Data sharing practices among authors of biomedical publications

Gabelica, Mirko

Doctoral thesis / Disertacija

2021

Degree Grantor / Ustanova koja je dodijelila akademski / stručni stupanj: **University of Split, School of Medicine / Sveučilište u Splitu, Medicinski fakultet**

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:171:593273

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-03-29**

SVEUČILIŠTE U SPLITU
MEDICINSKI FAKULTET
UNIVERSITAS STUDIORUM SPALATENSIS
FACULTAS MEDICA

Repository / Repozitorij:

MEFST Repository

UNIVERSITY OF SPLIT

dabar
DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

UNIVERSITY OF SPLIT

SCHOOL OF MEDICINE

Mirko Gabelica

Data sharing practices among authors of biomedical publications

DOCTORAL THESIS

Mentor: Prof. Livia Puljak, MD, PhD

Split, Croatia, December 2021

This doctoral thesis was made at the University of Split, School of Medicine, as a final part of the PhD program TRIBE (Translational Research in Biomedicine).

Mentor: Professor Livia Puljak, MD, PhD

Acknowledgements

I am thankful to all the people that helped me achieve this, especially:
To Maja, Matea, Ivana and Hana, you are the spring of my strength and persistence, my life
To Slave and Iće, thank you for endless support, love and money
To Ana and Zvone, I love you; keep shining you crazy diamonds
To Mare, thank you for the childhood, and Rajka, thank you for coming home
To Jasna, thank you for watching over my wife and kids, great ideas and cooking
To Damir Sapunar and Matko Marušić, I should've listened to you more
To Ana Marušić, thank you for your faith in me
To Stipe, Marko, Ivan, Andro, Toni and Filip, I hope we never grow up
Words cannot express the gratitude towards my mentor Livia, for guiding me and frequently pushing me to work harder; I hope we stay friends for life

Ive, Riba and Željko, I hope you can see us and see this; I miss you with all my heart.
Dear grandparents, I wish I have had known you better

Table of Contents

1 INTRODUCTION 1

1.1 **OPEN DATA** 1
1.2 **DATA SHARING** 3
1.3 **BENEFITS AND CONCERNS REGARDING OPEN DATA AND DATA SHARING** 4
1.4 **CLINICAL TRIALS** 6
1.4.1 Definition of a clinical trial 6
1.4.2 History of clinical trials 7
1.4.3 Importance of clinical trials in the hierarchy of evidence 9
1.5 SHARING DATA FROM CLINICAL TRIALS 12
1.5.1 Importance of clinical trial data sharing 12
1.5.2 Existing repositories for hosting clinical trial data 14
1.5.3 Results of previous studies that tried to access data from clinical trials 14
1.6 **RESEARCHERS' WILLINGNESS TO SHARE SCIENTIFIC DATA** 16
1.7 **DATA AVAILABILITY STATEMENT** 19
1.8 **RESEARCH PROBLEM** 21

2 AIMS 22

2.1 **AIMS OF THE FIRST STUDY** 22
2.2 **AIMS OF THE SECOND STUDY** 22
2.3 **HYPOTHESES OF THE FIRST STUDY** 23
2.4 **HYPOTHESES OF THE SECOND STUDY** 23

3 METHODS 24

3.1 **FIRST STUDY** 24
3.1.1 Study design 24
3.1.2 Ethics 24
3.1.3 Unit of analysis 24
3.1.4 Sample 24
3.1.5 Search 24
3.1.6 Outcomes 25
3.1.7 Data extraction 25
3.1.8 Author survey 25
3.1.9 Data analysis 26
3.2 **SECOND STUDY** 26
3.2.1 Study design 26
3.2.2 Ethics 27
3.2.3 Unit of analysis 27
3.2.4 Sample 27
3.2.5 Outcomes 27

3.2.6 Data Extraction 28
3.2.7 Author Survey 29
3.2.8 Data Analysis 30

4 RESULTS 31

4.1 **FIRST STUDY RESULTS** 31
4.1.1 Results 31
4.1.2 Data sharing in manuscripts describing RCTs 32
4.1.3 Requesting data from trial authors 32
4.1.4 Response from authors who had a data-sharing statement 33
4.1.5 Raw data that was shared 33
4.2 **SECOND STUDY RESULTS** 34
4.2.1 Data extraction analysis 34
4.2.2 Contacting the authors 36
4.2.3 Response analysis 37
4.2.4 Randomized controlled trials among the articles that shared the data 40
4.2.5 Usable datasets 41

5 DISCUSSION 41

5.1 **FIRST STUDY** 41
5.2 **SECOND STUDY** 44

6 ABSTRACT 50

7 SAŽETAK 52

8 REFERENCES 54

9 APPENDICES 66

9.1 APPENDIX 1. SCANNED APPROVAL OF THE STUDY PROTOCOL BY THE ETHICS COMMITTEE OF THE UNIVERSITY OF SPLIT SCHOOL OF MEDICINE, FOR STUDY 1 66
9.2 APPENDIX 2. PERSONALISED EMAILS TO EACH POTENTIAL PARTICIPANT IN STUDY 1 67
9.3 APPENDIX 3. PERSONALISED EMAILS TO EACH POTENTIAL PARTICIPANT IN THE STUDY 2 68
9.4 APPENDIX 4. NON-DISCLOSURE AGREEMENT 69
9.5 APPENDIX 5. SCANNED APPROVAL OF THE STUDY PROTOCOL BY THE ETHICS COMMITTEE OF THE UNIVERSITY OF SPLIT SCHOOL OF MEDICINE, FOR STUDY 2 70

10 CURRICULUM VITAE 71

List of abbreviations

AIDS - Acquired immunodeficiency syndrome.

B2Share - Repository that is a part of European collaborative data infrastructure.

BBC - British Broadcasting Corporation, radio and television broadcasting company.

BMC - BioMed Central, for-profit scientific open access publisher.

BMJ - British Medical Journal, a weekly peer-reviewed medical journal published by the British Medical Association.

CHORUS - a platform for monitoring open access to content and datasets reporting on funded research.

CI - Confidence Interval

CSDR - clinicalstudydatarequest.com, a consortium of clinical study sponsors

DAS - Data Availability Statement, a statement about where to access data supporting the results reported in a published article

DRUM - The Data Repository for the University of Minnesota, publicly available collection of digital research data generated by University of Minnesota researchers, students, and staff

DOI - digital object identifier, electronic fingerprint for digital objects

DTA - Data Transfer Agreement, a contract between the providing and recipient institutions that manages legal commitments and limitations in compliance with laws and regulations of such activity.

EASY/DANS - EASY – digital repository, DANS – Data Archiving and Network Services, Dutch national centre of expertise and repository for research data.

EBM - Evidence-based medicine, using up-to-date evidence in the healthcare decision-making process

EMA - European Medicines Agency, European institution in charge of the evaluation and supervision of medicinal products.

Enago - ENAGO Academy.

FAIR - acronym describing scientific data as Findable, Accessible, Interoperable and Reusable.

FAIR/O - database is compliant with FAIR principles and also carries an open licence.

FDA - The United States Food and Drug Administration is a federal authority ensuring efficacy, safety, and security of human and veterinary drugs, biological products, etc.

GDPR - General Data Protection Rule, 2016/679, EU legal regulation on data protection.

GISAID - Global Initiative On Sharing Avian Influenza Data, a global science initiative and primary source established in 2008 that provides open access to genomic data of influenza viruses.

GRID - Gay-Related Immunodeficiency.

GSK - GlaxoSmithKline, a pharmaceutical industry company.

HIV - Human Immunodeficiency Virus.

ICF - Individual Consent Form, participant affirmation form on which they acknowledge understanding of the research procedure and consent to it.

ICPSR - Inter-university Consortium for Political and Social Research, maintains and provides access to a vast archive of social science data for research.

ICSU - International Council of Scientific Unions, a global non-governmental organization dedicated to international cooperation in advancing science.

IPD - Individual Participant Data, all raw data gathered from one participant during the research.

IOM - Institute of Medicine, American non-profit, non-governmental organization, now called National Academy of Medicine.

JČ - Jakica Čavar, author.

JCR - Journal Citation Reports, academic journals impact factor report.

LSHTM - London School of Hygiene & Tropical Medicine.

MB - Megabyte.

MEDLINE - a bibliographic database of life sciences and biomedical information.

MG - Mirko Gabelica, author.

N - Number.

GenBank – United States National Institute of Health genetic sequence databank.

NDA - Non-Disclosure Agreement.

NY - New York, USA.

OSF - Open Science Framework, an open-source software project that facilitates open collaboration in science research.

PLoS - The Public Library of Science, open-access non-profit science, technology, and medicine publisher.

PubMed - acronym, Public and Medline - search engine accessing MEDLINE database of references and abstracts on life sciences and biomedical topics.

RCT - Randomised Controlled Trial.

RCSB PDB - Research Collaboratory for Structural Bioinformatics Protein Data Bank.

SARS-CoV2 - Severe Acute Respiratory Syndrome Corona Virus 2

SND - Swedish National Data Service.

SOAR - Supporting Open Access for Researchers initiative.

USA - United States of America.

Vivli - a global clinical research data sharing platform.

YODA - Yale University Open Data Access.

WA, USA - Washington state in the United States of America.

WMA - World Medical Association.

1 INTRODUCTION

1.1 Open data

1.1.1. Definition of open data

European Commission communication to European Parliament in 2014 defines the term "open data" as [quote]: "*a subset of data, namely to data made freely available for re-use to everyone for both commercial and non-commercial purposes*" (1).

According to Krzysztof Izdebski, open data is defined through three key factors: (i) open access means everyone can obtain data without being discriminated against data access for any reason, (ii) database format means that data is accessible in digital format compatible for reading with most common software for the determined file type, and (iii) freedom of re-use means that anyone can use fully or partially, re-use, build on or redistribute data without bureaucratic obstacles (2).

1.1.2. History of open data movement

The history of modern scientific journals and peer review begins in 1665 in London, when the secretary of Royal Society, the oldest independent scientific academy globally, Henry Oldenburg established the first and still publishing scientific journal, *Philosophical Transactions of Royal Society*. His idea was to open his correspondence to other scientists. In order to be published, the scientist had to accompany the correspondence with appropriate evidence, i.e. he demanded data be sent along with the article. Scientists were permitted to closely inspect and criticise the alleged logical correlation between data and proposed evidence. The intention was to scrutinise published data or replicate the experiment and observations and ultimately re-use the data for other research. The process proved to be the most powerful peer review form, even more powerful than pre-publication peer-review. Openness to refutation is the core element in developing scientific knowledge, that is, by definition, tentative and indefinite (3).

International Council of Scientific Unions (today International Science Council) (ICSU) was founded in 1931, assembling world academies to strengthen science for society's benefit. In 1955 the ICSU recommended that data should be made available in machine-readable form. In preparation for International Geophysical Year 1957-1958, the ICSU established several World Data Centres (today World Data System) to minimise the risk of data loss and to maximise data accessibility as well to initiate open access to scientific data, and to facilitate the adoption of standards for data exchange (4).

In 1964 MEDLINE was created as the first large-scale computerised search service, managed by the National Institute of Health (USA) and the National Library of Medicine with bibliographical citations from journals in the biomedical area, accessible only through institutions. In 1996 PubMed was launched, providing free public access to MEDLINE content (5).

The first mention of the term open data was in 1995 in a document regarding environmental and geophysical data disclosure among scientists. This document defined that open data should be freely available for re-use and uploaded online (6).

Although the formalised definition of open data is relatively recent, the idea behind open data is not new. The idea of sharing knowledge, i.e., that data from scientific research should be accessible to all was popularised much earlier by Robert King Merton, founder of the sociology of science. In the 1940s, he posited that knowledge was the common good and that every researcher should contribute to the collective knowledge pool, thus waiving intellectual property (7).

