

# How Good Research Data Management Enables **Reproducibility and Data Reuse**: **A Case Study of Clinical Trial Data**

**Leonhard Held**

University of Zurich

16.02.2026

- Professor of Biostatistics  
[www.biostat.uzh.ch](http://www.biostat.uzh.ch)
- UZH Open Science Delegate  
[www.openscience.uzh.ch](http://www.openscience.uzh.ch)
- Director, Center for Reproducible Science and Research Synthesis [www.crs.uzh.ch](http://www.crs.uzh.ch)
- Steering Committee, Swiss Reproducibility Network [www.swissrn.org](http://www.swissrn.org)

# 20 years ago ...

## Abstract

# Trials

### Background:

genomics being  
routinely refuse

Commentary

Open Access

### Discussion:

Designing secondar  
design of future  
Clinical trialists,  
misrepresentati

## Whose data set is it anyway? Sharing raw data from randomized trials

Andrew J Vickers\*

inclusion of the original trialists as co-authors on any publication resulting from data sharing. Moreover, if we treat any data set as belonging to the patients who comprise it, rather than the investigators, such concerns fall away.

2006, Trials

**Conclusion:** Technological developments, particularly the Internet, have made data sharing generally a trivial logistical problem. Data sharing should come to be seen as an inherent part of conducting a randomized trial, similar to the way in which we consider ethical review and publication of study results. Journals and funding bodies should insist that trialists make raw data available, for example, by publishing data on the Web. If the clinical trial community continues to fail with respect to data sharing, we will only strengthen the public perception that we do clinical trials to benefit ourselves, not our patients.

# 20 years ago ...

---

## Reproducible Epidemiologic Research

**Roger D. Peng, Francesca Dominici, and Scott L. Zeger**

From the Biostatistics Department, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD.

*Received for publication November 4, 2005; accepted for publication January 10, 2006.*

---

The replication of important findings by multiple independent investigators is fundamental to the accumulation of scientific evidence. Researchers in the biologic and physical sciences expect results to be replicated by independent data, analytical methods, laboratories, and instruments. Epidemiologic studies are commonly used to quantify small health effects of important, but subtle, risk factors, and replication is of critical importance where results can inform substantial policy decisions. However, because of the time, expense, and opportunism of many current epidemiologic studies, it is often impossible to fully replicate their findings. An attainable minimum standard is “reproducibility,” which calls for data sets and software to be made available for verifying published findings and conducting alternative analyses. The authors outline a standard for reproducibility and evaluate the reproducibility of current epidemiologic research. They also propose methods for reproducible research and implement them by use of a case study in air pollution and health.

# Replication: Misuse of scarce expertise?

## Commentary

D. R. COX\*

*Nuffield College, University of Oxford, Oxford OX1 1NF, UK*  
david.cox@nuffield.ox.ac.uk

CHRISTL DONNELLY

*MRC Centre for Outbreak Analysis and Modelling, Imperial College, London W2 1PG, UK*

Professor Keiding makes an important distinction between using data to illustrate the performance of statistical methods and using statistical methods to extract important subject-matter information from data. The former may be useful but, as Professor Keiding emphasizes, the latter may involve subtle considerations about the interplay between subject-matter and statistical aspects and the detailed nature of the data and its compilation.

From this perspective, the suggestion of requiring independent replication of specific statistical analyses as a general check before publication seems not merely unnecessary but a misuse of relatively scarce expertise.

The position of Donnelly that [17] ‘the suggestion of requiring independent replication of specific statistical analyses as a general check before publication seems not merely unnecessary but a misuse of relatively scarce expertise’, needs revisiting. The present case underlines the obligation not only for rigorous checks of statistical analysis but also for validation of the statistical models and assumptions used within submitted manuscripts to verify them. Accordingly, a very substantial number of publications that rest extensively or completely on RBCT statistical analyses may require major qualification or retraction. The justification for lethal control of badgers to date appears to have been based upon basic statistical oversight.

