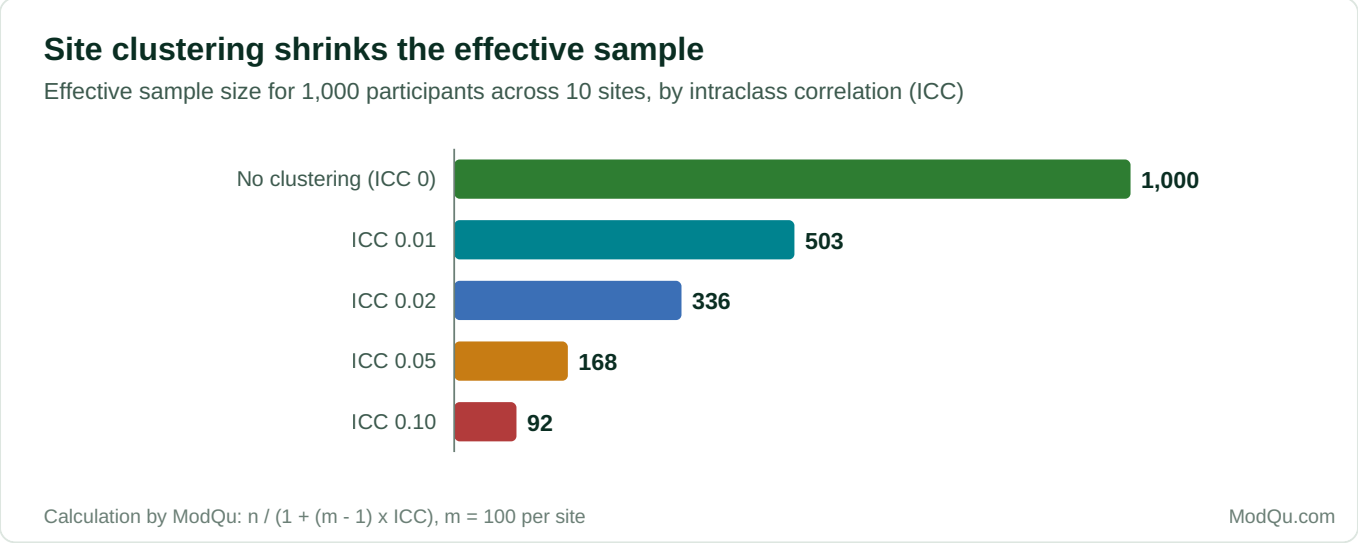


Figure 2. How clustering by site reduces the effective sample size for an overall estimate: 1,000 participants across 10 sites of 100. Illustrative calculation by ModQu using the design effect 1 + (m - 1) x ICC (Kish, 1965).

The illustrative calculation above shows the effect for a study of 1,000 participants spread evenly across 10 sites, when the aim is to estimate an overall average. With no clustering, the effective sample size is 1,000. With an intraclass correlation of 0.05, which is common in practice, it falls to about 168. The study is still useful, but it is much less precise than its raw size suggests.

When the study compares treatments within sites, as in a multi-site randomised trial, site effects can work in the researcher's favour, because each site acts as its own comparison. Localio et al. (2001) reviewed methods for adjusting for site, including fixed effects, random effects and generalised estimating equations. Kahan and Morris (2013) showed that when randomisation is stratified by site, the analysis should also account for site; otherwise, the standard errors are too large and power is lost. Raudenbush and Liu (2000) showed that in multi-site trials the power to detect an average effect depends on both the number of sites and the variation in the treatment effect across sites, so adding sites can be more valuable than adding participants within each site.

Because sites offer many possible ways to cut the data, the main analysis, including how site will be handled, should be specified in advance. Preregistering the analysis plan prevents the temptation to choose, after seeing the results, the subset of sites or the model that gives the most attractive answer (Nosek et al., 2018).

Site differences as findings

Variation between sites is not only a nuisance. When a programme works well in some sites and poorly in others, the pattern is often the most useful result of the study. Weiss et al. (2014) argued that multi-site evaluations should be designed to explain variation in effects, by measuring the treatment contrast, implementation, participant characteristics and context at each site, rather than reporting only the average. Kraemer (2000) similarly argued that differences between sites in a trial should be examined as evidence about the conditions under which a treatment works.

This requires planning. If the study is to explain why results differ between sites, it must collect data on the factors that might explain the differences, and it needs enough sites for the comparison to mean something. A study with three sites can describe their differences but can rarely explain them statistically. A study with twenty or more can begin to.

Multi-site research in commercial settings

Businesses run multi-site studies more often than they realise. A customer survey fielded in several countries, a test of a new service model in a sample of branches, and mystery shopping across a store network all face the same issues. A brief hypothetical illustrates them. A company runs a satisfaction survey in five countries and finds that scores in one country are 12 points lower. Before concluding that customers there are less satisfied, the research team should check whether the questionnaire was translated with the same care, whether the response scale is used differently in that culture, whether the sample was drawn in the same way and whether fieldwork took place at the same time. Each of these can produce a gap of that size on its own.

The practical lessons from academic multi-site research apply directly: a single protocol, a clear division of core and adaptable elements, careful translation, common data definitions, central monitoring during fieldwork and an analysis that accounts for site differences rather than ignoring them.

