INTRODUCTION TO FAIR PRINCIPLES AND DATA MANAGEMENT

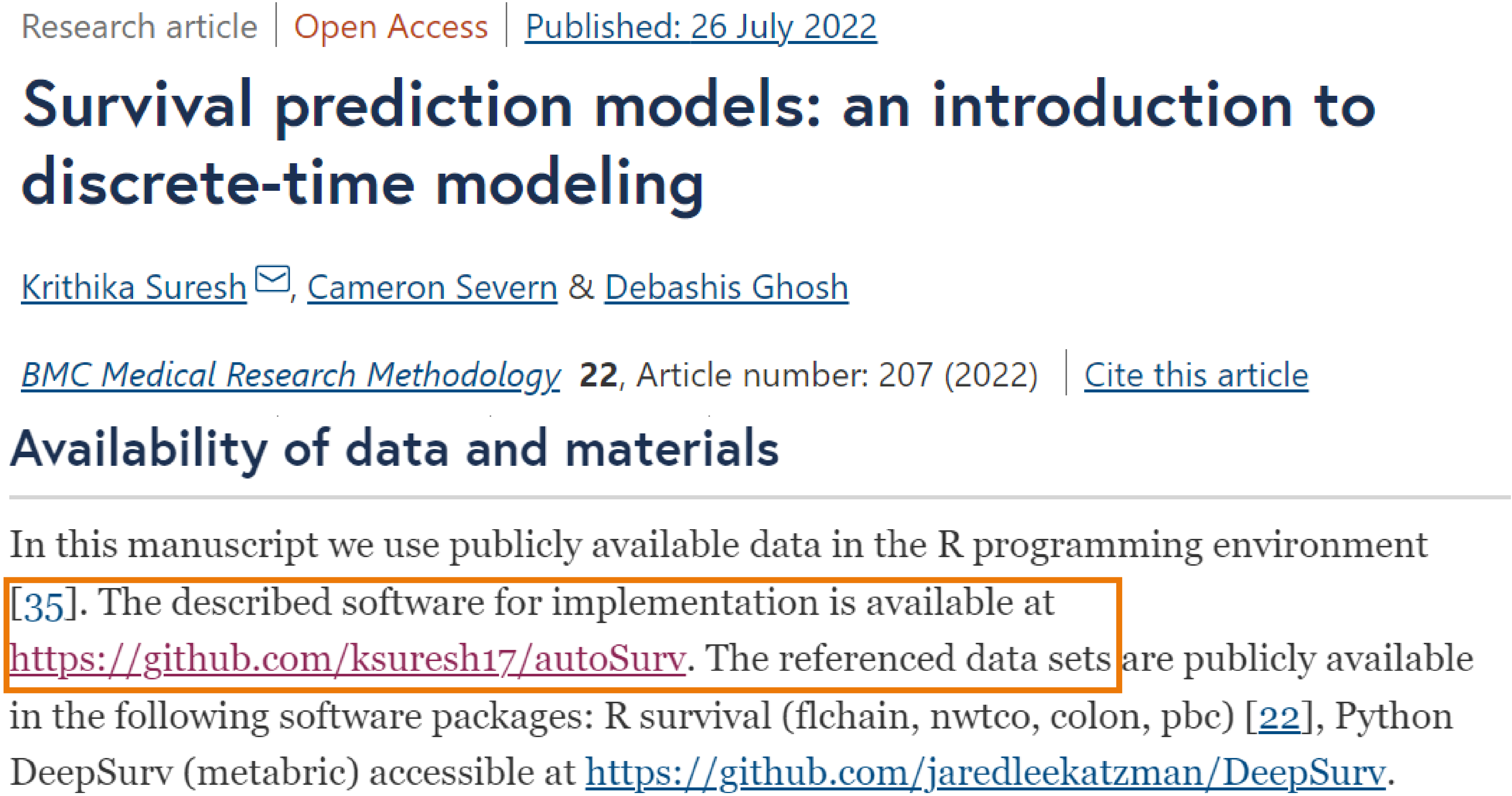
WARM UP

- What does FAIR stand for in the scientific research context?
    - **F**indable, **A**ccessible, **I**nteroperable, **R**eusable
- What do you think is needed to make your data FAIR?
    - Persistent link / identifiers
    - Metadata (human and machine readable)
    - The data…?
    - …
- Why is it important?

LEARNING OBJECTIVES

- After this lecture, you should be able to:
  - Explain what the FAIR principles are
  - Explain the value of metadata and data archiving

A STORY OF MY LIFE...