ROYAL SOCIETY  
OPEN SCIENCE

royalsocietypublishing.org/journal/rsos

CC BY

Comment



**Cite this article:** Torgerson PR, Hartnack S, Rasmussen P, Lewis FI, O'Donnell P, Langton TES. 2025 Randomised Badger Culling Trial—no effects of widespread badger culling on tuberculosis in cattle: comment on Mills, Woodroffe and Donnelly (2024a, 2024b). *R. Soc. Open Sci.* 12: 241609.  
<https://doi.org/10.1098/rsos.241609>

Received: 1 October 2024

Accepted: 9 May 2025

**Subject Category:**

Ecology, conservation, and global change biology

Randomised Badger Culling Trial—no effects of widespread badger culling on tuberculosis in cattle: comment on Mills, Woodroffe and Donnelly (2024a, 2024b)

Paul R. Torgerson<sup>1</sup>, Sonja Hartnack<sup>1</sup>, Philip Rasmussen<sup>1,2</sup>, Fraser I. Lewis<sup>1,3</sup>, Peter O'Donnell<sup>4</sup> and Thomas E. S. Langton<sup>5</sup>

<sup>1</sup>Veterinary Epidemiology, University of Zurich, Zurich, Switzerland

<sup>2</sup>Department of Veterinary and Animal Sciences, University of Copenhagen, Copenhagen, Denmark

<sup>3</sup>Private Statistical Consultant, London, UK

<sup>4</sup>Department of Mathematics, University of Cambridge, Cambridge, UK

<sup>5</sup>Herpetofauna Consultants International, Halesworth, Suffolk, UK

# Errors may happen ...

Research

## This article has been retracted

JAMA | Original Investigation

### Effect of a Program Combining Transitional Care and Long-term Self-management Support on Outcomes of Hospitalized Patients With Chronic Obstructive Pulmonary Disease A Randomized Clinical Trial

Hanan Aboumatar, MD, MPH; Mohammad Naqibuddin, MBBS, MPH; Suna Chung, MPH; Hina Chaudhry, MPH; Samuel W. Kim, BA; Jamia Saunders, MD, MS; Lee Bone, MPH; Ayse P. Gurses, MS, MPH, PhD; Amy Knowlton, ScD, MPH; Peter Pronovost, MD, PhD; Nirupama Putcha, MD, MHS; Cynthia Rand, PhD; Debra Roter, DrPH; Carol Sylvester, RN, MS; Carol Thompson, MS, MBA; Jennifer L. Wolff, PhD; Judith Hibbard, PhD, MPH, FCCM; Robert A. Wise, MD

**CONCLUSIONS AND RELEVANCE** In a single-site randomized clinical trial of patients hospitalized due to COPD, a 3-month program that combined transition and long-term self-management support resulted in significantly fewer COPD-related hospitalizations and emergency department visits and better health-related quality of life at 6 months after discharge. Further research is needed to evaluate this intervention in other settings.

**CONCLUSIONS AND RELEVANCE** In a single-site randomized clinical trial of patients hospitalized due to COPD, a 3-month program that combined transition and long-term self-management support resulted in significantly greater COPD-related hospitalizations and emergency department visits, without improvement in quality of life. Further research is needed to determine reasons for this unanticipated finding.

Aboumatar, JAMA 2018, [doi:10.1001/jama.2018.17933](https://doi.org/10.1001/jama.2018.17933)

The identified programming error was in a file used for preparation of the analytic data sets for statistical analysis and occurred while the variable referring to the study “arm” (ie, group) assignment was recoded. The purpose of the recoding was to change the randomization assignment variable format of “1, 2” to a binary format of “0, 1.” However, the assignment was made incorrectly and resulted in a reversed coding of the study groups. Even though the data analyst created and conducted some test analysis programs, they were of the type that did not show any labeling of the arm categories, only the “arm” variable in a regression, for example. After detecting this error, we promptly reported it to our institutional review board and appropriate offices within our university, alerted JAMA, and proceeded to confirm whether the error had affected the analytic data sets, which we found to be the case. We therefore started a complete data reanalysis, with 2 biostatisticians performing double programming and an independent analysis of study primary outcomes to ensure the validity of the reported results. As noted here, this reanalysis showed reversed study findings, with a higher number of hospitalizations and emergency department visits in the intervention compared with the usual care group.

