Supplementary material for: A Metadata Schema for Data from Experiments in the Social Sciences [*]

Jack Cavanagh[†] Jasmin Fliegner[‡] Sarah Kopper[†] Anja Sautmann[§]

December 28, 2022

C Additional Information

C.1 Other Resources

The Yale Application for Research Data (YARD) is an open-source web application that structures the process of curating research data for the ISPS Data Archive, a repository for affiliates of the Yale Institution for Social and Policy Studies (ISPS) (ISPS, 2021). YARD uses generic DDI fields to store a self-built vocabulary that can be read by DDI, providing a template for mapping our schema to DDI (Colectica, 2022).

The Registry of Efficacy and Effectiveness Studies (REES) at ICPSR is a registry for causal inference studies, including RCTs, in education research (Anderson et al., 2019).

Ohmann et al. (2017) identify current barriers and propose a set of principles related to sharing and reuse of clinical trial data.

The Clinical Data Interchange Standards Consortium (CDISC) develops data documentation standards for medical research that are used by many regulatory agencies such as the US Food and Drug Administration (National Cancer Institute (2021)).

The World Health organization manages the International Clinical Trial Repository Platform and defines the WHO Trial Registration Data Set, a minimum set of 24 items that a trial must report in order to be considered registered WHO (2021).

Raftery et al. (2015) build a metadata database with 429 data fields containing 125 clinical trials, with very detailed information on study conduct and results (as opposed to the data generated by the study),

[*] Additional materials for this project can be found on an associated GitHub repository.

[†] J-PAL/MIT, email: jcavanagh@povertyactionlab.org and skopper@povertyactionlab.org

[‡] The University of Manchester, email: jasmin.fliegner@manchester.ac.uk

[§] Development Economics Research Group, World Bank, email: asautmann@worldbank.org.

including e.g. information on adherence to protocol, quality of statistical and economic analyses, and costs.

Other platforms aimed at facilitating access to clinical studies harvest registries such as ClinicalTrials.gov: Vivli (2021), the Yoda Project (2021), Duke Clinical Research Institute (2021) and Clinical Study Data Request (CSDR (2021)).

C.2 Systematic Development of Controlled Vocabularies

We followed the procedure of the DDI Controlled Vocabulary working group to make changes to existing and develop new controlled vocabularies. We first defined the coverage of the metadata field, then reviewed existing controlled vocabularies, and finally made changes, such as expanding the available options or creating sub-categories of existing options, or (in some cases) by removing options. CV development was aided by inputs from the working group, multiple J-PAL staff, and the advisory group. We considered CVs from ADA, AidGrade, CESSDA, Clinicaltrials.gov,the Cross-National Equivalent File PSID Codebook, DDI, J-PAL, Gesis Demographic Variables, the IHSN, IPA, REES, and the World Bank. Although all newly developed CVs were tested extensively on the datasets below, we believe further testing with a larger body of data is required to fully cover the universe of values. The current version (version 0.9) of the proposed CVs is therefore hosted on the GitHub repository for this project, and we consider many of them subject to change.

C.3 Datasets included in the testing process

Note: Included are the 26 *public* datasets used in the testing process. Three more datasets were used from ongoing projects and will be added once the data are made public.

1. Ashraf, Nava; Field, Erica; Lee, Jean, 2018, “Replication Data for: Household Bargaining and Excess Fertility: An Experimental Study in Zambia”, https://doi.org/10.7910/DVN/6MSJHK, Harvard Dataverse, V1
2. Attanasio, Orazio; Augsburg, Britta; De Haas, Ralph; Fitzsimons, Emla; Harmgart, Heike, 2019-10-12, “Replication Data for: The Impacts of Microfinance: Evidence from Joint-Liability Lending in Mongolia.” Nashville, TN: American Economic Association [publisher], 2015. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E113597V1
3. Banerji, Rukmini; Berry, James; Shotland, Marc, 2019-10-12, “Replication data for: The Impact of Maternal Literacy and Participation Programs: Evidence from a Randomized Evaluation in India.” Nashville, TN: American Economic Association [publisher], 2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E113658V1
4. Beg, Sabrin; Lucas, Adrienne; Halim, Waqas; Saif, Umar, 2021, “Replication Data for: Engaging Teachers with Technology Increased Achievement, Bypassing Teachers Did Not”, https://doi.org/10.7910/DVN/Y1UZTT, Harvard Dataverse, V2, UNF:6:fpaSKwID+YXoTg8dulNYzg== [fileUNF]