Research article | Open Access | Published: 26 July 2022

Survival prediction models: an introduction to discrete-time modeling

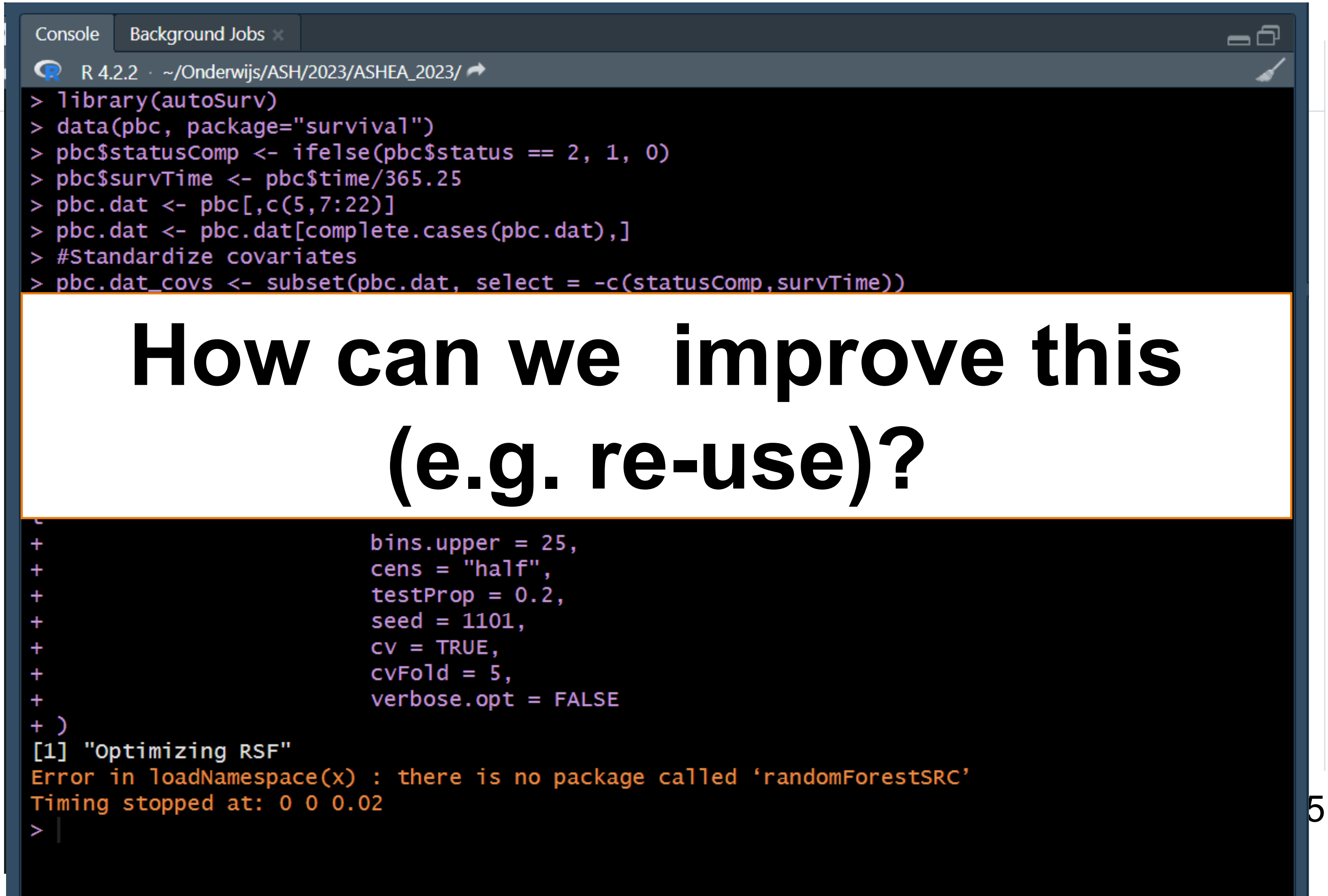
Krithika Suresh, Cameron Severn & Debashis Ghosh

BMC Medical Research Methodology **22**, Article number: 207 (2022) | Cite this article

Availability of data and materials

In this manuscript we use publicly available data in the R programming environment [35]. The described software for implementation is available at https://github.com/ksuresh17/autoSurv. The referenced data sets are publicly available in the following software packages: R survival (flchain, nwtco, colon, pbc) [22], Python DeepSurv (metabric) accessible at https://github.com/jaredleekatzman/DeepSurv.

A STORY OF MY LIFE...



```
> library(autoSurv)
> data(pbc, package="survival")
> pbc$statusComp <- ifelse(pbc$status == 2, 1, 0)
> pbc$survTime <- pbc$time/365.25
> pbc.dat <- pbc[,c(5,7:22)]
> pbc.dat <- pbc.dat[complete.cases(pbc.dat),]
> #Standardize covariates
> pbc.dat_covs <- subset(pbc.dat, select = -c(statusComp,survTime))
```

How can we improve this (e.g. re-use)?

```
+                          bins.upper = 25,
+                          cens = "half",
+                          testProp = 0.2,
+                          seed = 1101,
+                          cv = TRUE,
+                          cvFold = 5,
+                          verbose.opt = FALSE
+ )
[1] "Optimizing RSF"
Error in loadNamespace(x) : there is no package called ‘randomForestSRC’
Timing stopped at: 0 0 0.02
> 
```

THE FAIR PRINCIPLES – WHY?