To ensure standard in opening the data, in 2016, a consortium of scientists and organisations published "FAIR Guiding Principles for scientific data management and stewardship" using FAIR as an acronym meaning – Findability, Accessibility, Interoperability and Reusability. Findability means the data should be undemanding to find by both machines and humans. Accessibility suggests data should be available possibly in a non-restrictive manner, enabling processes such as authentication of the third party and authorisation from the provider, or sometimes, not even that. Interoperability defines data integrations with its metadata and association with other applications and analysis platforms. Reusability is the optimal achievement of FAIR principles, meaning that existing data and metadata are strictly defined and described to be replicated or built upon for a different purpose (8). The evolution of FAIR

principles is today's FAIR/O abbreviation indicating that the data has been prepared accordingly and contains an explicit open license. The importance is emphasised in adopting mentioned criteria by the Association of European Research Libraries in 2018 (9).

1.2 **Data sharing**

1.2.1. Definition of data sharing

Data sharing in the academic sense makes research data available to third parties (10). Data sharing is not synonymous with open data. According to the Open Data Institute, [quote] "*Data sharing is providing restricted data to restricted organisations or individuals*", while "*Open data is providing unrestricted data to everyone*" (11).

1.2.2. History of data sharing

Data were certainly shared in history in smaller scientific communities. However, data sharing as an example of a worldwide standard first took place in Vienna in 1873 when an international standard for weather observation data was adopted. An incentive to advance the field of meteorology was described in an essay Suggestions on a Uniform System of Meteorological Observations, by Buys Ballot, [quote]:" *It is elementary to have a worldwide network of meteorological observations, free exchange of observations between nations and international agreement on standardised observation methods and units in order to be able to compare these observations*" (12). Meteorological data from around the globe has been shared daily ever since. Other geophysical sciences now use meteorological data to complement/enrich their measurements and observations. Today scientists also regularly share equipment and samples, which is even more challenging than sharing raw data. However, it is acknowledged that sharing of biomedical data raises specific concerns such as patient confidentiality, governed by regulations such as General Data Protection Regulation (GDPR), and ethical issues (13).

One of the examples of acceptable data sharing practices are diseases that pose a threat on a global scale, such as pandemic SARS-CoV2 infection, HIV, Ebola, Zika virus disease,

malaria and tuberculosis. In public health emergencies, data sharing enables researchers worldwide to analyse data, improve conclusions, and strengthen the facts to help find the optimal solution promptly and in a cost-effective way (14). However, data sharing is essential even outside public health emergencies. Infrastructure is needed to ensure that data is available to researchers to prevent and treat present and emerging threats. The scientific community is well aware of the economic, political, and scientific implications of data withholding.

An example of such a platform for global data sharing has existed since May 2008, and it is called the Global Initiative on Sharing All Influenza Data (GISAID). It was made to meet the scientific needs regarding scattered comprehensive data on bird flu in 2006; the platform was designed to overcome researcher concerns regarding intellectual property on data, fear that someone else may not credit them for contributing to data, or partially publish findings without their knowledge. In addition, the database contains free public access to genetic sequences of avian influenza (15).

Nowadays, funding agencies and peer-reviewed journals may require researchers to disclose and share supplemental data such as raw data, source code or statistical methods essential to understand or replicate published research. However, most scientific data is not subjected to data sharing, mainly because there are no obligatory stipulations to ensure data sharing, placing the process at the researchers' discretion. Sometimes governments or institutions place an embargo on data sharing to protect the national interest, national security, patient or victim confidentiality, or simply to protect the institution from using data for political purposes (16).

1.3 **Benefits and concerns regarding open data and data sharing**

The benefits of open data and data sharing include accelerated scientific progress and increased research relevance and visibility. Furthermore, such openness and sharing create opportunities for additional publications through collaboration, increase citation rate that is important for academic progress and future research funding, and strengthen the reputation of educational institutions. In clinical terms, data sharing is beneficial because it decreases time transferring knowledge from the laboratory to clinical practise (17).

There are numerous projects and initiatives in the scientific community that could not be managed without data sharing. One of the flagship data sharing projects in biomedicine is the Human Genome Project, which took fourteen years to complete. It included twenty institutions, and numerous experts in chemistry, engineering, informatics, physics and biology. The complete human genome data is hosted at the University of California, Santa Cruz GenomeBrowser, Ensembl, and GenBank websites, and it is open to everyone without restrictions (18).

According to Wojick, the lack of budget is the biggest issue in data sharing processes regarding data preparation and curation; prior disclosure, and archiving require substantial funding (19). However, a group of biostatisticians from the University of Riyadh found no connection between data quality and funding. The lack of training, consulting with experts, and using electronic data capturing methods was responsible for initially flawed datasets (20).

When sharing data, researchers may fear that other researchers can find a mistake in the data; however, the idealistic reasoning is that by placing the data online, researchers will be compelled to closely examine and prepare the data before publishing (21). Furthermore, suppose someone spots a problem with the original data or their analyses. In that case, this will contribute to the self-correcting mechanisms in science and will help the scientific society rectify the initial issues eventually. However, not every researcher will share this idealistic vision that open data is ultimately beneficial for society, as multiple problems are considered, such as credit issues, responsibility, and consequences (22).

Regarding the credit questions, the potential negative backside of data sharing may include misappropriation of data and work. Some scientists are being daunted by the idea that someone else can receive a monetary prize or scientific acclamation using parts of their data, someone else uncovering the pieces of the puzzle that have missed the eye of the first researcher. In the context of science, it does not matter who found the answer or solution, but it matters from a competitive scientific viewpoint (23). Wallis et al. proposed that data exchange among scientists should be treated as a good scientific practice, and data should not be treated as a commodity because the data itself is inseparable from science (24,25).

Another contentious issue in this context is responsibility and consequences if re-analysis of open/shared data shows problems with the original data/results. When someone re-analyses open or shared data, some of the original scientific conclusions based on that data may be

discarded as misleading or false. The question in those cases is whether a journal that has published the initial results should intervene and retract the troublesome paper. Retractions may have serious personal consequences for researchers, such as loss of tenure, funding withdrawal, possible scientific prize withdrawal, and finally, legal actions from the third parties, along with dishonouring the institution where the research took place. Mistakes in data analyses or making erroneous conclusions can result from honest mistakes, or if one or more authors have deliberately corrupted the original data or purposefully made misleading conclusions (26–29). However, a possibility that someone can discover unintentional mistakes in the original data or analyses may hinder the idea of data sharing/open data among researchers.

1.4 **Clinical trials**

1.4.1 Definition of a clinical trial

A clinical trial is defined as a [quote]: "a *research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioural outcomes*" (30).

There are four phases of biomedical clinical trials. Phase I examines new potential medicines for the first time among a small group of healthy people, usually twenty to eighty participants, to establish a safe dose range and evaluate side effects. Phase II examines the efficacy of the proposed medicine on a smaller scale, often versus placebo group; it is commonly divided into two stages, stage IIa evaluates how much drug should be given, it is a dose assessment phase, stage IIb determines the efficacy of the medicine in prescribed dose, and establishes therapeutic range, it is the effect assessment phase. Phase III determines the safety and efficacy of the medicine; in this phase, a large group of people, hundreds or thousands of participants, are tested. Phase III needs to include outcomes on efficacy and safety and comparison to approved treatments, if applicable. Finally, phase IV initiates after the medicine receives marketing authorisation. This phase continues throughout the medicines' active lifetime. Phase IV addresses optimal use and continuously evaluates the therapy's risk/benefit ratio (31).

1.4.2 History of clinical trials

Some consider that the Book of Daniel, from the Bible's Old Testament, contains the first description of an intervention resembling a clinical trial. The King Nebuchadnezzar, to keep his servants healthy, proscribed daily rations of meat and wine. However, Daniel asked the king if some servants could preserve their vegetarian diet, consisting primarily of beans and legumes. The king allowed it for ten days and ordered Melzar to watch over the vegetarian and meat-eating groups and file a report after ten days. The vegetarian group appeared to be healthier, according to Melzar's report (32).

In the 11th century, one of the earliest interventional studies with a control group was recorded in China. The Atlas of Materia Medica, written by Ben Cao Tu Jing and edited by Song Su, documented a trial of ginseng. Two athletes were asked to run along. One consumed the ginseng while the other ran without ginseng. After running for approximately 2000 meters, the athlete who did not receive ginseng developed severe shortness of breath, while the one who received ginseng did not. While this "trial" included only two individuals, and we do not know much about their characteristics and comparability, it is essential to emphasise that the use of a control group is the critical aspect of modern randomised controlled trials (RCT) (33).

In 1537, the French military barber-surgeon with a particular interest in wound healing, Ambroise Pare, wrote to his Captain [quote]: "*Je le pansai, Dieu le guérit"*, meaning: "*I bandaged him, and God healed him".* Having run out of elder oil for bandages, he devised a mixture of egg white, turpentine and rose oil. Injured soldiers treated with the mixture suffered no agony and recovered faster than the other because of the turpentine's antiseptic property (34).

In 1747, James Lind conducted the famous scurvy trial, hailed by many as the "first clinical trial in history". In that era, long sea journeys proved to be more perilous than the enemies encountered at sea. Various descriptions of the scurvy disease were noted. Most common were lack of strength, periodontal bleeding, loose teeth and bruising. Sailors consuming ship rats were protected because rats can synthesise ascorbic acid. In his trial, James Lind took 12 soldiers who were in the same stage of the disease and separated them into six groups of two, each of whom was assigned one of the following six interventions that they were supposed to take for 14 days: 1.1 L of cider, twenty-five mL of vitriol (sulfuric acid), thirteen mL of

vinegar three times a day before meals, two hundred eighty-four mL of seawater, two oranges and lemon, and paste made up of garlic, mustard seed, gum myrrh and dried radish root. The group that was supposed to take two oranges and one lemon per day continued taking this treatment for six days only because the supply was exhausted. At the end of the experiment, Lind concluded the following [quote]:

"*The most sudden and visible good effects were perceived from the use of oranges and lemons; one of those who had taken them being at the end of six days fit for duty … The other was the best recovered of any in his condition, and being now deemed pretty well, was appointed nurse to the rest of the sick"* (35).

In 1786, in Bath, England, physician Caleb Parry conducted a study to assess whether locally grown rhubarb, commonly used as a laxative, was as effective as the more expensive Turkish variety. During the trial, he switched the type of rhubarb given to each patient at different times. Parry then compared an individual patient's symptoms while consuming different types of rhubarb. He concluded that the Turkish version was not superior to locally grown rhubarb. This study is considered the first example of a crossover trial (a study where two or more test groups receive placebo and the medicament at a different time) (36).

In 1836, in Buffalo, NY, USA, Austin Flint described a trial on patients with rheumatic arthritis. They were all given some ointments for the wrists and a tincture of quassia. He reported that the beneficial effects of the treatment were solely attributed to patients' beliefs. This is considered the first described use of a placebo (37).

In 1905, the beriberi outbreak was recorded at the Kuala Lumpur Lunatic Asylum. Dr William Fletcher assigned a number to each patient, and then gave the groups different treatments. Patients with even numbers were given unpolished brown rice. Patients with odd numbers were given polished white rice. At the end of the experiment, the patients who ate white rice developed beriberi, and 15% of the patients died, while none of the patients given brown rice developed beriberi or died. His experiment mimicked multiple features of a modern randomised trial, including quasi-randomisation (38).

