Data Cleaning

CCTS Biostatistics Core

June 2023

This page includes publications and tools that our consultants have found useful. For more information on this topic, including advice about how to apply it in your research, consider scheduling a consultation with a biostatistician.

While we hope this resource list serves as a helpful starting point for other researchers, we provide no guarantee of its comprehensiveness or of the accuracy or reliability of the works cited. If you have concerns or suggestions to improve this page, please contact us.

Resources

Bonner, Anne (2019). *The complete beginner's guide to data cleaning and preprocessing*. Python, simple. https://towardsdatascience.com/the-complete-beginners-guide-todata-cleaning-and-preprocessing-2070b7d4c6d.

de Jonge, Edwin and Mark van der Loo (2013). *An introduction to data cleaning with R*. PDF, extensive. https://cran.r-project.org/doc/contrib/de_Jonge+van_der_Loo-Introduction_to_data_cleaning_with_R.pdf.

Elgabry, Omar (2019). *The ultimate guide to data cleaning*. https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4.

Kantarci, Atakan (2020). *Data cleaning: What it is, why it matters, best practices & tools*. Business case, overview, colorful charts. https://research.aimultiple.com/data-cleaning.

Pomerantseva, Vera (2009). *Clinical data cleaning and validation steps*. Not language-specific. https://pharmaceuticalprocessingworld.com/clinical-data-cleaning-and-validation-steps/.

Regional Educational Laboratory Central (2021). *Common Sources of Data Errors and Error-Checking Techniques*. Many resources for data management. https://ies.ed.gov/ncee/rel/Products/Region/central/Resource/100644/26.

Sciforce (2019). *Data cleaning and processing for beginners*. Python, simple. https://medium.com/sciforce/data-cleaning-and-processing-for-beginners-25748ee00743.

Society for Clinical Data Management (SCDM) (2013). *Good Clinical Data Management Practices (GCDMP)*. PDF, 524 pages. https://scdm.org/wp-content/uploads/2019/10/21117-Full-GCDMP-Oct-2013.pdf.

Van den Broeck, Jan, Solveig Argeseanu Cunningham, Roger Eeckels, et al. (2005). "Data Cleaning: Detecting, Diagnosing, and Editing Data Abnormalities". In: *PLoS Medicine* 2.10, p. e267. DOI: 10.1371/journal.pmed.0020267. https://doi.org/10.1371/journal.pmed.0020267.

Willems, Karlijn (2017). *An introduction to cleaning data in R*. R, simple. https://www.datacamp.com/community/blog/an-introduction-to-cleaning-data-in-r.