- We do not maximise the benefits of data publication
- Increasing number of non-special-purpose repositories (ZENODO, DANS, Dataverse)
- Different data formats and documentation

→ Difficult to re-use data, software, algorithms, workflows

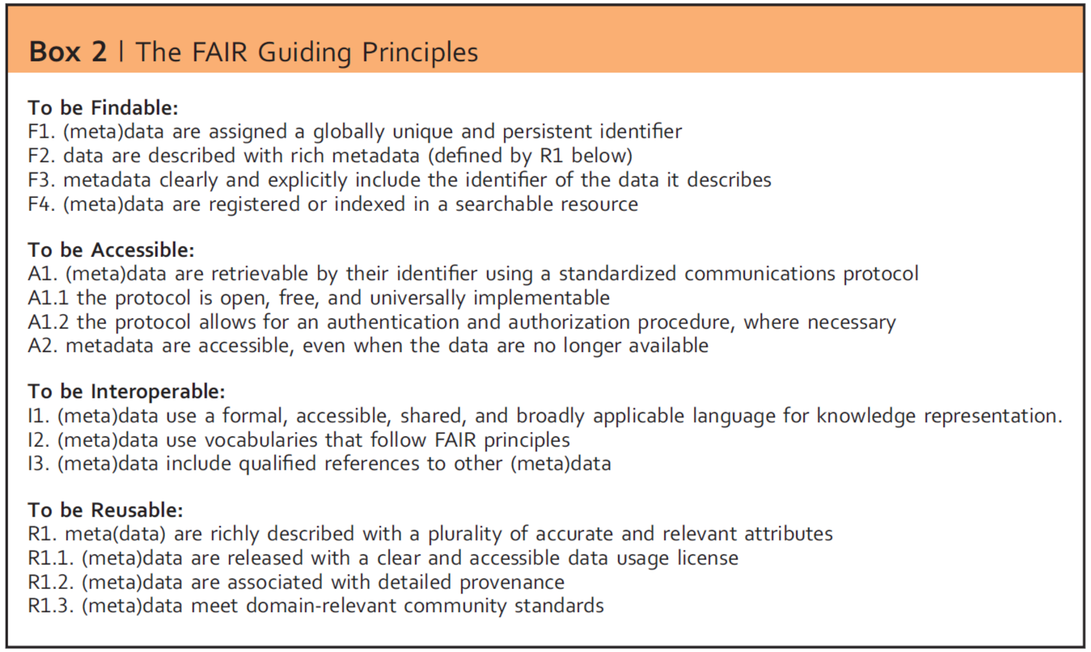
  - Not only for humans, also for machines

→ Obstacles to data discovery and re-use!

→ Loss of efficiency

THE FAIR PRINCIPLES – WHAT?

- Guidelines to improve the Findability, Accessibility, Interoperability, and Reuse of digital objects
- Scientific domain independent & high-level
- Principles are related but independent and separable
- Apply to different types of digital objects
- Distinction between data & metadata
- No standard in themselves!
- **FAIR data ≠ open data!**



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Source: Wilkinson et al. 2016

FINDABLE

F1. (Meta)data are assigned a **globally unique and persistent identifier**

F2. Data are described with rich **metadata**

F3. Metadata clearly and explicitly include the identifier of the data they describe

F4. (Meta)data are registered or **indexed in a searchable resource**

ACCESSIBLE

A1. (Meta)data are retrievable by their identifier using a standardised communications protocol

- A1.1 The protocol is open, free, and universally implementable
- A1.2 The protocol allows for an authentication and authorisation procedure, where necessary

A2. Metadata are accessible, even when the data are no longer available

INTEROPERABLE

I1. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (Meta)data use vocabularies that follow FAIR principles

I3. (Meta)data include qualified references to other (meta)data

REUSABLE

R1. (Meta)data are richly described with a plurality of accurate and relevant attributes (> to avoid this)

- R1.1. (Meta)data are released with a clear and accessible data usage **license**
- R1.2. (Meta)data are associated with detailed provenance
- R1.3. (Meta)data meet domain-relevant community standards

METADATA
SOURCE: DEUTZ ET AL. 2020

- Data about other digital objects
- 3 types of metadata:
  - Administrative
    - Management-level data: e.g. project/ resource owner, principal investigator, project collaborators, funder, project period, etc.
  - Descriptive
    - Data allowing discovery and identification: e.g. for example, authors, title, abstract, keywords, persistent identifier, related publications, etc.
  - Structural
    - Data on how dataset was created and is internally structured. For example: the unit of analysis, collection method, sampling procedure, sample size, categories, variables, etc.
- Example: Dublin Core metadata & DataCite

ORIGINAL DUBLIN CORE ELEMENTS

SOURCE: WIKIPEDIA

1. Contributor – "An entity responsible for making contributions to the resource".
2. Coverage – "The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant".
3. Creator – "An entity primarily responsible for making the resource".
4. Date – "A point or period of time associated with an event in the lifecycle of the resource".
5. Description – "An account of the resource".
6. Format – "The file format, physical medium, or dimensions of the resource".
7. Identifier – "An unambiguous reference to the resource within a given context".
8. Language – "A language of the resource".
9. Publisher – "An entity responsible for making the resource available".
10. Relation – "A related resource".
11. Rights – "Information about rights held in and over the resource".
12. Source – "A related resource from which the described resource is derived".
13. Subject – "The topic of the resource".
14. Title – "A name given to the resource".
15. Type – "The nature or genre of the resource".