Randomisation is a cornerstone in modern clinical trial design. Sir Austin Bradford Hill, an English epidemiologist and statistician, conducted the first randomised controlled trial in 1948. The trial was designed to treat tuberculosis; he used a table of random numbers to

decide whether a patient should be treated with the antibiotic streptomycin plus bed rest or bed rest alone. Patients were not informed that they were participating in a trial. The researchers were not informed which participant belonged to each group; allocation details were hidden in sealed envelopes. A method used to hinder selection bias by hiding the allocation sequence from those assigning participants to intervention is called allocation concealment. Assuring neither researchers nor participants know which treatment they are receiving is called blinding. Allocation concealment and blinding are now standard characteristics of randomised controlled trials. A randomised controlled trial is nowadays the "gold standard" for clinical trial design as it is considered a fair test (39).

1.4.3 Importance of clinical trials in the hierarchy of evidence

In a constant scientific pursuit for quality evidence and minimising the risk of bias in biomedical studies, there appeared a need to stratify the strength of evidence. This was particularly emphasised by the evidence-based medicine (EBM) movement. As its name suggests, EBM is a concept that encourages an effort to produce evidence and use that evidence to make clinical decisions. The foundation of EBM is the ranking system of categorising evidence, known as the levels of evidence. Physicians are encouraged to incorporate the highest level of evidence in clinical decision making (40).

The pyramid that depicts evidence levels helps put each study type results into perspective based on specific study design strengths and weaknesses. At the base of the pyramid is an expert opinion as the least credible form of medical evidence. Moving up the pyramid, each study design is treated with more scrutiny and rigour, providing a greater level of confidence in results, where there is a lesser chance of statistical error, and confounding factors and bias influencing the results are minimised. Finally, RCTs are at the top of the pyramid as the highest evidence level among primary studies. At the very top, there are systematic reviews of RCTs as the pinnacle of medical research evidence (41).

Unlike RCTs, evidence that is ranked lower in the evidence pyramid has inherent issues hindering its objectivity. Expert opinion, ranked low in the pyramid of evidence, represents a scientific view or comment from one or more experts based on an appraisal of scientific evidence or another expert opinion. When viewed on its own, expert opinion can be heavily

influenced by beliefs, opinions and politics, thus providing a low level of confidence in the decision-making process (41).

There are observational studies in the middle of the pyramid, such as case reports, case series, case-control and cohort studies. Observational studies are vulnerable to various biases and structural limitations, but they are still valuable for collecting evidence, mainly when RCTs are not possible or feasible (42).

Case reports and case series describe one or more patients of interest regarding pathophysiological or operational aspects of a disease, treatment or diagnostic procedures. Case reports and case series represent basic types of clinical study designs, in which researchers observe and describe the experience of a small group of participants. The primary difference between case reports/series and the single-subject experiment is that the researcher does not intervene in a case report/series but simply documents occurrences during the usual clinical practice. A case series is a cluster of clinically equal or similar case reports in which the researcher describes several cases and their relation to one another. However, no causal deductions should be made from the case series regarding the efficacy of the investigated treatment. A case series includes patients with a specific outcome and a specific exposure or includes patients with a specific outcome and patients irrespective of exposure. Case series are often valuable in the early identification of clinical problems (43). For example, case reports were essential for the recognition of congenital rubella syndrome. Observation of a series of infants born with congenital cataracts and additional cardiac abnormalities in Australia in 1941 inspired Sir Norman Gregg to hypothesise a causal link between an epidemic rubella infection that had happened six to nine months before the children were born and the following deformities. It is now known that rubella infection during pregnancy may cause severe embryopathy (44).

A report of a series of five cases of *Pneumocystis carinii* pneumonia that occurred in young, previously healthy, homosexual men in Los Angeles hospitals from 1980–81 (45), and a month later case series report which described twenty-six homosexual men who developed Kaposi sarcoma from 1978-81, raised concern that an unknown disease led to these disorders (46). These case series were very curious because the condition almost exclusively developed among elderly Jewish/Mediterranean men and immunosuppressed. The disease was linked to the homosexual lifestyle and was stigmatised as GRID (Gay-Related Immunodeficiency). Not long after, a similar disease pattern appeared in intravenous addicts and haemophiliacs who

needed blood transfusions and Haitians. Consequently, another stigma appeared, as the "4H club", Homosexual, Heroin, Haemophiliac and Haitian. In 1983, American and French scientists independently discovered that the virus caused the disease, now known as Acquired Immunodeficiency Syndrome (AIDS) (47).

Case-control studies evolved during the 19th and 20th centuries, combining medical concepts (caseness, disease aetiology, and a focus on the individual), together with medical procedures (patient history, a grouping of cases into series, and differentiation of the diseased and the healthy) (48). The first major case-control study was the one that Janet Lane-Claypon published in 1926, which yielded the first epidemiologic evidence that low fertility increases the risk of breast cancer. Her report titled *A Further Report on Cancer of the Breast, With Special Reference to Its Associated Antecedent Conditions,* described a study of 500 hospitalised cases and 500 controls (49).

Cohort studies are placed just below RCTs in the evidence pyramid. The cohort was a standard Roman tactical military unit composed approximately of 480 soldiers (50). The cohort study is an observational study that monitors a large group of participants over an extended period to see how their exposures affect their outcomes; it is also called longitudinal or epidemiological study. Cohort studies enlist and monitor participants who share a common feature, such as a particular occupation or demographic similarity. This type of research is often used to observe the outcome of questioned risk factors that cannot be measured empirically – for example, the influence of smoking on lung cancer. These analyses are frequently used to determine the long-term effects of a lifestyle, diet, or other interventions. Cohort studies may involve a control group that did not participate in the same intervention. Although these studies are a step up in reliability and generalizability, they can be challenging to blind, cannot be controlled for outside variables, and are usually not randomised. Nevertheless, cohort studies are of particular value in epidemiology, helping to understand what factors increase or decrease the likelihood of developing the disease (51–53).

Some of the most notable cohort studies are the British Doctors Study and the Framingham Heart Study. In 1951, The British Doctors Study recruited and followed up over 40 000 participants, monitoring mortality rates and causes of death over the subsequent years and decades. The first set of preliminary results in 1954 presented evidence linking smoking with lung cancer and increased mortality. Over the following decades, the study provided

additional evidence of the health risks from smoking and was extended to investigate other causes of death (e.g., stroke) and other behavioural variables (e.g., alcohol intake) (54).

The Framingham Heart Study commenced in 1948 and is now following up a third generation that includes grandchildren of the original cohort of participants from a Massachusetts town. The study has provided extensive data on the risk factors for cardiovascular disease and fortified international guidelines on prevention (55).

RCTs are considered the highest among primary studies as the hierarchies of evidence rank studies according to the probability of bias. Ideally, they are designed to be unbiased and have a lower risk of systematic errors. For example, by randomly allocating subjects to two or more treatment groups, RCTs also randomise confounding factors that may bias results. Nevertheless, randomisation alone is not sufficient for minimising bias; it is expected that RCTs will also use allocation concealment and blinding of the key individuals involved with a trial (56).

However, it is recognised that RCTs are not always adequately designed, conducted and reported. Therefore, it has been recently suggested that study design alone may be insufficient on its own as a surrogate for the risk of bias. Methodological limitations of a study, such as imprecision, inconsistency and indirectness, are unfettered from study design and can affect the quality of evidence derived from any study design. The quality of evidence may be rated down due to the methodological limitations of RCTs and imprecision (wide CI that includes considerable benefit and harm). Likewise, the quality of evidence traditionally considered "lower-level" in the pyramid can be rated up if high-quality observational studies are available (57).

1.5 Sharing data from clinical trials

1.5.1 Importance of clinical trial data sharing

A randomised controlled trial (or RCT) is a scientific (often medical) experiment that aims to decrease bias when examining the effectiveness of new therapeutics; this is conducted by randomly allocating participants to two or more groups, exposing them to different treatments, and comparing their measured response. There are usually two groups or clusters. The experimental group receives the assessed intervention, and the control group receives a placebo, standard treatment or no intervention. The groups are carefully monitored to

determine the benefits of the experimental intervention, and efficacy is assessed compared to the control group. An RCT is perceived to produce the most rigorous evidence of effectiveness without biases, assumptions and limitations (58).

The growing interest in open science is driven by ethical and scientific imperatives and technological possibilities. The availability of individual patient data (IPD) from clinical trials allows further analyses of clinical trial data, increasing completeness, accuracy and fidelity of evidence about medical interventions, thus the reliability of evidence needed for health decision-making. They can also improve research integrity and reduce research waste, which would benefit patients and society (59,60).

It is imperative to share data that has been obtained via public funding, following the argument that such results should become public property that is adequately deposited and safeguarded (61). Clinical trial data can sometimes be obtained on request from clinician trialists and sponsors, but such requests can be ignored or denied (62). In such cases, authors or sponsors often provide unconvincing reasons for refusing to provide full access to clinical trial reports (63). Even access to summary data from trials can be rejected without providing a reason (64,65).

The total value of any clinical trial can be achieved only if the obtained research data are accessible to the research community and others who might use them (66). Data sharing involves deposition and preservation of data, and it is primarily associated with enabling access for the use of previously collected data. Consequently, there has been a rich ongoing discussion on data preparation for public sharing (67,68).

Numerous stakeholders have been taking initiatives aiming at improving the reliability of evidence and reducing research waste by broader sharing and reuse of clinical trial data, including the International Committee of Medical Journal Editors (ICMJE) (69), Institute of Medicine (IOM) (70), World Medical Association (WMA) Declaration of Helsinki (71), CORBEL project (72), regulators, such as the European Medicine Agency (EMA) and others (73).

To be fully open and reusable, i.e., analysable, clinical trial data sets need to be published in an open access repository that allows text mining and other forms of unrestricted access for reuse. Research data repositories (repositories) are electronic databases that host research data

and facilitate their re-use (74). For data to be reusable, clinical trial datasets published in a repository should contain anonymised clinical trial data and metadata, information on the statistical methods, sample definition details, and any other relevant supporting information (75).

1.5.2 Existing repositories for hosting clinical trial data

In an unpublished study conducted at the University of Split School of Medicine during 2012-2016, repositories hosting clinical trial data were identified, among other types of data. For the study, Google, Re3data and Databib were searched. Researchers found fourteen (ArrayExpress, B2Share, DRUM, Dryad, EASY/DANS, Edinburgh Data Share, Figshare, Harvard Dataverse, ICPSR, LSHTM Data Compass, Open Science Framework (OSF), Swedish National Data Service (SND), University of Bath Research Data Archive, Zenodo) repositories that among other types of data hosted clinical trial data.

In 2019, Banzi et al. published a manuscript that analysed data repositories available in 2018 and assessed their suitability for hosting clinical trial data. Initially, they identified 55 repositories as potentially relevant and narrowed the analysis to 25 repositories. The authors reported that half of those repositories were generic, meaning that they were not limited to a particular disease or clinical topic, and more than half were launched within the past eight years. The repositories were highly heterogeneous in their characteristics, including entities that developed them. The authors identified multiple shortcomings in those repositories, highlighting that more work is needed on repositories to facilitate sharing clinical study data (76).

1.5.3 Results of previous studies that tried to access data from clinical trials

Several studies reported the results of their authors to obtain data from clinical trials from various information sources and platforms. Ross et al. analysed inquiries from 2013 until August 2018 for accessing clinical trial data on the YODA platform. The summary result from the first five years of operating stated that the YODA platform approved 19.3% of inquiries that led to complete clinical trial data sharing (77).