5. Benhassine, Najy; Devoto, Florencia; Duflo, Esther; Dupas, Pascaline; Pouliquen, Victor, 2015, “Turning a Shove into A Nudge? A ‘Labeled Cash Transfer’ for Education”, https://doi.org/10.7910/DVN/29014, Harvard Dataverse, V1

6. Blair, Robert A.; Karim, Sabrina M.; Morse, Benjamin S., 2019, “Replication Data for: Establishing the Rule of Law in Weak and War-torn States: Evidence from a Field Experiment with the Liberian National Police”, https://doi.org/10.7910/DVN/ZIXH95, Harvard Dataverse, V1

7. Brune, Lasse; Chyn, Eric; Kerwin, Jason, 2020, “Replication Data for: Peers and Motivation at Work: Evidence from a Firm Experiment in Malawi”, https://doi.org/10.7910/DVN/CAXO8Y, Harvard Dataverse, V1, UNF:6:D5wZ+NVkUhWnHOd7uPrUdw== [fileUNF]

8. Cai, Jing; de Janvry, Alain; Sadoulet, Elisabeth, 2019-10-12, “Replication Data for: Social Networks and the Decision to Insure.” Nashville, TN: American Economic Association [publisher], 2015. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E113593V1

9. Deming, David J.; Yuchtman, Noam; Abulafi, Amira; Goldin, Claudia; Katz, Lawrence F., 2018, “Replication Data for: The Value of Postsecondary Credentials in the Labor Market: An Experimental Study”, https://doi.org/10.7910/DVN/P8ETJ0, Harvard Dataverse, V1

10. Dobbie, Will; Fryer, Jr., Roland G., 2018, “Are High-Quality Schools Enough to Increase Achievement among the Poor? Evidence from the Harlem Children’s Zone”, https://doi.org/10.7910/DVN/B7YLTV, Harvard Dataverse, V1

11. Duflo, Esther; Dupas, Pascaline; Ginn, Thomas; Pouliquen, Victor; Sharma, Vandana; Makana Barasa, Grace; Baraza, Moses, 2019, “HIV Prevention Among Youth: A Randomized Controlled Trial of VCT and Male Condom Distribution in Rural Kenya”, https://doi.org/10.7910/DVN/CVOPZL, Harvard Dataverse, V1

12. Duflo, Esther; Dupas, Pascaline; Kremer, Michael, 2013, “Data on Kenyan Youths”, https://doi.org/10.7910/DVN/TDFJ7X, Harvard Dataverse, V4

13. Duflo, Esther; Dupas, Pascaline; Kremer, Michael, 2016, “School Governance, Teacher Incentives, and Pupil–Teacher Ratios: Experimental Evidence from Kenyan Primary Schools”, https://doi.org/10.7910/DVN/9534YA, Harvard Dataverse, V1

14. Emerick, Kyle, 2020, “Trading frictions in Indian village economies”, https://doi.org/10.7910/DVN/90WZHM, Harvard Dataverse, V1, UNF:6:lQeC+ogeRN2HDrI5Mkg/EA== [fileUNF]

15. Fafchamps, Marcel; Quinn, Simon, 2019, “Replication Data for: Networks and Manufacturing Firms in Africa: Results from a Randomized Field Experiment”, https://doi.org/10.7910/DVN/HW9JQY, Harvard Dataverse, V1, UNF:6:MklmBxnQIjbQ4/OqEhNtdA== [fileUNF]