INCREASING REUSABILITY
README FILE

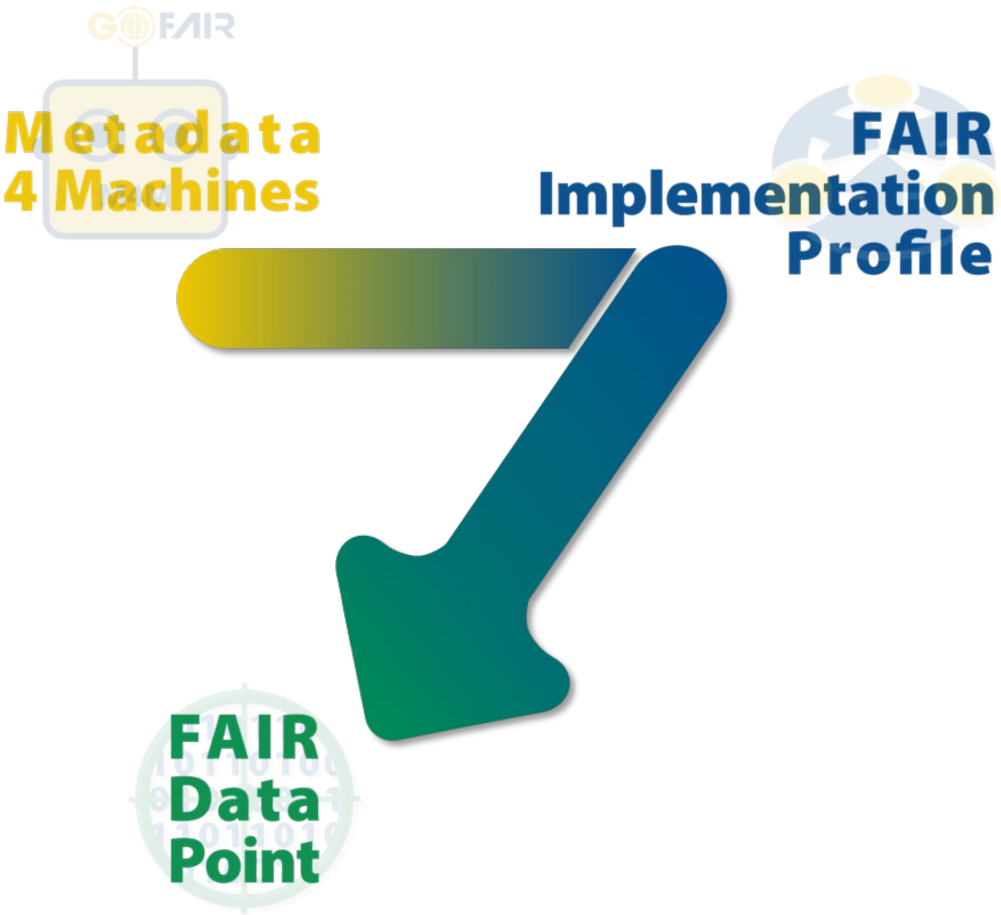
- Audience: end-user
- Some potential content:
  - Title of the project
  - Aim of the project & introduction
  - System / software requirements
  - Installation
  - Workflow
  - Origin of data
  - License
  - Troubleshooting & FAQ
  - Contact person (developer and maintainers)
  - Any other relevant information…

LICENSING

- Important for the re-use of data
- What are you (not) allowed to do with the data?
  - Legal point of view
- Commonly used licenses:
  - MIT
  - GNU GPLv3
  - Creative Commons (CC) >> not recommended for softwares!
- Need help? https://choosealicense.com/

THE FAIR PRINCIPLES – HOW?

- Originally, no guidance
- This led to practice variation
- Initiatives to improve the implementation of FAIR principles
- Implementation networks
- FAIR Evaluation Services:
  - https://fairsharing.github.io/FAIR-Evaluator-FrontEnd/#!/#%2F!



Source: Schultes, 2021

RESEARCH DATA MANAGEMENT

- Data Management Plan (DMP)
  - Compulsory for new research (depending on funder)
  - UT DMP: https://apps.utwente.nl/dmpstartpage/home
  - Asks how researchers will:
    - Perform data collection
    - Document data
    - Store data [during research]
    - Secure access to the data
    - Preserve data (archiving) [after research]
    - Make data (publicly) available
- Think in advance about research data lifecycle!