Navar et al. analysed how many RCTs were available through three publicly available platforms, including SOAR, YODA and Clinicalstudydatarequest.com (CSDR), from the inception (first in 2013) until the end of 2015. Major pharmaceutical companies finance those

platforms and use them to disclose clinical trial data purportedly. Out of 3255 trials, requests for data from 505 trials were filed in the analysed period. There were 234 data requests; 177 were adequately filed and met proposed requirements, four were withdrawn, ten were under consideration, and 12 were rejected by the review panel, leaving 154 approved requests. The end of the process that leads to data sharing was completed for 113 requests, leaving 41 requests unanswered about what happened to them. Analytic goals of those requests differed. The proposals for validation studies were rare, making only 4.4% of the requests (78). The authors found only one publication (79) that emerged from such validation proposals. This sole published validation study (79) found contradictory results from initially published findings (80) in Study 329 about the efficacy and harms of paroxetine and imipramine in the treatment of major depression in adolescence (78).

Vaduganathan et al. performed similar analyses to determine the availability and use of data generated in different phases of cardiometabolic clinical trials. Their only resource was Clinicalstudydatarequest.com that hosts patient-level data from thirteen prominent pharmaceutical companies. They found that among all trial records, 16% evaluated cardiometabolic interventions. The average time from study completion to data availability was six years and seven months. Out of 318 proposals (the proposal is an official inquiry or request towards the data holder), 163 have signed a data use agreement, meaning they have met all requirements defined by the data holder to send the data to the inquirer. Data use agreement is the last step before the data is sent to the third party. Only 30 data use agreements were related to cardiometabolic inquiries. Half of the data-sharing proposals were unfounded; most proposals were secondary hypothesis-generating questions with only one proposal for data validation (reanalyses) of the original primary hypothesis. Furthermore, only three publications arose from the shared data (81).

Miller et al. in 2019 published an article reporting that the authors developed a tool to measure pharmaceutical companies' data sharing policies and practices. They reported that only 5 out of 20 analysed pharmaceutical companies made participant-level clinical trial data accessible to external researchers for new drug approvals. They also found that the two companies had no data sharing policies. When examining results for new medicines applications to regulatory agencies, 42% of them had results publicly available in some form, six months after FDA (United States Food and Drug Administration) approval, meaning that other 58% were not in compliance with the Institute of Medicine's recommendations (82).

In 2015, Le Noury et al. (79) published in BMJ a reanalysed raw data from SmithKline Beecham's Study 329, originally published by Keller et al. in 2001 in the *Journal of the American Academy of Child and Adolescent Psychiatry* (80). The primary objective of Study 329 was to compare the efficacy and harms of paroxetine and imipramine with placebo in the treatment of major depression in adolescents. Authors that reanalysed raw data from Study 329 found that the effectiveness of imipramine and paroxetine was not clinically or statistically considerably different from placebo for any prespecified primary or secondary efficacy outcome. However, there were clinically substantial increases in harms, including cardiovascular problems in the imipramine group, along with suicidal thoughts and behaviour and other adverse events in the paroxetine group. These findings of Noury et al. were opposite of the results published in the manuscript by Keller et al., which promoted paroxetine use in adolescents (79).

Safety concerns regarding paroxetine use among adolescents were first raised in the BBC documentary Panorama: The Secrets of Seroxat broadcasted in 2002. Events that followed the documentary, initiated by the British Medicines and Healthcare products Regulatory Agency, led to the pharmaceutical company (now called GSK, after the merger of Smith Kline Beecham with Glaxo Wellcome) being fined three billion dollars in the United States in 2012, for withholding the data that adolescents taking paroxetine are prone to suicide, behavioural changes and self-harm (83). Unfortunately, until today, the controversial publication of Keller et al., which reported data in favour of paroxetine use among adolescents, has not been retracted (80).

1.6 **Researchers' willingness to share scientific data**

Several surveys have explored researchers' willingness to share their research data. Weng et al. did a two-site survey of medical centres personnel, comprising faculty, staff and students, on willingness to share clinical data for research. The article did not report the date when the study was conducted. The results showed that 56% of respondents were "somewhat/definitely willing" to share clinical data *with* identifiers, while 89% were "somewhat/definitely willing" to share *without* identifiers. They concluded that a considerable fraction of potential patient participants would be willing to donate their de-identified clinical data to a shared research repository once educated about benefits and risks. However, this survey only explored self-

professed willingness to share data and did not ask participants actually to share their data (84).

Tenopir et al. surveyed 1329 scientists regarding their data-sharing practices from October 27, 2009, to July 31, 2010. They found that the most common denominator for withholding data is a lack of time and funding for data preparation. Most research organisations do not provide the required infrastructure for short-term or even long-term data preservation, which is controversial because the data lifecycle is not independent of the research lifecycle. Respondents agreed they would be willing to share data if some stimulus is provided, for example, in the form of authorship or citation. The authors concluded that "*Barriers to effective data sharing and preservation are deeply rooted in the practices and culture of the research process as well as the researchers themselves*." The authors also warned that new mandates for data management plans from federal agencies and global attention, aimed towards the need to share and preserve data, could lead to changes (67).

Savage and Vickers requested data from ten scientists who had published articles in PLoS (Public Library of Science) journals, which have specific data sharing policies. However, the article did not report when the study was conducted. Two email addresses were invalid; three authors did not respond; four responded and refused to share the data; only one out of ten contacted had sent an original dataset. The authors' reasons for withholding data were concerns regarding patient privacy and deidentification, rights to retain exclusivity to data that took them a considerable amount of time to produce, and future publication opportunities (85).

In 2009, Weitzman et al. surveyed 261 patients about their disposition to share a wide variety of personal medical data with "outside providers" (clinicians) as well as with public-health researchers at the state or local public health department. Participants were young adults and parents who manage a personally controlled health record (i.e., an electronic health record that holds detailed medical data and permits a patient to decide who can access certain information). Weitzman et al. found that most users were willing to share all information categories with the state or local public health authority (63.3%) rather than with an outside provider (54.1%). The authors further concluded that only a few patients would not share any information category (86).

In another study performed by the same group of authors, 151 personally controlled health record participants (meaning patients themselves have full access to their health record and can control who can access their medical data and for what purpose) were asked about sharing personal data for health research; generally, Weitzman et al. found that 91% (N=138) were willing to share medical information. Participants' willingness to share was mandated by anonymity, use in research purposes, involvement with a trusted intermediary, transparency around personally controlled health records, access and use, and payment (87).

In March 2018, publisher Springer Nature posted on it website a whitepaper of practical challenges for researchers regarding data sharing. The whitepaper examines the results of a large-scale survey with over 7700 participants, which analysed the challenges researchers encounter in sharing their data. The study findings confirm that researchers' efforts to archive, publish and share data continue to be hindered by time restrictions and a lack of knowledge regarding data standards, metadata and deficiency in data curation, repository options, and funder requirements (88). In the Springer Nature survey, 63% of participants responded that they generally submit data files as supplementary information or deposit the files in a repository when submitting their work in a journal; 76% of participants rated 7.3 out of 10 on the importance of enabling their data discoverable; 25% of participants rated data discoverability importance as 10 out of 10. The most significant challenges to data sharing were identified as: ‘Organising data in a presentable and useful way’ (46%), ‘Unsure about copyright and licensing - 37%, ‘Not knowing which repository to use’ - 33%, ‘Lack of time to deposit data’ - 26%, ‘Costs of sharing data’ - 19% (88).

The survey noted a difference between researchers' seniority on time and knowledge data-sharing issues. When asked about data sharing problems, time was a more significant issue with senior researchers (29% for senior and 23% for early career researchers); 40% of early career researchers state they do not know where to share data, as opposed to 30% for the most senior researchers. Uncertainty about copyright and licensing worried 43% of early career researchers and 33% senior researchers. Cost concerns remain low as a stated factor throughout various career stages (18-20%), issues with organising data in a presentable and practical way stay high throughout (48-49%) (88).

Dataset size may also affect whether data are shared; 42% of respondents that generate the smallest data files (less than 20MB; n = 2,036) had the highest amount of data that are neither deposited in a repository nor shared as supplementary information in the manuscript. On the

contrary, 70% of researchers with data files larger than 50GB (n = 700) shared their data, with a strong predisposition for sharing through repositories (59%) (88). The discoverability of research data was rated highest in the biological sciences (7.8 out of 10), geographical sciences (7.7), medical sciences (7.2), and physical sciences (6.6), which matches with data sharing practices in these fields. The biological sciences had an admirable segment of researchers who share data associated with publications (75%), followed by the geographical sciences (68%), medical sciences (61%), and physical sciences (59%) (88). The lack of time was more significant concern to researchers in Europe, Australia and North America, while data sharing costs are more of a problem to respondents in Asia and South America (88). Distinguished problems to data sharing also varied between subject areas. Organising data helpfully ranged from 57% in the physical sciences to 40% in the medical sciences. Copyright and licensing issues varied from 44% in the medical sciences to 31% in the physical sciences. Not knowing which repository to use varied from 37% in the medical sciences to 27% in the physical sciences (88).

Only 54% of researchers who produced specific biological and medical data, where area-specific repositories exist, used available repositories to share their data (88).

1.7 **Data Availability Statement**

Data Availability Statement (DAS), sometimes also called a Data Access Statement or Availability of Data and Materials, is a statement that informs the reader where research-associated data is obtainable and whether there are restrictions in accessing the data. If it is manageable, DAS should contain a hyperlink to a publicly available repository that hosts datasets generated or analysed during the study (89–91).

The first initiative for declaring DAS in a research manuscript came from the PLoS publisher. The purpose of the DAS was to promote data visibility, encourage data re-use and increase the reproducibility of published research. The initiative was first posted on the PLoS website in February 2014, announcing a new era in the PLoS publishing policy where every author will have to declare a data availability statement in the manuscript. The DAS requires authors to make all data needed for study replication publicly available without any restrictions at the time of publication. Alternatively, if there is a specific barrier opposing data sharing, the author is obligated to explain the mechanism of how another researcher can access data (92).

The PLoS' editorial team also guides authors to what data should be submitted. They named it "the minimal dataset", meaning there is no need to submit all raw data, images, simulations and early models or iterations. Instead, it means that the data supporting the final model or discovery presented in the manuscript, with enough information (data and metadata) to replicate the study, should be submitted along with the manuscript. In specific cases, if the researchers have too much data, PLoS offers help to authors in finding an optimal solution for making even large file data available to other researchers. PLoS also invites researchers with delicate patient-level data or legal data to cooperatively find a solution for publishing even that type of data, and this primarily concerns researchers conducting randomised clinical trials distressed regarding patient deidentification (92).

The PLoS also encourages researchers to use subject area repositories such as Clinicaltrials.gov, Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) or National Institute of Health genetic sequence databank (GenBank), or in unstructured publicly accessible repositories such as Figshare, Dryad, and Open Science Framework (OSF) to deposit their scientific data. In addition, smaller amounts of supporting data could be included entirely within the supplementary files in the article (92).

PLoS editors rationally explain probably the most significant researcher concern of all, a fear that some other researcher may benefit from the data. They firmly believe that this concern applies solely to the time before publication. According to them, post-publication, the data should be available to others for reuse. They trust that most of the authors published their work precisely so that others can benefit from it (91,92).

Federer et al. analysed DAS in articles published in PLoS One between March 1, 2014, and May 31, 2016. They retrieved 62,589 articles and identified that 14,928 had not had a DAS. Those 14,928 manuscripts were accepted before the policy came into effect on May 1, 2014, even though they were published later. Further analyses on 47,661 articles showed that 20% of articles direct the reader to a particular repository, 20% state that privacy concerns prevent them from data sharing, and 60% are contained within the text or supplemental material (93).

CHORUS initiative keeps track of publishers and journals that mandate DAS in their article submission policies. Until May 15 2020, CHORUS listed thirty publishers and one journal that implemented DAS as one of the manuscript's integral elements (94).