# Clinical data sharing

Statistical controversies in clinical research: data access and sharing—can we be more transparent about clinical research? Let's do what's right for patients

F. W. Rockhold

Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University Medical Center, Durham, USA

Correspondence to: Prof. Frank W. Rockhold, Department of Biostatistics and Bioinformatics, Duke Clinical Research Institute, Duke University Medical Center, PO Box 17969, Durham, NC 27715, USA. Tel: +1-919-668-1073; E-mail: frankrockhold@duke.edu

Calls for greater transparency and 'open data access' in clinical research are widespread, from sources including the Executive Office of the President, which in 2013 called for increased access to the results of federally funded research. In 2015, The Institute of Medicine issued a report advocating for a multi-stakeholder effort to foster responsible data sharing, and there are many others. Open science is good for researchers, good for innovation, and good for patients. The question at the center of the open-science efforts for clinical trials should not be *whether* data should be shared, but rather how we can usher in responsible methods for doing so. Unfortunately, there remain numerous perceived barriers to complete transparency around clinical trial data. This paper reviews the current status of data disclosure, the barriers to achieving it and a suggestion for the future.

**Key words:** open science, transparency, data sharing, data access

**Ethical obligation towards participating patients**

But data sharing platforms require:

- **common data standards**
- **standards for data use agreements**

*Annals of Oncology* 28: 1734–1737, 2017  
doi:10.1093/annonc/mdx123  
Published online 5 April 2017

# Data sharing platforms



The screenshot shows the Vivli website homepage. At the top left is the Vivli logo, a globe with a grid, and the text "Vivli" in a bold, sans-serif font. Below the logo is a small yellow banner that reads "CENTER FOR GLOBAL CLINICAL RESEARCH DATA". To the right of the logo is a navigation menu with links for "Home", "About", "Members", "News & Events", "Resources", "Portals", and a "LOG IN" button. The main content area features a large, abstract background image of a globe with glowing data points. The text "A global clinical research data sharing platform" is centered in a large, white, sans-serif font. Below this is a smaller line of text: "The Vivli team is dedicated to helping researchers share and access data from clinical trials to advance science." At the bottom of the main content area are two yellow buttons: "SEARCH FOR STUDIES" and "SUBMIT YOUR STUDY". The footer is a solid blue bar with four white boxes, each containing a statistic: "8,000+" (CLINICAL TRIALS), "55+" (MEMBERS), "5.2 MILLION" (PARTICIPANTS), and "530+" (PUBLICATIONS).

**Vivli**  
CENTER FOR GLOBAL CLINICAL RESEARCH DATA

Home About Members News & Events Resources Portals LOG IN

A global clinical research data sharing platform

The Vivli team is dedicated to helping researchers share and access data from clinical trials to advance science.

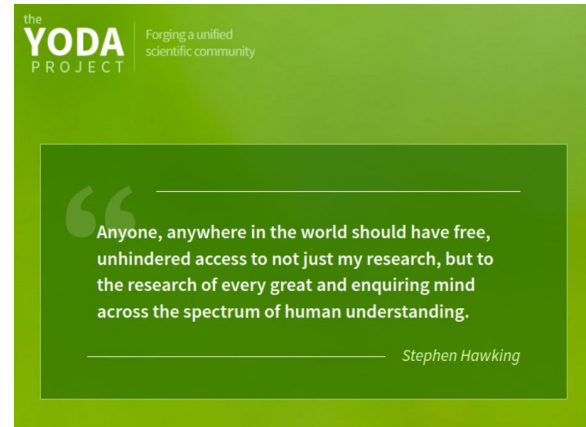
SEARCH FOR STUDIES SUBMIT YOUR STUDY

8,000+ CLINICAL TRIALS

55+ MEMBERS

5.2 MILLION PARTICIPANTS

530+ PUBLICATIONS



The banner for "the YODA PROJECT" is set against a green background. The text "the YODA PROJECT" is in white, with "YODA" in a larger font. To the right is the tagline "Forging a unified scientific community". Below this is a quote in white text: "Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding." The quote is attributed to "Stephen Hawking" at the bottom right.

the **YODA** PROJECT  
Forging a unified scientific community

“  
Anyone, anywhere in the world should have free, unhindered access to not just my research, but to the research of every great and enquiring mind across the spectrum of human understanding.  
”  
Stephen Hawking