DMP

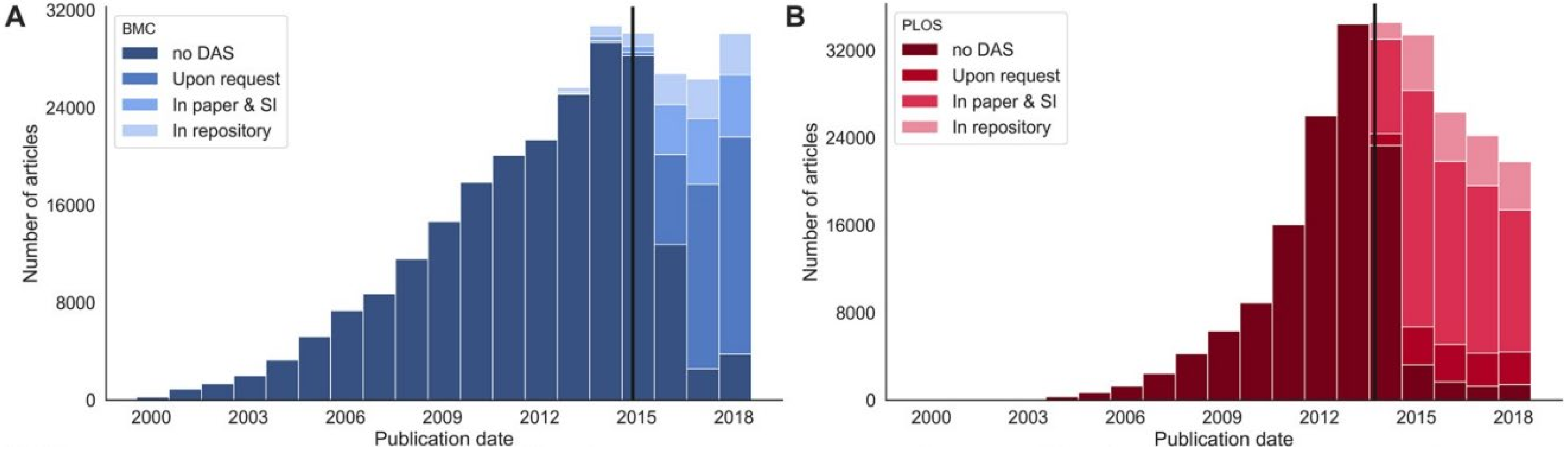
- Aims of DMP are to improve:
  - Adherence to legal and funding body requirements
  - Adherence to FAIR principles
  - The quality and security of data

- Software Management Plan (equivalent for softwares):
  - eScience centre guidelines: https://doi.org/10.5281/zenodo.7038280

DATA ARCHIVING (& SHARING)

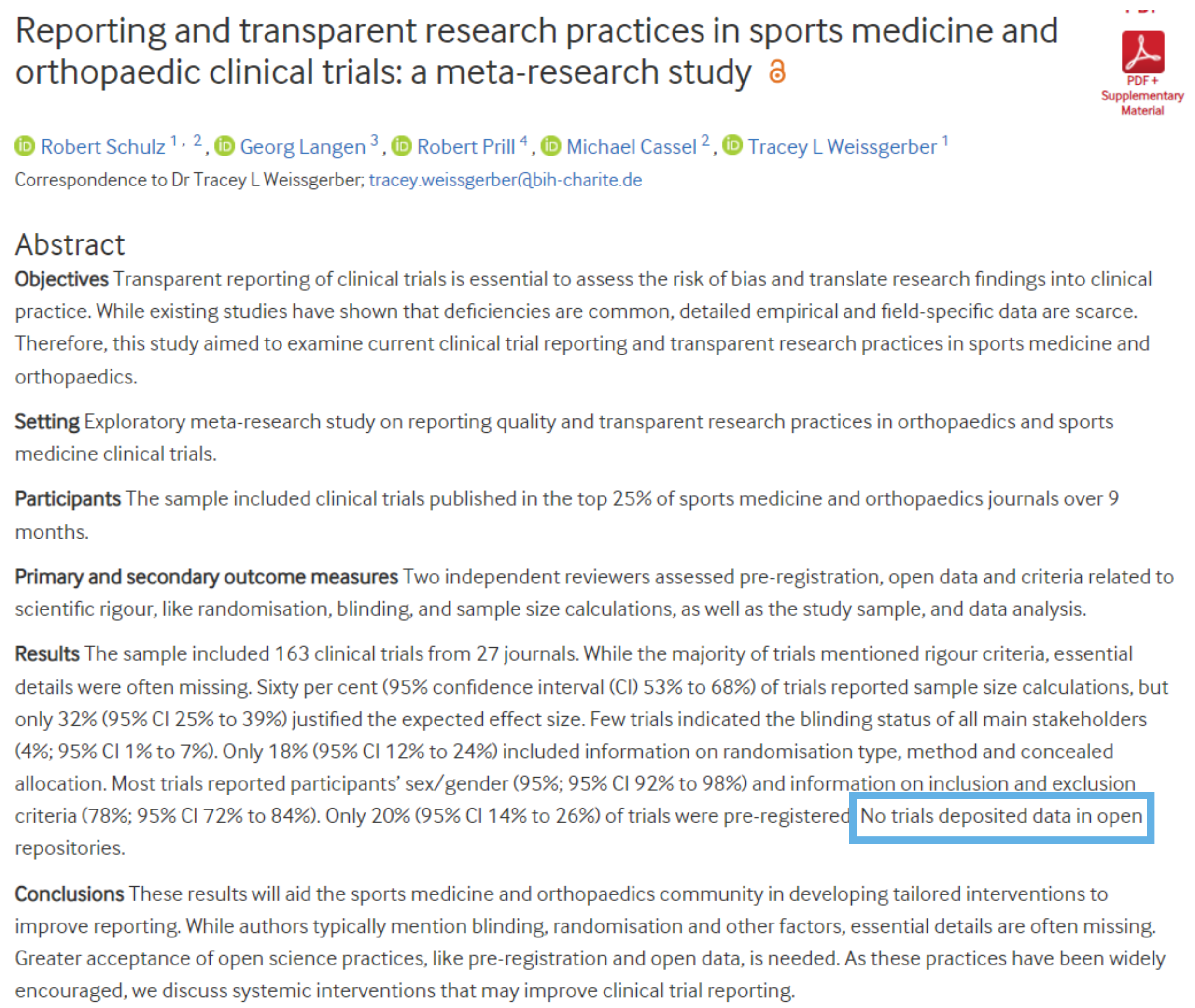
- Some existing repositories
  - 4TU.ResearchData
  - DANS-EASY
  - ZENODO
  - Certified repositories: https://www.coretrustseal.org/why-certification/certified-repositories/
- Research-field specific repositories
- Advantages:
  - Generate a DOI
  - Generate information required for citation
  - Allow Findability (and Accessibility)