In 2017, the ICMJE published Data Sharing Statements for Clinical Trials, which included the following requirement: "*As of July 1, 2018 manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement*." The statement contains a table of examples illustrating a range of acceptable data sharing statements; it means that authors must disclose a certain level of data sharing. The second requirement is: "*Clinical trials that begin enrolling participants on or after 1 January 2019 must include a data sharing plan in the trial's registration*." (95)

In 2019, Sun Huh opined that there is still insufficient evidence that data sharing stimulates scientific development in clinical medicine. Four significant challenges were identified: first, there are not many clinical studies that have used data sharing; second, consensus regarding data structure and consistency in the data is still a significant challenge; and third, research design regarding a plan for further raw data sharing has to be created as a new standard in research planning, and finally it is still not mandatory to deposit clinical data in the publicly accessible repository (96).

However, the prevalent opinion is that there is no doubt that data sharing is valuable to the scientific community; it significantly advances the research field allowing other researchers to use original data presented in the manuscript. Meaningful manipulation with data can be helpful to researchers who want to conduct a similar study; insight on the data would allow them to alter their research or abandon the topic entirely; therefore, resources were not wasted on unnecessary research. Reanalysis of original data could provide new insight on the matter, positive as well as negative. Meta-analyses strongly rely upon previous data, and particularly on individual participant data meta-analyses, accessibility of raw data is critical. Moreover, enabling raw data access is part of the solution regarding the reproducibility crisis (97).

1.8 **Research problem**

The ICMJE data sharing policy is as follows [quote]: "*as of July 1 2018, manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement"*(95). It is unclear what proportion of clinical trials that were published before this cut-off date had had a data-sharing statement and the proportion of trialists willing to share their data on request. Furthermore, when the articles contain data sharing statements, it is unknown whether they would share their data on request if this is what they have indicated in their writing. It is also unknown whether authors of clinical trials would be

willing to share their data on request if they did not indicate in their manuscript that they are willing to do so. Our study aims to fill those gaps and help in revising data sharing statements and expectations in the future.

2 AIMS

2.1 Aims of the first study

The first study aimed to examine the effectiveness of data sharing statements and reports of the availability of open data in RCTs published in high-ranking anesthesiology journals from 2014 to 2016 and to explore authors' willingness to share raw data from those trials if they were not openly available (98).

RCTs were chosen because they are considered the highest level of evidence among primary studies and ICMJE policies regarding data sharing. By repeating a similar survey of trials published after 2018, it was possible to analyse whether the ICMJE statement has had any effect. The field of anesthesiology was chosen because issues of pain and anaesthesia are applicable across all clinical areas. The study was initiated in the year 2017, and we chose the period 2014-2016 for analysis because of two reasons; firstly, we wanted to analyse very recent period, and secondly, by taking into analysis studies published before that, the possibility of email decay is increasing, thus diminishing the likelihood of successful email contact of trial investigators. We chose high-ranking journals because we anticipated that they should be of higher quality in terms of critical observation regarding the scientific contribution of submitted manuscripts. We also presumed that those journals should be better compared to low-ranking journals in terms of fostering open data and data sharing.

2.2 Aims of the second study

The second study aimed to examine the effectiveness of DAS categories in Nature Springer BioMed Central (BMC) open access journals and to examine whether authors who indicated they would share their data on request would comply with their DAS. These journals were chosen because we were unaware of another large set of journals with the exact requirements regarding data sharing statements. Since DAS is mandatory for all manuscripts submitted to

BMC journals, this selection criterion provided hundreds of journals appropriate for the analysis.

2.3 **Hypotheses of the first study**

Hypotheses for the first study (conducted on a sample of clinical trials published in high-ranking anaesthesiology journals) were:

- Majority of researchers who published clinical trials in high-ranking anesthesiology journals did not engage in open data, i.e. did not share openly raw data collected within their trial, either as a manuscript supplement or in an online repository.

- Majority of researchers who published clinical trials in high-ranking anesthesiology journals will share raw data from their trial on request via email.

2.4 **Hypotheses of the second study**

Hypotheses for the second study (conducted on a sample of open-open access journals mandating the use of data availability statements) were:

- Majority of researchers who published manuscripts in an open-access journal that mandates the usage of DAS will indicate in their DAS that data is available only on request.
- Majority of researchers who indicate in their DAS that they will share data on request will be willing to share their data following their DAS.

3 METHODS

3.1 First study

3.1.1 Study design

The first study included a primary methodological study and a cross-sectional survey of corresponding authors.

3.1.2 Ethics

The Ethics Committee of the University of Split School of Medicine approved the study protocol. Scanned approval of the Ethics Committee was available to the study participants on request, which is available in Appendix 1.

3.1.3 Unit of analysis

We analysed RCT reports published in high ranking anesthesiology journals (a methodological study) and responses of corresponding authors of those RCTs (a cross-sectional survey).

3.1.4 Sample

We included RCTs of interventions published from January 1, 2014, to December 31, 2016, in seven journals from the Journal Citation Reports category Anesthesiology that belongs to Q1, that is, the highest-ranking 25% of journals in this category, according to the Clarivate Analytics' Journal Impact Factor distribution. Based on the 2015 Journal Citation Reports (JCR) Journal Impact Factor and in alphabetical order, the analysed journals were: Anaesthesia, Anesthesia and Analgesia, Anesthesiology, Pain, British Journal of Anaesthesia, European Journal of Anaesthesiology, and Regional Anesthesia and Pain Medicine. Information about corresponding authors of those RCTs was extracted manually from included publications.

3.1.5 Search

We searched MEDLINE (via PubMed) via advanced search using journal names, a filter for RCTs, and a filter for publication dates 2014 - 2016. We exported titles and abstracts into

reference management software. Two researchers MG and JČ independently screened titles and abstracts, and if necessary, full texts to verify that those studies were indeed RCTs. We resolved any discrepancies in opinion via discussion. Full texts from all included RCTs were downloaded for further analysis.

3.1.6 Outcomes

The primary outcomes of the first study were:

- Number of RCTs with openly shared data published in the journal as supplementary materials or in the data repository
- Number of RCTs with data sharing statement or data availability statement
- Number of RCT authors who shared their raw data on request.

The secondary outcome of the first study was the number of RCT authors who indicated in the trial report that they would share their data on request who complied with the data sharing request.

3.1.7 Data extraction

We analysed full texts of the included RCTs and extracted data into a Microsoft Excel worksheet (Microsoft Inc., Redmond, WA, USA). The worksheet was piloted with five studies to make sure that it is suitable to extract target data. We extracted the following data: name of the first author, year of publication, email address of the corresponding author, source of funding (commercial/industry or non-profit), presence of data sharing information (a statement that raw data set is available in a specific repository, or available on request, or no such statement), and names of repositories where raw data sets were made public.

3.1.8 Author survey

All RCT authors that did not indicate that data were available in a publicly available repository (i.e., those that indicated that data were available on request and those that did not have any statements regarding open data sharing) were contacted via email. The first author (MG) sent personalised emails to each potential participant from a personal email account. A de-identified copy of the email sent to the RCT authors is available in Appendix 2. If the authors did not respond after the initial email, they received only one reminder two weeks after the first email. Likewise, if the authors reacted positively with a willingness to share raw data sets but did not provide data within two weeks, they received an additional email

reminder. Corresponding emails were obtained directly from the manuscripts of the examined studies. All emails and initial reminders were sent between January 26 and February 27 of 2018. Communication with several authors who requested additional information continued throughout March and April of 2018. If the corresponding authors suggested we contact another team member to obtain data and provided their email addresses, we contacted them. If the authors indicated that additional regulatory or approval procedures were required for obtaining raw data sets, we did not engage in those processes. This was due to our previous experience of such requests taking years to receive responses. Even then, there is a possibility for a data request to be refused without explanation. If the message sent to the corresponding authors was returned undelivered, we did not attempt to find their alternative email address. If the corresponding authors did not respond, we did not attempt to contact other co-authors.

After accessing raw data sets, we checked whether data were available in a way that enabled reanalysis, that is, published in a file that enabled data use and whether sufficient metadata (description of data and variables that would permit re-use) were included.

3.1.9 Data analysis

We used a convenience sample of the most recently published RCTs within three years, considering that this would be a large enough sample to notice the current state of open data sharing. We presented descriptive data as frequencies and percentages. Differences in proportions were analyzed using a chi-square test. Analyses were conducted using MedCalc statistical software, v 15.2.1 (MedCalc Software bvba, Ostend, Belgium). Statistical significance was set at $P < 0.05$. We used Fisher test to explore differences in data sharing between trials that had commercial versus noncommercial funding.

3.2 **Second study**

3.2.1 Study design

The second study included a primary methodological study and a cross-sectional survey of corresponding authors.

3.2.2 Ethics

The Ethics Committee of the University of Split School of Medicine approved the study protocol. Scanned approval of the Ethics Committee was available to the study participants on request.

3.2.3 Unit of analysis

We analysed data availability statements (DAS) from manuscripts published in open-access journals and responses of corresponding authors of those RCTs (a cross-sectional survey).

3.2.4 Sample

We included manuscripts published during January 2019 in all open-access journals from BioMed Central (BMC; part of Nature Springer). Based on the BMC webpage, there were 333 journals in March 2020. We did not separate these manuscripts into groups based on study type. Extraction of journal name, article title and availability of data and materials was done using a computer web scraping tool (available: https://github.com/bojcicm/bmc-scrapper). One thousand articles were manually extracted to verify the computed findings. Computer-extracted data were in complete concordance with the data extracted manually. Only one shortcoming was detected in the data extraction tool; it did not recognize journals with zero published articles. That flaw was manually adjusted in the spreadsheet with extracted data.

3.2.5 Outcomes

The primary outcomes of the second study were:

- Categories of DAS in analyzed open-access journals.
- The number of authors whose DAS indicate they will share their data on request who will comply with the data sharing request.

The secondary outcome of the second study was the number of RCTs and open-access publications for which authors shared data that are eligible for re-analysis. We analysed RCTs specifically because we investigated RCTs in the first study.

3.2.6 Data Extraction

The corresponding author name and corresponding author email were extracted only if DAS is category 2, 4 and 6. We extracted data into a Microsoft Excel worksheet (Microsoft Inc., Redmond, WA, USA). Two authors piloted the worksheet on a sample of twenty articles to make sure that it is suitable to extract target data.

From manuscripts in eligible journals, we extracted the following data: Journal name, ISSN, DOI for each article, article title, DAS copied verbatim, DAS category 1.-6.

DAS categories:

1. The authors have indicated in which repository they deposited datasets, and they should provide a web link to the datasets.

2. The datasets are available from the corresponding author on reasonable request.

3. All data generated or analyzed is included in this published article and its supplementary information files.

4. The datasets generated are not publicly available due to disclosed reasons but are available from the corresponding author on reasonable request.

5. Data sharing does not apply to this article because no datasets were generated or analyzed during this study.

6. The data is available from a third party, and restrictions apply regarding data availability because data were used under license and therefore are not publicly available. Data is, however, available upon reasonable request and with the permission of the licence holder (9).

These categories were defined according to the Springer Nature Data Availability Statements guidance for authors and editors.

For manuscripts that were not eligible for classification according to Springer Nature Data Availability Statements guidance for authors and editors, we created a seventh and eighth category: 7. Not available - for statements that claim data is not available to the third party

under any circumstances. 8. Other – for statements that cannot be classified as categories one to seven.

Some manuscripts had dual or triple coding because DAS had elements of several categories. All double and triple codes containing codes 2, 4 or 6 were included for contacting corresponding authors.