# Share-CTD: An EU-funded Doctoral Training Network

## Aims:

- to train a new generation of biomedical researchers in **clinical trial data sharing**
- implementing **Open Science** in medical research

## Research projects:

- Impact of preparing data for sharing
- Using shared data

<https://sharectd.eu/>



SHARE-CTD: Start 01.01.2024, End: 31.12.2027



# Share-CTD

- Provides structured professional development through schoolings, datathons and professional trainings
- **Datathon:** intensive workshop where participants use clinical trial data to answer specific research questions

	Main Training Events & Conferences	ECTS	Lead Institution / Place
1	Kick-off propaedeutic training	2	LMU / Munich
2	School (S1)	4	UHZ / Zürich
3	Datathon (D1)	4	UMG / Göttingen
4	Professional Training (PT1)	1	
5	School (S2)	4	UMCU / Lyon
6	Datathon (D2)	4	UMCU / Utrecht
7	Professional Training (PT2)	1	
8	School (S3)	4	UR1 / Online
9	Datathon (D3)	4	UR1 / Paris
10	Professional Training (PT3)	1	
11	Final conference	1	LMU + BIH/ Berlin

# Original study

---

## Oxaliplatin added to fluorouracil-based preoperative chemoradiotherapy and postoperative chemotherapy of locally advanced rectal cancer (the German CAO/ARO/AIO-04 study): final results of the multicentre, open-label, randomised, phase 3 trial

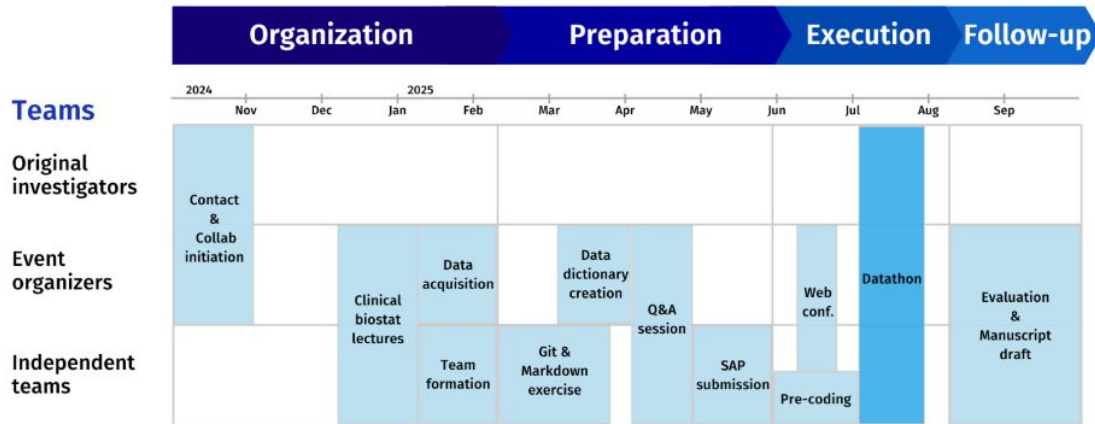


*Claus Rödel\*, Ullrich Graeven\*, Rainer Fietkau, Werner Hohenberger, Torsten Hothorn, Dirk Arnold, Ralf-Dieter Hofheinz, Michael Ghadimi, Hendrik A Wolff, Marga Lang-Welzenbach, Hans-Rudolf Raab, Christian Wittekind, Philipp Ströbel, Ludger Staib, Martin Wilhelm, Gerhard G Grabenbauer, Hans Hoffmanns, Fritz Lindemann, Anke Schlenska-Lange, Gunnar Folprecht, Rolf Sauer\*, Torsten Liersch\*, on behalf of the German Rectal Cancer Study Group†*

**Lancet Oncol 2015; 16: 979–89**

# Data availability

- **Anonymized data** was acquired in Jan 2025 in .rda format.
- Access to the material secured on a **closed cloud storage** was provided.
- Each team was required to complete an exercise on **Git and Markdown**.
- The original dataset was not accompanied by a data dictionary, thus a **centralized data dictionary** was prepared by the event organizer.