USE OF REPOSITORIES



Source: Colavizza et al, 2020

SHARING DATA IN REPOSITORIES
SPORTS MEDICINE & ORTHOPAEDIC TRIALS



Reporting and transparent research practices in sports medicine and orthopaedic clinical trials: a meta-research study

Robert Schulz [1, 2], Georg Langen [3], Robert Prill [4], Michael Cassel [2], Tracey L Weissgerber [1]

Correspondence to Dr Tracey L Weissgerber; tracey.weissgerber@bih-charite.de

Abstract

Objectives Transparent reporting of clinical trials is essential to assess the risk of bias and translate research findings into clinical practice. While existing studies have shown that deficiencies are common, detailed empirical and field-specific data are scarce. Therefore, this study aimed to examine current clinical trial reporting and transparent research practices in sports medicine and orthopaedics.

Setting Exploratory meta-research study on reporting quality and transparent research practices in orthopaedics and sports medicine clinical trials.

Participants The sample included clinical trials published in the top 25% of sports medicine and orthopaedics journals over 9 months.

Primary and secondary outcome measures Two independent reviewers assessed pre-registration, open data and criteria related to scientific rigour, like randomisation, blinding, and sample size calculations, as well as the study sample, and data analysis.

Results The sample included 163 clinical trials from 27 journals. While the majority of trials mentioned rigour criteria, essential details were often missing. Sixty per cent (95% confidence interval (CI) 53% to 68%) of trials reported sample size calculations, but only 32% (95% CI 25% to 39%) justified the expected effect size. Few trials indicated the blinding status of all main stakeholders (4%; 95% CI 1% to 7%). Only 18% (95% CI 12% to 24%) included information on randomisation type, method and concealed allocation. Most trials reported participants' sex/gender (95%; 95% CI 92% to 98%) and information on inclusion and exclusion criteria (78%; 95% CI 72% to 84%). Only 20% (95% CI 14% to 26%) of trials were pre-registered. No trials deposited data in open repositories.

Conclusions These results will aid the sports medicine and orthopaedics community in developing tailored interventions to improve reporting. While authors typically mention blinding, randomisation and other factors, essential details are often missing. Greater acceptance of open science practices, like pre-registration and open data, is needed. As these practices have been widely encouraged, we discuss systemic interventions that may improve clinical trial reporting.

SHARING DATA IN REPOSITORIES SPANISH COVID-19 STUDIES

- 28 papers (out of 167 with linked data) shared data in a repository
- Only 7 out of 28 had a persistent identified (e.g. DOI)