3.2.7 Author Survey

All corresponding authors of manuscripts with DAS category 2, 4 and 6 were contacted via email and asked to share their raw data sets. A de-identified copy of the email sent to the authors is available in Appendix 3. We have also prepared a Non-Disclosure Agreement (NDA) and Ethical Committee Approval for the second study for researchers that might request one or both of those documents from us (available as Appendix 4 and 5, respectively). The first author (MG) sent personalized emails to each corresponding author from a personal email account. If the authors did not respond after the initial email, they received only one reminder. If the authors responded positively with a willingness to share raw data sets but did not provide data within two weeks, they received an additional email reminder. Corresponding emails were obtained directly from manuscripts included in the first part of the study. All emails and initial reminders were sent between January 18th and May 18th in 2021. If corresponding authors suggested we should contact another team member to obtain data and provided their email addresses, we contacted those persons. If the authors indicated that additional regulatory or approval procedures were required for obtaining raw data sets, we did engage in those processes, such as signing nondisclosure or data transfer agreements or sending an official letter of request signed by the University of Split School of Medicine official. If the message sent to the corresponding authors was returned undelivered, we did not try to find their alternative email address. If the corresponding authors did not respond, we did not attempt to contact other co-authors.

After accessing raw data sets, we checked whether data were available in a way that enabled reanalysis, that is, published in a file that enabled data use and whether relevant metadata were included.

3.2.8 Data Analysis

We used a convenience sample of all BMC publications from January 2019, considering that this would be a large enough sample to notice the current state of data sharing. We presented descriptive data as frequencies and percentages. Differences in proportions were analyzed using a chi-square test. Analyses were conducted using MedCalc statistical software, v 15.2.1 (MedCalc Software bvba, Ostend, Belgium). Statistical significance was set at $P < 0.05$.

4 RESULTS

4.1 First study results

4.1.1 Results

Our study included 619 RCTs published in seven anaesthesiology journals between January 1, 2014, and December 31, 2016. The maximum number of RCTs was published in the Anesthesia and Analgesia (N = 112), after that in the British Journal of Anaesthesia (N = 103), Pain (N = 97), Anesthesiology (N = 90), Anaesthesia (N = 86), European Journal of Anaesthesiology (N = 66), and Regional Anesthesia and Pain Medicine (N = 65) (98).

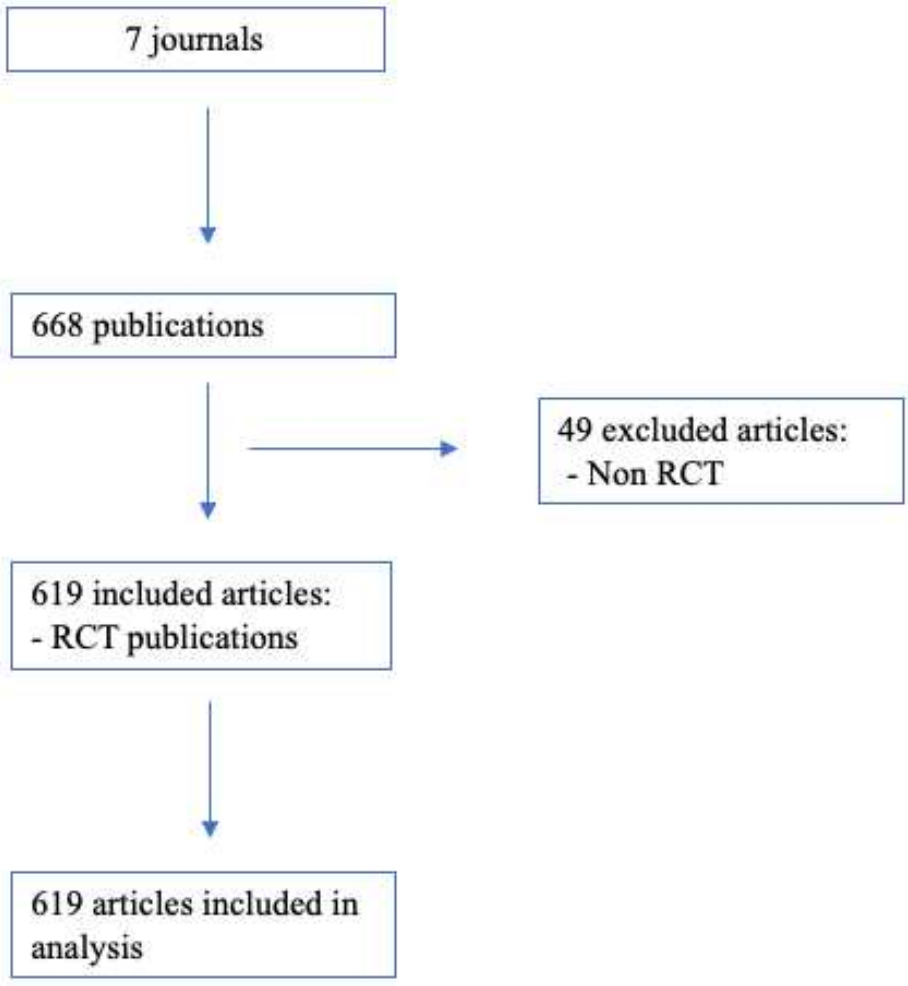


Figure 1. Flow chart of the first study

Most of the RCTs had non-profit funding (N = 439; 70.92%); the remaining used commercial funding (N = 80; 12.92%) or declared no funding (N = 51; 8.23%). In 49 (7.91%) trials, statements regarding funding were unreported.

4.1.2 Data sharing in manuscripts describing RCTs

Among the 619 analyzed RCTs, not one provided raw data within the manuscript or an internet link to the repository containing raw data. We found 24 studies with data sharing statements. Four studies were published by two different research groups (each of the two groups published two of those studies) the other 20 were published by 20 different research groups. One research group posted a total of six studies from our cohort, and they provided data sharing statements in two out of six studies, but upon request, they did not provide data for any of them.

4.1.3 Requesting data from trial authors

We contacted 619 RCTs corresponding authors. We received 31 emails that bounced back as undelivered. Out of the remaining 588 manuscripts, we received responses from 86 (14.63%); 502 (85.37%) authors did not answer our query. From the 86 responses, further raw data were only obtained from 24 (3.83%), whereas 62 were unwilling to share raw data.

Seventeen (2.75%) corresponding authors stipulated that they do not have raw data for sharing and gave us the email address of another person to address our inquiry. Six of those 17 other individuals responded. Twelve (1.94%) automatic messages designated that the recipients were temporarily away; those authors were contacted again on April 21, 2018; three out of twelve responded.

Among 62 authors who declined to provide data, not all provided a reason. Table 1 summarizes the responses of those authors. The most common two explanations were that they do not own data and participant privacy concerns. All those who responded that data were not theirs to share indicated that this was imposed by the rules of their country, institution, or study sponsor. Three researchers stated their study was not an RCT, though it was described in manuscripts.

Twelve corresponding authors who expressed privacy concerns as reasons for withholding raw data had the following explanations: local research ethics approval prevents or might

prevent disclosure of data in any way (N = 3); country-specific legislation prohibits sharing of raw data; this was the instance for Norway (N = 1) and Denmark (N = 3); trial participants were not asked for consent to share their anonymized data with other investigators(N = 2); university regulations disallow raw data sharing as this information is not considered public property (N = 2), and the raw data are not anonymized hence inadequate for sharing (N = 1).

Table 1. Responses received by authors of randomized controlled trials who did not share raw data on request

Responses	**N (%)**
Request for raw data refused immediately without reason provided	18 (29.03)
The corresponding author requested more data about our study, but after receiving it, they declined to provide raw data	13 (20.97)
Data are not mine, and therefore, I cannot share raw data	12 (19.35)
Participant privacy concern	12 (19.35)
Our research was not a randomized controlled trial	3 (4.84)
Trials are still ongoing, and the data are still in use; therefore, they cannot share the data	2 (3.23)
An author wanted co-authorship in exchange for raw data	1 (1.61)
The corresponding author wrote back that another person would provide raw data and gave us an email address; we contacted this other person but never received a response	1 (1.61)
Total	62 (100)

4.1.4 Response from authors who had a data-sharing statement

Of the 24 corresponding authors with manuscripts that included a statement regarding raw data availability, only one sent the data upon request. Eighteen of those related authors did not reply, three replied they could not share their data, one stated they would get back to us in 4 months (but did not), and one sent an affirmative email that they were sending data, but without data attached, and did not reply on request to kindly provide the attachment.

4.1.5 Raw data that was shared

In total, we received 24 raw data sets from 24 manuscripts, provided by 19 authors; however, two data sets were sent in pdf format; 22 (3.6%) usable data sets from a total of 619 manuscripts were retrieved. The data sets we obtained were in Microsoft Excel and SPSS format. The median response time was six days. Among the 24 manuscripts whose authors provided raw data sets, 20 (83.33%) had non-profit funding, whereas the remaining 4

(16.66%) had commercial funding. Statistical significat difference was found in the proportion of data sharing between manuscripts with commercial financing and those without commercial funding (Fisher test, $p < 0.001$).

4.2 **Second study results**

4.2.1 Data extraction analysis

In January 2019, BMC had 333 registered journals; 51 (15,31) did not publish a single article in the observed period, while 282 (84.68%) published at least one paper. From the 282 journals, we extracted 3556 articles. We excluded 68 articles from the analysis because they were not primary publications; namely, 63 were corrections to previously published articles, and five were reports from conferences and symposiums. Among the remaining 3488 pieces, we further excluded 72 articles that did not have a DAS. Thus, we included in our analysis 3416 articles with DAS (Figure 2).

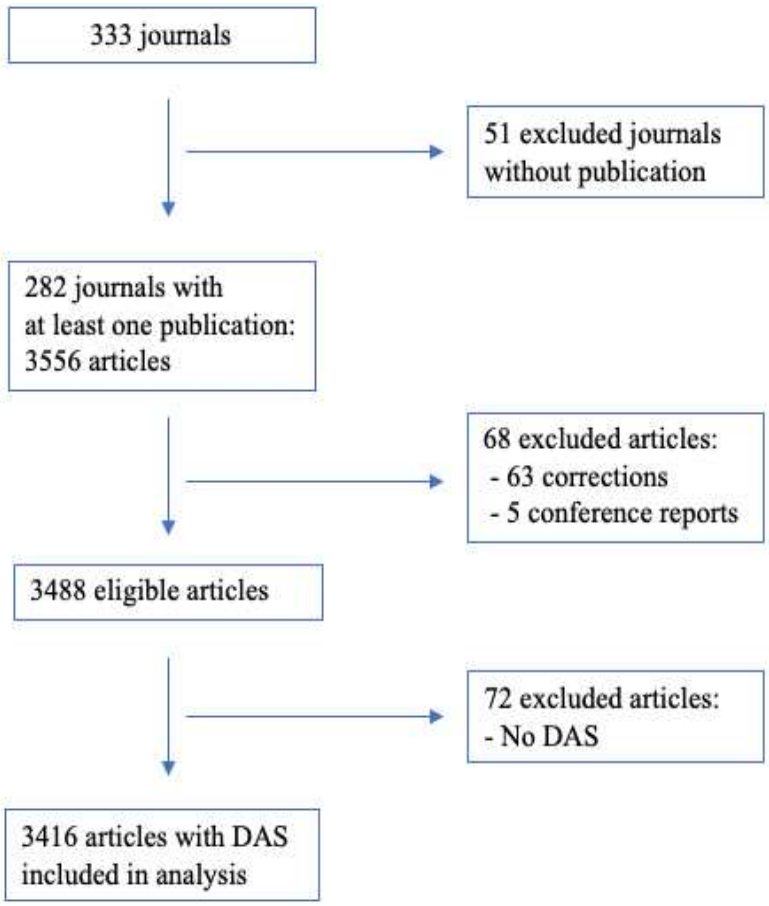


Figure 2. Flow chart for the second study

The most commonly used DAS category was the second one (the datasets are available from the corresponding author on reasonable request), followed by the third one (all data generated or analyzed is included in this published article and its supplementary information files). The minority of the articles were classified into multiple categories (Table 2).