16. Fischer, Greg; Karlan, Dean, 2018, “The Catch-22 of External Validity in the Context of Constraints to Firm Growth”, https://doi.org/10.7910/DVN/MNW5LP, Harvard Dataverse, V1
17. Galiani, Sebastian; Gertler, Paul; Ajzenman, Nicolás; Orsola-Vidal, Alexandra, 2017, “Promoting Handwashing Behavior: The Effects of Large-Scale Community and School-Level Interventions”, https://doi.org/10.7910/DVN/0OTVLH, Harvard Dataverse, V1
18. Gerber, Alan S.; Karlan, Dean; Bergan, Daniel, 2018, “Replication Data for: Does the Media Matter? A Field Experiment Measuring the Effect of Newspapers on Voting Behavior and Political Opinions”, https://doi.org/10.7910/DVN/67UGGG, Harvard Dataverse, V1
19. Gilligan, Daniel O; Karachiwalla, Naureen; Kasirye, Ibrahim; Lucas, Adrienne M.; Neal, Derek, 2021, “Replication Data for: Educator Incentives and Educational Triage in Rural Primary Schools”, https://doi.org/10.7910/DVN/FJOL7N, Harvard Dataverse, V1, UNF:6:9T9h8aWC5AXitExRjEjZPg== [fileUNF]
20. Hanna, Rema; Duflo, Esther; Greenstone, Michael, 2018, “Replication Data for: Up in Smoke: The Influence of Household Behavior on the Long-Run Impact of Improved Cooking Stoves”, https://doi.org/10.7910/DVN/T4APMN, Harvard Dataverse, V1
21. Hartman, Alexandra; Blair, Robert; Blattman, Christopher, 2019, “Replication Data for: Engineering Informal Institutions: Long Run Impacts of Alternative Dispute Resolution on Violence and Property Rights in Liberia”, https://doi.org/10.7910/DVN/MN1OON, Harvard Dataverse, V1
22. Malesky, Edmund, 2019, “Replication Data for: Participation, Government Legitimacy, and Regulatory Compliance in Emerging Economies: A Firm-Level Field Experiment in Vietnam”, https://doi.org/10.7910/DVN/IANHOG, Harvard Dataverse, V1, UNF:6:45Z4bwhxkL4Z3NZZTFLn9Q== [fileUNF]
23. Olken, Benjamin A.; Onishi, Junko; Wong, Susan, 2019-10-12, “Replication data for: Should Aid Reward Performance? Evidence from a Field Experiment on Health and Education in Indonesia.” Nashville, TN: American Economic Association [publisher], 2014. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], https://doi.org/10.3886/E113906V1
24. Oreopoulos, Philip, 2018, “Replication Data for: Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes”, https://doi.org/10.7910/DVN/WXMJCN, Harvard Dataverse, V1
25. Pons, Vincent; Liegey, Guillaume, 2018, “Increasing the Electoral Participation of Immigrants: Experimental Evidence from France”, https://doi.org/10.7910/DVN/DDCNEW, Harvard Dataverse, V1
26. Rockoff, Jonah E.; Staiger, Douglas O.; Kane, Thomas J.; Taylor, Eric S., 2018, “Information and Employee Evaluation: Evidence from a Randomized Intervention in Public Schools”, https://doi.org/10.7910/DVN/1OFPOU, Harvard Dataverse, V1

References

Anderson, D., J. Spybrook, and R. Maynard (2019). Rees: A registry of efficacy and effectiveness studies in education. *Educational Researcher 48*(1), 45–50.

Colectica (2022). Colectica 7.0 - DDI Lifecycle in Colectica curation tools. URL: https://docs.colectica.com/curation/technical-documentation/ddi-mapping/ (04/27/2022).

CSDR (2021). Clinicalstudydatarequest.com. URL: https://clinicalstudydatarequest.com/Default.aspx (11/05/2021).

Duke Clinical Research Institute (2021). SOAR DATA. URL: https://dcri.org/our-work/analytics-and-data-science/data-sharing/soar-data/ (11/05/2021).

ISPS (2021). YARD at ISPS. URL: https://isps.yale.edu/research/data/deposit/yard (12/16/2022).

National Cancer Institute (2021). CDISC Terminology. URL: https://datascience.cancer.gov/resources/cancer-vocabulary/cdisc-terminology (11/05/2021).

Ohmann, C., R. Banzi, S. Canham, S. Battaglia, M. Matei, C. Ariyo, L. Becnel, B. Bierer, S. Bowers, L. Clivio, M. Dias, C. Druml, H. Faure, M. Fenner, J. Galvez, D. Ghersi, C. Gluud, T. Groves, P. Houston, G. Karam, D. Kalra, R. L. Knowles, K. Krleža-Jerić, C. Kubiak, W. Kuchinke, R. Kush, A. Lukkarinen, P. S. Marques, A. Newbigging, J. O'Callaghan, P. Ravaud, I. Schlünder, D. Shanahan, H. Sitter, D. Spalding, C. Tudur-Smith, P. van Reusel, E.-B. van Veen, G. R. Visser, J. Wilson, and J. Demotes-Mainard (2017). Sharing and reuse of individual participant data from clinical trials: principles and recommendations. *BMJ Open 7*(12).

Raftery, J., A. Young, L. Stanton, R. Milne, A. Cook, D. Turner, and P. Davidson (2015). Clinical trial metadata: defining and extracting metadata on the design, conduct, results and costs of 125 randomised clinical trials funded by the National Institute for Health Research Health Technology Assessment programme. *Health Technology Assessment 19*(11), 1–166.

Vivli (2021). Center for clinical research data. URL: https://search.vivli.org/ (11/05/2021).

WHO (2021). Who trial registration data set (version 1.3.1). URL: https://www.who.int/clinical-trials-registry-platform/network/who-data-set (11/05/2021).

Yoda Project (2021). Welcome to the Yoda Project. URL: https://yoda.yale.edu/welcome-yoda-project (11/05/2021).