→ This may hamper (long-term) accessibility

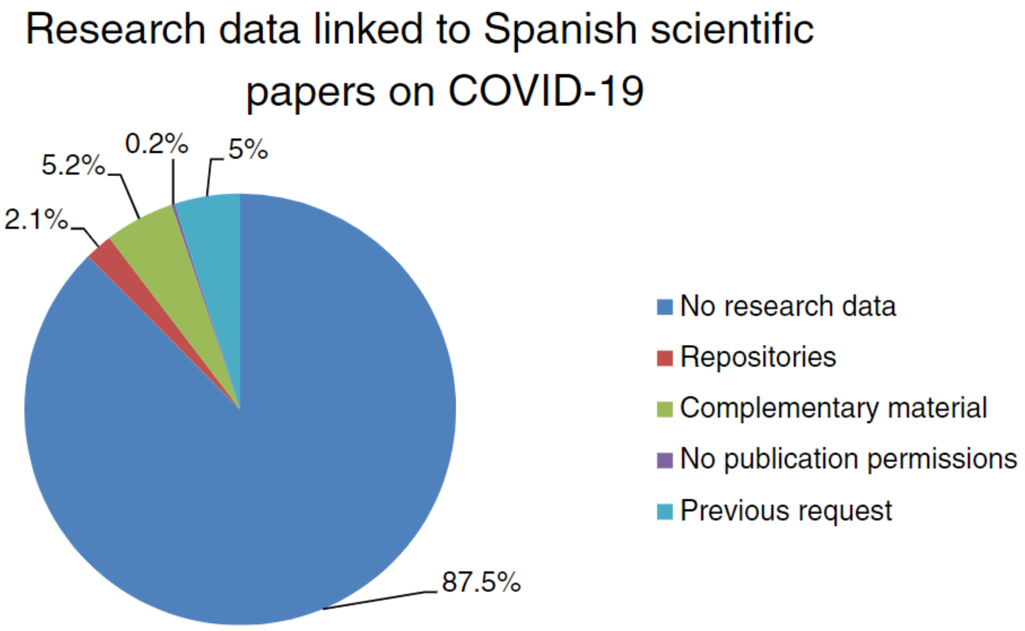


FIGURE 1 Research data linked to Spanish scientific papers on COVID-19

Source: Cerda-Cosme & Mendez, 2022

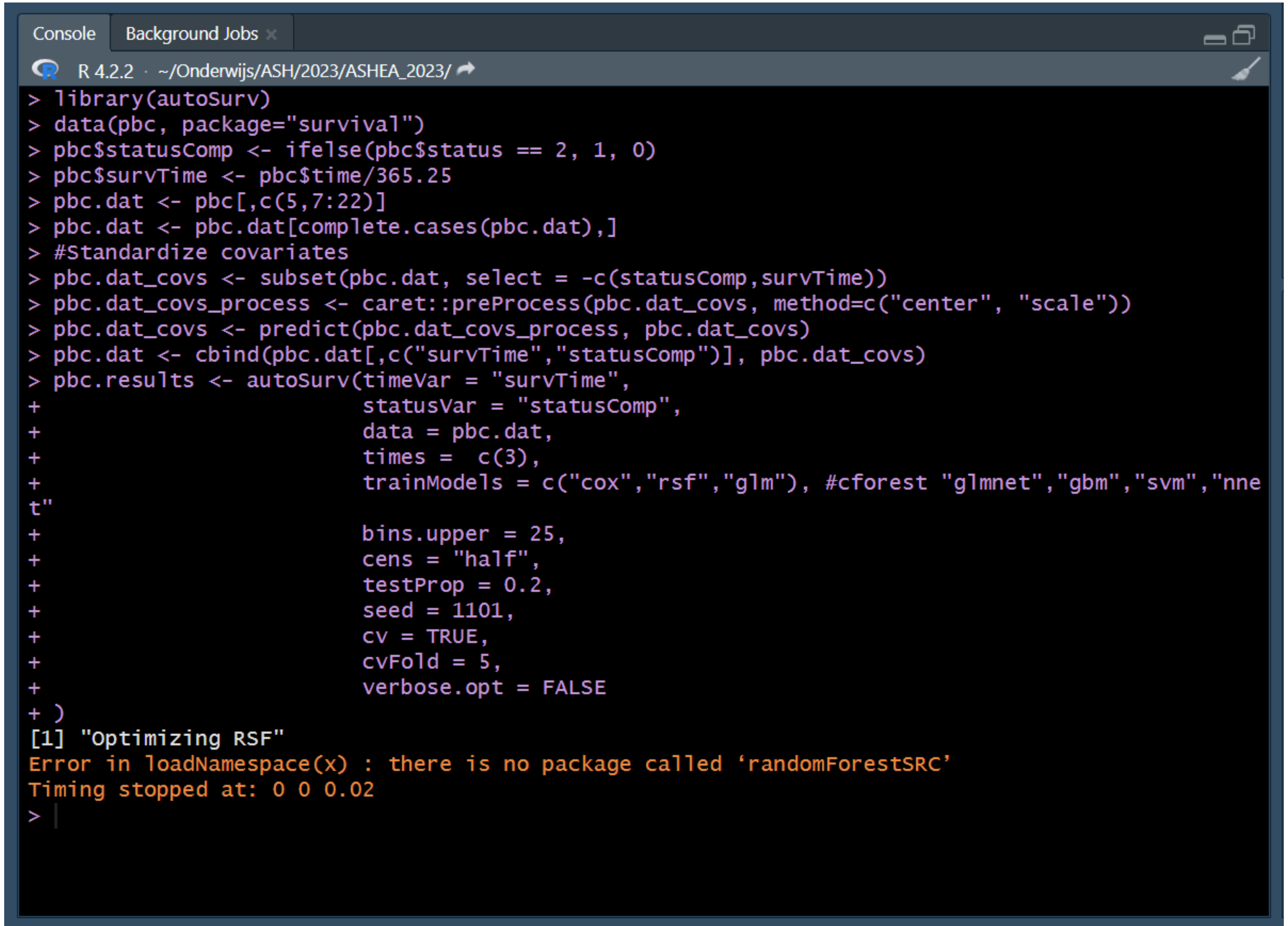
SHARING HEALTH ECONOMIC MODELS?