Table 2. Frequency of different categories of data availability statement (DAS) (N=3416)

DAS category	N (%)
1. The authors have indicated in which repository they deposited datasets, and they should provide a web link to the datasets	381 (11.15)
2. The datasets are available from the corresponding author on reasonable request.	1437 (42.07)
3. All data generated or analyzed is included in this published article and its supplementary information files.	577 (16.89)
4. The datasets generated during are not publicly available due to disclosed reasons but are available from the corresponding author on reasonable request.	166 (4.86)
5. Data sharing does not apply to this article because no datasets were generated or analyzed during this study.	419 (12.27)
6. The data is available from the third party; restrictions apply regarding data availability because data was used under license and therefore are not publicly available. Data is, however, available upon reasonable request and with permission of the licence holder	119 (3.48)
7. Not available - for statements that claim data is not available to the third party under any circumstances	95 (2.78)
8. Other – for statements that cannot be classified as categories one to seven	93 (2.72)
DAS categorized into two categories	127 (3.72)
DAS categorized into three categories	2 (0.06)

Among 129 articles with DAS categorized into more than one category, the most common combination category of DAS was 1 and 3 (Table 3).

Table 3. Frequency of data availability statements categorized into more than one category (N=129)

DAS categories	N (%)
1 and 2	10 (7.75)
1 and 3	48 (37.21)
1 and 8	7 (5.42)
2 and 3	38 (29.45)
2 and 4	4 (3.10)
2 and 6	1 (0.77)
2 and 7	2 (1.55)
2 and 8	1 (0.77)
3 and 4	2 (1.55)
3 and 7	1 (0.77)
4 and 6	4 (3.10)
4 and 7	1 (0.77)
6 and 7	2 (1.55)
6 and 8	3 (2.32)
7 and 8	3 (2.32)
1 and 2 and 3	2 (1.55)

4.2.2 Contacting the authors

The total number of manuscripts eligible for contact was 1792 out of 3416. We contacted all the 1792 corresponding authors from the eligible manuscripts to request their data. After our initial e-mail, we received no reply from 1461 (81.53%) contacts, and 77 (4.30%) e-mails bounced back as undelivered. There were 38 corresponding authors that instructed us to contact another researcher responsible for data management, retrieval and sharing, and provided us with a forwarding address; 29 of those contacts did not respond. From 17 researchers, we received an automated email stating they were unavailable at the time, and we should contact them after the "away" period, which we did; only one researcher responded to the reminder e-mail sent after the "away" period. In summary, of 1792 e-mails sent, we did not receive any response for 1538 articles because messages were either undelivered (N=77; 4.29%) or the author did not reply (N=1461; 81.53%). We received responses from 254 (14.17%) contacted individuals. A flow chart depicting the outcome of contacting the authors is shown in Figure 3.

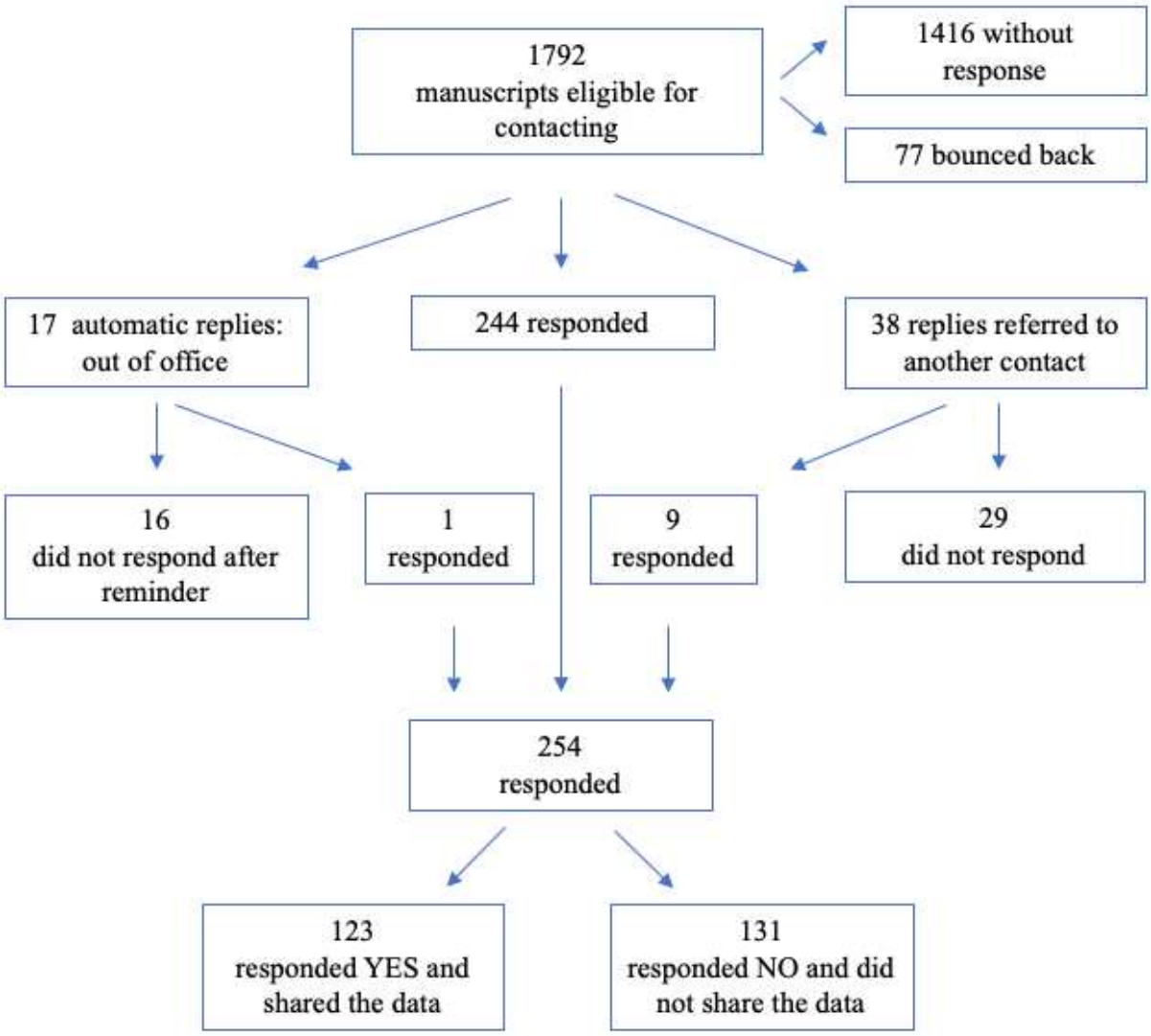


Figure 3. Flow chart depicting the outcome of contacting the authors in the second study

4.2.3 Response analysis

Among the 254 authors that responded, 123 shared the data, which corresponds to 7,17% of 1715 articles for which authors were contacted, and e-mails did not bounce back undelivered. Reasons for not sharing data among the remaining 131 authors who responded to our request for data are provided in Table 4.

Table 4. Reasons for not sharing requested data (N=132)

Reason	N (%)
The authors asked for more information about our study, but after our detailed response and clarification, we did not receive further response from them	23 (17.42)
Their informed patient consent did not include sharing data with other researchers, or the ethical committee prohibited external data sharing and use	14 (10.6)
They cannot access the data, either because they are no longer in the institution that conducted the research or they are no longer active on the project	12 (9.09)
They do not want to share the data or in any way participate in our study without a specific explanation	11 (8.33)
No reply after we signed and sent an NDA or DTA that the authors requested	10 (7.57)
The corresponding author instructed us to use a web service to request the data	10 (7.57)
The data does not belong to them	6 (4.54)
The study was still ongoing	5 (3.78)
Privacy concerns; specifically, they did not want to share de-identified patient data, and other concern was third party data storage and safekeeping from external access to patient data	4 (3.03)
This was a summary article, and there are no data to share	3 (2.27)
We were unable to meet their specific requests regarding NDA	2 (1.51)
The author requested reimbursement for data sharing	2 (1.51)
The author wanted ethics approval translated into English	2 (1.51)
I will send you the data (however, the authors did not send the data subsequently)	2 (1.51)
After both parties signed the NDA, the author wrote back to say that data sharing is against their Ethical Committee suggestions	1 (0.75)
A conceptual article, no data to share	1 (0.75)
Dataset has 8 Gb and is unable to share such a large file	1 (0.75)
The most important data and the dataset is in the supplementary file in the manuscript (however, this was not the case; the supplementary file did not contain raw data from the study)	1 (0.75)
The author misplaced the data	1 (0.75)
The author was not sure what data we needed (the response remained the same even after we sent three emails to explain what information we needed)	1 (0.75)
Repeat your inquiry in six months	1 (0.75)
The author was sick and did not go to the office	1 (0.75)
Another scientist will send you the data (but they did not)	1 (0.75)
The author wanted to schedule an online meeting but was unable to set the date and time regarding the time zone difference	1 (0.75)
The author wrote they would get back to us (but did not)	1 (0.75)
The author requested vast and lengthy procedures and authorisations	1 (0.75)
I will respond to your message in due course	1 (0.75)
I am writing a grant application and cannot help you	1 (0.75)
The author sent us published supplementary materials	1 (0.75)

"We did not use any of our own data for this publication. The data was gleaned from peer-reviewed publications that are widely available and referenced in the reference section."	1 (0.75)
Qualitative research conducted in Ukrainian and Russian language	1 (0.75)
"What advantage will I get from sharing?"	1 (0.75)
"Too many technical aspects from our side; cannot comply with."	1 (0.75)
Qualitative research conducted in the Finnish language	1 (0.75)
Awaiting ethical committee approval	1 (0.75)
It will take too long	1 (0.75)
All data is within the article (however, the data was not within the article)	1 (0.75)
The author asked if our institution has MATLAB software; no reply after two reminders	1 (0.75)
The author wanted us to sign their DTA, which we agreed to. However, they did not send the DTA or replied to our messages even after we sent them two reminders.	1 (0.75)
Wants that NDA be signed by the rector of the University of Split	1 (0.75)
Do not want to prepare all this for a study that is not interested in the data itself	1 (0.75)
It may be not feasible to do this in the near future	1 (0.75)

Acronyms: DTA = data transfer agreement; NDA = non-disclosure agreement

Thirty authors asked for more information about the study. We replied to them all, and the clarifications sometimes took several e-mails. Of those 30 authors, seven eventually provided their data.

Among 22 authors who requested that we sign an NDA or data transfer agreement (DTA), two authors accepted NDA that we have prepared, while 20 sent their own version of NDA or DTA. We received data sets from eight of those 22 authors. Ten authors did not reply at all after we sent them an NDA or DTA. Two authors had requests regarding NDA that we could not accommodate; namely, one author wanted an NDA signed by the principal (rector) of the University of Split, and another one wanted NDA signed by an official from the university technology transfer office. One author informed us that data sharing is against their Ethical Committee suggestions after both parties signed the NDA (Table 4).

Two authors demanded reimbursement, and one author requested co-authorship for providing us with data from their research. Ten researchers directed us to a web portal we should access and register, with instructions that afterwards we should specifically describe variables we need, and only then would their decision-making start. We did not engage in those activities. Two authors asked us to send them the official letter, on the School letterhead, in which we

will request the data. One of those authors shared the data after sending the official letter, while the other did not.

Various other reasons for not sharing the data, including health condition and retirement of the author, requests for translating the ethics approval for our research into non-English language, large datasets, misplacing the entire data from the study, etc., are shown in Table 4.