# Datathon 1



	Sunday 7/6/2025	Monday 7/7/2025	Tuesday 7/8/2025	Wednesday 7/9/2025	Thursday 7/10/2025	Friday 7/11/2025
06:30-07:45		breakfast	breakfast	breakfast	breakfast	breakfast
08:30-10:00	Arrival (Sunday or Monday)	Arrival (Sunday or Monday)	DT 2, Session 1: Understanding the data and the process how they were collected [UM]	Biostatistical Aspects and Reproducibility 3: Subgroup Analysis & Competing Risks [LH]	CT: methods of data sharing – from locking in, data transfer to federated learning - examples [US]	*ADMINISTRATION* CT: Update Protocols, regulatory affairs, EHDS including DT: news from the YODA application + paper [UM]
10:00-10:30			coffee break	coffee break	coffee break	coffee break
10:30-12:00		DT 2, Session 2: Spirit 2025 & Consort 2025 [UM]	Biostatistical Aspects and Reproducibility 4: working on a current trial [LH]	DT 4: time to discuss and fix presentations	Feedback Round with Pls and DCs [US]	
12:00-13:00		lunch	lunch	lunch	lunch	lunchbox
13:30-15:00		DT 1, Session 1: introduction into the medical field by a medical doctor [TL]	Biostatistical Aspects and Reproducibility 1: Analysis of Primary Outcome [LH]	DT 3, Session 1 DCs groupwork & <b>PI Meeting</b>	DT 4: present and discuss results I - discussant and peer review	Departure (Friday or Saturday)
15:00-15:30		coffee break	coffee break	coffee break	coffee break	
15:30-17:00		DT 1, Session 2: introduction into the data set, PICO [JK, TL]	Biostatistical Aspects and Reproducibility 2: working on a current trial [LH]	DT 3, Session 2 DCs groupwork	DT 4: present and discuss results II - discussant and peer review	
17:00		Barbecue at GWDG	transfer	DT 3: Deadline report for the discussants		
18:00			17:30-18:30 City Tour			
19:00-21:00			dinner at Bullerjahn	dinner at Tante Giulia	dinner at Zum Szützenbürger	

DT: Datathon  
Rep. & CB: Reproducibility & Clinical Biostatistics  
CT: Clinical Trials

TL: Torsten Liersch  
JK: Johanna Kreutzer  
UM: Ulrich Mansmann  
LH: Leonhard Held  
US: Ulrich Sax

















# Datathon 1, Göttingen 2025



# Results

 Follow this preprint

## Supporting Reanalysis and Reuse of Clinical Trial Data: A Case Study

 Cora Burgwinkel,  Han Chang Chiam,  Ka Hin Tai,  Jifan Wang,  Mian Haider Ali, Salman Soleiman Fallah,  Minoo Matbouriahi,  Tobechei Obinwanne,  Grigoriu Papapostolou,  Muhammad Riedha,  Giulia Varvara,  Yazid Zalai,  Ulrich Mansmann,  Ulrich Sax,  Leonhard Held

doi: <https://doi.org/10.1101/2025.11.06.25339683>

	Reproducibility				Robustness
	Team 1	Team 2	Team 3	Team 4	Team 5
Log-rank test and Cox proportional hazard model for DFS	R	R	R	P	R'
Table 1: Baseline characteristics	P§	P§	P§	-	P'§
Table 2: First events for primary endpoint DFS	P*	P*	P*	-	P*
Table 3: All-cause deaths	R	R	R	-	R
Figure 1: Trial profile	P	R	R	-	R
Figure 2: Kaplan-Meier curves of DFS (A) and OS (B)	R	R	R	P	P'
Figure 3: Cumulative incidence of locoregional recurrences (A) and distant recurrences (B)	p*	p*	p*	p*	p'*
Figure 4: Subgroup analyses on DFS	R	R	R	P	P'
Pathological complete response (pCR)	-	-	P	P	-
R0 resection rate, Sphincter preservation, grade 3 acute toxicity	-	-	-	P	-

R: reproducible; P: partially reproducible; R': robust; P': partially robust -: did not reproduce

§ Median age in the control group is not reproducible across teams

\* Partially reproducible or robust due to the lack additional data

# Discussion

Reproducing the findings of an RCT is feasible and provides **meaningful scientific value**:

- validating the original findings
- revealing discrepancies that warrant further investigation

Availability of data, code and statistical analysis plan (SAP) and **appropriate reporting** is vital.