- Fragmented practice
  - As appendices
  - In repository
  - On github
  - No standard documentation and metadata templates
- Open-Source model clearinghouse
  - 16 available health economic models
  - Originally asked 248 authors to submit: 4 did it… (Emerson et al. 2019)
  - http://ghcearegistry.org/orchard/open-source-model-clearinghouse

FAIR @ UT

- Digital Competence Centre
  - FAIR Data Steward
- UT data archive: AREDA
- FAIR Data Fund (4TU-level)

THUS, HOW WOULD YOU DO IT NOW?



```
> library(autoSurv)
> data(pbc, package="survival")
> pbc$statusComp <- ifelse(pbc$status == 2, 1, 0)
> pbc$survTime <- pbc$time/365.25
> pbc.dat <- pbc[,c(5,7:22)]
> pbc.dat <- pbc.dat[complete.cases(pbc.dat),]
> #Standardize covariates
> pbc.dat_covs <- subset(pbc.dat, select = -c(statusComp,survTime))
> pbc.dat_covs_process <- caret::preProcess(pbc.dat_covs, method=c("center", "scale"))
> pbc.dat_covs <- predict(pbc.dat_covs_process, pbc.dat_covs)
> pbc.dat <- cbind(pbc.dat[,c("survTime","statusComp")], pbc.dat_covs)
> pbc.results <- autoSurv(timeVar = "survTime",
+                         statusVar = "statusComp",
+                         data = pbc.dat,
+                         times =  c(3),
+                         trainModels = c("cox","rsf","glm"), #cforest "glmnet","gbm","svm","nne
t"
+                         bins.upper = 25,
+                         cens = "half",
+                         testProp = 0.2,
+                         seed = 1101,
+                         cv = TRUE,
+                         cvFold = 5,
+                         verbose.opt = FALSE
+ )
[1] "Optimizing RSF"
Error in loadNamespace(x) : there is no package called ‘randomForestSRC’
Timing stopped at: 0 0 0.02
> 
```

DO-IT-YOURSELF

Some ways to apply today’s topic:

- Create a data management or software management plan for your health economic model
  - Does the content of these fall short for your model? Why (not)?
- Add metadata to (elements of) your health economic model
- Create a README file for your health economic model
  - What should someone know about your model in order to (re)use it properly?

ANY QUESTION?

RESOURCES

- Cerda-Cosme, R., & Méndez, E. (2022). Analysis of shared research data in Spanish scientific papers about COVID-19: A first approach. Journal of the Association for Information Science and Technology, 1– 13. https://doi-org.ezproxy2.utwente.nl/10.1002/asi.24716
- D.B. Deutz, M.C.H. Buss, J. S. Hansen, K. K. Hansen, K.G. Kjelmann, A.V. Larsen, E. Vlachos, K.F. Holmstrand (2020). How to FAIR: a Danish website to guide researchers on making research data more FAIR https://doi.org/10.5281/zenodo.3712065; link to website: https://www.howtofair.dk/, accessed on 18-01-2023.
- Emerson J, Bacon R, Kent A, Neumann PJ, Cohen JT. Publication of Decision Model Source Code: Attitudes of Health Economics Authors. PharmacoEconomics. 2019;37(11):1409-10. doi:10.1007/s40273-019-00796-3.
- GO FAIR Initiative website: https://www.go-fair.org/, accessed on 18-01-2023.
- Martinez-Ortiz, Carlos, Martinez Lavanchy, Paula, Sesink, Laurents, Olivier, Brett G., Meakin, James, de Jong, Maaike, & Cruz, Maria. (2023). Practical guide to Software Management Plans (1.1). Zenodo. https://doi.org/10.5281/zenodo.7589725

RESOURCES

- Schultes, Erik. (2021). Official GO FAIR Foundation icons for the Three-Point FAIRification Framework (Version 1). Zenodo. https://doi.org/10.5281/zenodo.4678333.
- Schulz R, Langen G, Prill R, et alReporting and transparent research practices in sports medicine and orthopaedic clinical trials: a meta-research studyBMJ Open 2022;12:e059347. doi: 10.1136/bmjopen-2021-059347
- Wikipedia, “Dublin Core”: https://en.wikipedia.org/wiki/Dublin_Core#Dublin_Core_Metadata_Element_Set, accessed on 18-01-2023
- Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18.