Conclusion

Multi-site research makes it possible to study larger samples, test whether findings hold across settings and learn from the differences between places. These benefits depend on coordination. A clear governance structure, a single protocol and manual, a deliberate choice of what to standardise and what to adapt, comparable measurement, consistent training, shared data systems and active monitoring together make the results from different sites comparable. An analysis that accounts for site effects, specified in advance, then turns a collection of local projects into one study whose conclusions are stronger than any single site could support.

References

Baigent, C., Harrell, F. E., Buyse, M., Emberson, J. R., & Altman, D. G. (2008). Ensuring trial validity by data quality assurance and diversification of monitoring methods. *Clinical Trials*, *5*(1), 49–55. https://doi.org/10.1177/1740774507087554

Brislin, R. W. (1970). Back-translation for cross-cultural research. *Journal of Cross-Cultural Psychology, 1*(3), 185–216. https://doi.org/10.1177/135910457000100301

Carroll, C., Patterson, M., Wood, S., Booth, A., Rick, J., & Balain, S. (2007). A conceptual framework for implementation fidelity. *Implementation Science, 2*, Article 40. https://doi.org/10.1186/1748-5908-2-40

Chan, A.-W., Tetzlaff, J. M., Altman, D. G., Laupacis, A., Gøtzsche, P. C., Krleža-Jerić, K., Hróbjartsson, A., Mann, H., Dickersin, K., Berlin, J. A., Doré, C. J., Parulekar, W. R., Summerskill, W. S. M., Groves, T., Schulz, K. F., Sox, H. C., Rockhold, F. W., Rennie, D., & Moher, D. (2013). SPIRIT 2013 statement: Defining standard protocol items for clinical trials. *Annals of Internal Medicine, 158*(3), 200–207. https://doi.org/10.7326/0003-4819-158-3-201302050-00583

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American Journal of Community Psychology, 41*(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0

Fowler, F. J., Jr., & Mangione, T. W. (1990). *Standardized survey interviewing: Minimizing interviewer-related error*. SAGE.

Groves, R. M., Fowler, F. J., Jr., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey methodology* (2nd ed.). Wiley.

Harkness, J. A., Braun, M., Edwards, B., Johnson, T. P., Lyberg, L., Mohler, P. Ph., Pennell, B.-E., & Smith, T. W. (Eds.). (2010). *Survey methods in multinational, multiregional, and multicultural contexts*. Wiley. https://doi.org/10.1002/9780470609927

Harkness, J. A., van de Vijver, F. J. R., & Mohler, P. Ph. (Eds.). (2003). *Cross-cultural survey methods*. Wiley.

Harris, P. A., Taylor, R., Thielke, R., Payne, J., Gonzalez, N., & Conde, J. G. (2009). Research electronic data capture (REDCap): A metadata-driven methodology and workflow process for providing translational research informatics support. *Journal of Biomedical Informatics*, *42*(2), 377–381. https://doi.org/10.1016/j.jbi.2008.08.010

Johnson, T. P., Pennell, B.-E., Stoop, I. A. L., & Dorer, B. (Eds.). (2019). *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)*. Wiley. https://doi.org/10.1002/9781118884997

Jowell, R., Roberts, C., Fitzgerald, R., & Eva, G. (Eds.). (2007). *Measuring attitudes cross-nationally: Lessons from the European Social Survey*. SAGE. https://doi.org/10.4135/9781849209458

Kahan, B. C., & Morris, T. P. (2013). Analysis of multicentre trials with continuous outcomes: When and how should we account for centre effects? *Statistics in Medicine*, *32*(7), 1136–1149. https://doi.org/10.1002/sim.5667

Kish, L. (1962). Studies of interviewer variance for attitudinal variables. *Journal of the American Statistical Association*, *57*(297), 92–115. https://doi.org/10.1080/01621459.1962.10482153

Kish, L. (1965). *Survey sampling*. Wiley.

Kraemer, H. C. (2000). Pitfalls of multisite randomized clinical trials of efficacy and effectiveness. *Schizophrenia Bulletin*, *26*(3), 533–541. https://doi.org/10.1093/oxfordjournals.schbul.a033474

Localio, A. R., Berlin, J. A., Ten Have, T. R., & Kimmel, S. E. (2001). Adjustments for center in multicenter studies: An overview. *Annals of Internal Medicine*, *135*(2), 112–123. https://doi.org/10.7326/0003-4819-135-2-200107170-00012

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213. https://doi.org/10.1037/1082-989X.5.2.199

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4–70. https://doi.org/10.1177/109442810031002

Venet, D., Doffagne, E., Burzykowski, T., Beckers, F., Tellier, Y., Genevois-Marlin, E., Becker, U., Bee, V., Wilson, V., Legrand, C., & Buyse, M. (2012). A statistical approach to central monitoring of data quality in clinical trials. *Clinical Trials*, *9*(6), 705–713. https://doi.org/10.1177/1740774512447898

Weinberger, M., Oddone, E. Z., Henderson, W. G., Smith, D. M., Huey, J., Giobbie-Hurder, A., & Feussner, J. R. (2001). Multisite randomized controlled trials in health services research: Scientific challenges and operational issues. *Medical Care*, *39*(6), 627–634. https://doi.org/10.1097/00005650-200106000-00010

Weiss, M. J., Bloom, H. S., & Brock, T. (2014). A conceptual framework for studying the sources of variation in program effects. *Journal of Policy Analysis and Management*, *33*(3), 778–808. https://doi.org/10.1002/pam.21760

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, Article 160018. https://doi.org/10.1038/sdata.2016.18