Data sharing and reuse generally assume the underlying data are of **high quality** and **correctly recorded**.

# Reporting Guidelines: 2025 Updates

nature medicine

Consensus Statement

<https://doi.org/10.1038/s41591-025-03668-w>

## SPRIT 2025 statement: updated guideline for protocols of randomized trials

nature medicine

Consensus Statement

<https://doi.org/10.1038/s41591-025-03635-5>

## CONSORT 2025 statement: updated guideline for reporting randomized trials

Consensus Statement

<https://doi.org/10.1038/s41591-025-03635-5>

Table 1 | CONSORT 2025 checklist to include when reporting a randomized trial

Section/topic	No	CONSORT 2025 checklist item description
<b>Title and abstract</b>		
Title and structured abstract	1a	Identification as a randomized trial
	1b	Structured summary of the trial design, methods, results, and conclusions
<b>Open science</b>		
Trial registration	2	Name of trial registry, identifying number (with URL) and date of registration
Protocol and statistical analysis plan	3	Where the trial protocol and statistical analysis plan can be accessed
Data sharing	4	Where and how the individual de-identified participant data (including data dictionary), statistical code and any other materials can be accessed
Funding and conflicts of interest	5a	Sources of funding and other support (eg, supply of drugs), and role of funders in the design, conduct, analysis and reporting of the trial
	5b	Financial and other conflicts of interest of the manuscript authors

# Data sharing at the review stage

For clinical trials, the **FDA** routinely access individual-level data and **independently reproduce analyses** as part of the regulatory process leading to possible marketing approval (This is not the case for investigator-initiated clinical trials).



## When should data and code be made available?

The sharing of data and code as part of a research publication is crucial for ensuring the computational reproducibility of scientific work. Statisticians and data scientists already have the skills and tools to share all research material at the manuscript submission stage. However, many obstacles need to be overcome before they can lead by example.  
By **Rachel Heyard** and **Leonhard Held**

# Further reading



Center for Reproducible Science and Research Synthesis

## Sharing Clinical Trial Data

[Hester van de Wiel](#), [Cora Burgwinkel](#), [Leonhard Held](#)

doi: [10.5281/zenodo.13860164](https://doi.org/10.5281/zenodo.13860164)



## Primer: Conducting an Individual Participant Data Meta-Analysis

[Gorka Fraga González](#), [Fabio Molo](#), [Leonhard Held](#)

doi: [10.5281/zenodo.17150854](https://doi.org/10.5281/zenodo.17150854)



## Clinical research data sharing in Switzerland in a nutshell

[Sarah R Haile](#), [Malwina Kowalska](#), [Hester van de Wiel](#), [Markus Golder](#), [Leonhard Held](#), [Ulrike Held](#)

doi: [10.5281/zenodo.13860399](https://doi.org/10.5281/zenodo.13860399)



## Primer: Dynamic Reporting

[Felix Hofmann](#), [Samuel Pawel](#), [Monika Hebeisen](#), [Leonhard Held](#)

doi: [10.5281/zenodo.7565735](https://doi.org/10.5281/zenodo.7565735)



# Data Sharing - Good Practice

<https://www.youtube.com/watch?v=wjWAUrvA6c4>



The image shows a slide from a video titled "UKRN Primer Animation Data Sharing". The slide is split into two main sections. On the left, there is a dark blue vertical panel containing the UKRN logo (a globe with "UKRN" text), the website "www.ukrn.org", and the Twitter handle "@ukrepro". On the right, the text "UKRN Primer Animation" is at the top, followed by "Data Sharing" in large pink letters. Below the text is a circular illustration of two blue cartoon figures sitting at a table with a laptop, with a speech bubble above them that says "RESEARCH DATA".

UKRN Primer Animation

## Data Sharing



www.ukrn.org

@ukrepro



# Data Sharing - Bad Practice Part 1

<https://www.youtube.com/watch?v=RVZbk3GEVSw&t=15s>



# Data Sharing - Bad Practice Part 2

<https://www.youtube.com/watch?v=RtSv0gSbCP8>



# Data Sharing - Bad Practice Part 3

<https://www.youtube.com/watch?v=-MIH8PkuUo4>