4.2.4 Randomized controlled trials among the articles that shared the data

Among the 123 articles for which the authors shared the raw data, there were 11 (8,94%) randomized controlled trials. Based on the authors' self-report, the most common study design of the articles for which authors shared raw data was cross-sectional (Table 5).

Table 5. Study design of articles for which the authors shared the raw data, as reported in the manuscript (N=123)

Study design	**N (%)**
Not reported	47 (38.22)
Cross-sectional	25 (20.32)
Randomized controlled trial	11 (8.94)
Cohort	6 (4.87)
Retrospective	5 (4.06)
Mixed Methods	4 (3.25)
Systematic Review	4 (3.25)
Prospective study	3 (2.43)
Protocol	3 (2.43)
Qualitative study	3 (2.43)
Questionnaire study	2 (1.62)
Descriptive study	2 (1.62)
Observational study	2 (1.62)
Post hoc analysis	1 (0.81)
Case-control study	1 (0.81)
Pre-test/post-test	1 (0.81)
Meta-analysis	1 (0.81)
Interventional study	1 (0.81)
Case study	1 (0.81)
Total	**123**

4.2.5 Usable datasets

Among 123 data sets shared with us, 118 (95.93%) were usable, i.e. they were sent in a format that would allow re-analysis of the data. The five unusable data sets were sent in a pdf. or .doc format.

5 DISCUSSION

Our results indicate that few clinical trial authors were willing to share their data, even if they wrote in their manuscript that data would be available on request. In the first study, 13.89% of authors responded to our query for data sharing, and eventually, 3.87% provided the requested data; these results defy the second hypothesis of the first study. However, very few of the trials included in the study had a DAS that indicated data would be available on request; this confirms the first hypothesis from the first study (98).

Thus, in our second study, we focused only on studies with DAS that indicated the availability of data on request. We received responses from 14.17% of the contacted authors, and data were shared by 6.8% of the contacted authors. These results disproved the second hypothesis of the second study. We confirmed the first hypothesis of the second study since 42.06% (1437) chosen DAS were type two, stating:" The datasets are available from the corresponding author on reasonable request".

The percentage of usable datasets were 3.55% and 6.86% in the first and second study, respectively.

Among the contacted authors in both studies, some provided reasons for not sharing data that offer insight into hurdles that individuals requesting data may face when accessing raw data from published articles.

5.1 First study

After attempting to procure anonymized raw data sets from 619 RCTs that were published in the field of anesthesiology from 2014 to 2016, we received data from 3.87% of trials, but 3.55% of the shared trial data would allow reanalysis. The majority of the contacted corresponding authors did not respond to our query for sharing raw data. The most provided

reasons for not sharing the raw data were issues related to ownership of the data and concerns related to the privacy of trial participants.

A study that analyzed trials published in BMJ and PLoS One, journals that mandate data sharing, concluded that authors' data sharing behaviour was ''not optimal'', despite the journals' ''strong policy for data sharing'' (99). These research groups also tried to reanalyse data obtained from those trials and concluded that they managed to reanalyse and replicate the original results for most analyzed studies. They figured that data sharing should be more widespread, as well as streamlined so that other independent author teams could reanalyse and reuse the data collected in clinical trials (99).

Although the International Committee of Medical Journal Editors in 2016 suggested that responsible data sharing is an ethical obligation of trialists because ''participants have put themselves to risk'', they now require that ''as of 1 July 2018, manuscripts submitted to ICMJE journals that report the results of clinical trials must contain a data sharing statement as described below'' (100); however, our study shows that having a data sharing statement is not indicative in actually providing data. Data were received from one of 24 manuscripts that incorporated a data sharing statement in their manuscript, indicating that authors' behaviour is not in line with their statements. Judging by the instructions for authors we analyzed, if we assume that these instructions did not change, the authors of those 24 manuscripts were not obliged by the journal's instructions to provide a data sharing statement.

The studies we analyzed were published in seven high-impact journals in the field of anesthesiology, and of those seven journals, only one, Anaesthesia, had a raw data sharing policy, which ''encourages'' authors to share data and indicates that authors ''should'' share their data; however, among reports that had data-sharing statement within or articles for which authors provided data, not a single one was published in the Anaesthesia journal.

Rathi et al. surveyed 317 corresponding authors of RCTs published in 2010 to 2011 in six high-impact general medicine journals; 74% of authors stated that sharing of de-identified data via repositories should be required, whereas 72% answered that investigators should be directed to share their de-identified data after receiving an individual request (101). Among 47% of authors who indicated that they had received requests for sharing their trial data, 77% responded that they had indeed granted the data to those who requested them (101).

Analysis of data sharing predictors, using data from the same survey, showed no significant differences between trialists who did or did not support data sharing (102). In 2018, Polanin and Terzian reported results of an RCT that was conducted among authors of studies included in recent meta-analyses via a web-based survey (103). They found that participants who were randomly assigned to receive a data-sharing agreement were more willing to share individual patient data of their primary study (103); however, in all those studies, the trialists were asked a hypothetical question. The real-world data indicate that authors' words and their behaviour do not necessarily match. A high number of authors who indicated that they provided their data on their request may simply be socially acceptable answers.

In the first study of Rathi et al., a survey asked authors to comment on their concerns regarding data sharing. Significant concerns were related to the appropriate use of shared data. These authors were less concerned with the interests of investigators and funders, whereas protection of research participants was among their most minor concerns (101). The two most common reasons for refusing data sharing in our study were data ownership issues, as corresponding authors indicated that they do not own the data and issues of concerns for participants' privacy. Although we highlighted to the contacted authors that we were asking for de-identified data, some authors still cited privacy concerns as the reason for the refusal.

Because it is assumed that recruitment of participants into clinical trials is an altruistic act that will contribute to the advancement of medicine and medical knowledge, it is easy to see why many authors argue that failure to publish trials and lack of data sharing is considered a violation of the trust of trial participants. Spence et al. (104) recently analyzed individual consent forms (ICFs) to see whether trial participants were informed about the investigators' plans related to contributing to medical knowledge, publishing trial results, and sharing de-identified trial data. Their study showed that ICFs seldom provide trialists' intentions regarding sharing of de-identified data or trial publication, and 91% of the ICFs did not indicate information regarding ownership of the trial data (104). This finding is essential in light of arguments suggesting that by refusing to share data, the trialists are protecting the privacy of participants.

In our study, we did not have any experience with authors requesting us to cover expenses related to data sharing, as described by Naudet et al. (99). When they asked corresponding authors of selected trials to share data, one research team that authored two targeted reports requested a sum of £607 (equivalent to $857) as a condition for sharing data, but the study

authors refused to cover these expenses, as the other teams shared their data without any charge; and we did have one corresponding author in our cohort who mandated co-authorship in exchange for data.

Data sharing transparency could be achieved by involving relevant stakeholders that can influence the behaviour of authors, such as editors, organizations such as ICMJE, and research funders. Our study indicates that simply requiring a data sharing plan is not sufficient. Instituting mandatory data repositories and requesting higher accountability from corresponding authors are also potential practical solutions (99,105).

Our study limitations include a scant sample of journals, a narrow time frame analyzed, and potential nonresponder bias. It is possible that sending a data-sharing agreement along with our email could have yielded additional responses (103). We aimed to contact only corresponding authors and other individuals that were specifically suggested by corresponding authors; we did not attempt to contact all study authors. Furthermore, in our invitation to the corresponding authors, we emphasized that we are studying open data sharing in RCTs from the field of anesthesiology and that our team is interested in re-examining RCT raw data sets. It is possible that this type of request can be considered too general, and our results may not generalize to more targeted requests.

In conclusion, authors should be required to disclose their de-identified trial data publicly. Journal encouragement for data sharing is not enough to elicit willingness to share when approached. Whether the authors should be required to make their trial data available at the time of manuscript submission, or manuscript publication or sometime after publication can be debated. Left to their own devices, authors would likely rather refuse to share their data unless they are required to.

5.2 **Second study**

Despite indicating in the DAS that they are willing to share data on request, our second study is concerning as the overwhelming majority of the authors (82%) did not even answer our invitation to share the data. Only 6.8% of authors shared the data, of which the majority would allow re-analysis, and we received a rich portfolio of excuses for not sharing the data.

Some authors imposed certain conditions before sharing the data, including providing NDA or DTA, requesting co-authorship, reimbursement, registration on a specific web platform. We were not willing to offer co-authorship to those authors as we actually did not plan to re-use their data and publish such a new analysis; our study was methodological and had a different aim. Also, we did not have funding to cover any reimbursement requests. If we had fulfilled those requests, we can only presume that those authors would share their data with us.

We fulfilled requests for an NDA or a DTA that depended only on our institution or us. Whenever authors asked for such documents, we offered them the NDA that we had prepared *a priori*. Some of them requested us to sign the NDA that they had prepared. However, even after signing those documents and sending them to the authors, the majority still did not share their data.

A scoping review and a practical guide of Ventresca et al. regarding obtaining and managing data sets for individual participant data (IPD) meta-analysis (106) addressed incentives that could be offered to authors in exchange for data, which correspond to the conditions we encountered. Ventresca et al. acknowledge that individuals who made considerable effort should be recognized by offering co-authorship to the original study. Their approach was to offer co-authorship or acknowledgement to a corresponding author and individuals that the corresponding author considered worthy of authorship or an acknowledgement (106).

There are many suggestions in the literature for rewarding trialists who share data, including payment, publication, recognition by funding agencies and academic institutions for promotions. Likewise, penalties for not sharing data were also suggested, including fines or suspension of a product's market authorization (60,72,105,107–113). In addition, offering authorship to those who share data in the context of an IPD or another study where data would be used was proposed as a sensible approach as the authors of data thus receive an incentive, and they can control the data and the future manuscript before publication (110,111). However, in our study, we did not offer co-authorship to anyone because, firstly, we contacted hundreds of individuals, and secondly, we did not intend to re-analyse and re-publish their data in any form.

Furthermore, offering co-authorship for sharing data has been questioned as potentially unethical. For example, Devriendt et al. warn that although co-authorship in return for

providing data is expected, this might not be compatible with the internationally recognized authorship guidelines and that, furthermore, it raises concerns over the capability of secondary analysts to potentially contest the proposed research methods or conclusions that were initially drawn from the data (114).

In our study, we did not offer any reimbursement to the authors for their efforts related to data sharing. Ventresca et al. described that they tried an approach of offering reimbursement for minimal expenses related to data sharing, for example, shipping fees for data that the corresponding authors did not want to send electronically (106).

Veroniki et al. tried offering a small financial incentive of 100 Canadian dollars to the authors of trials eligible for an IPD meta-analysis, but this intervention did not improve IPD retrieval rates (115,116).

We acknowledge that the search for data and data preparation for sharing may be a significant burden to the researchers, which may not be possible without funding. However, as we did not have any funding for this study, we did not offer any reimbursement in advance in exchange for data, and we were unable to accommodate the one request for reimbursement that we received.

It has been reported that some of the available platforms for data sharing are very costly, based on the articles published in 2016 and 2018, ranging from 30,000 to 50,000 USD annually to list up to 20 studies on CSDR, and from 2,000 to 4,500 USD per listed study on Vivli (117). We tried to find out the prices they were charging in July 2021; however, these prices were not listed publicly for CSDR, and their representatives did not answer our question on this topic. Vivli, on the other hand, has a transparent charging policy. They charge 4000$ and 9500$ for long term hosting per academic study; the higher price includes an independent review panel for the data request proposals.

In our first study, we did not prepare an NDA or a DTA. In the second study, we prepared the NDA to be used if needed. Data sharing agreements describe the conditions that the research team requesting the data should respect in exchange for the data (118–120). Such agreements describe the aim of the study, analysis plan, data that is being exchanged, confidentiality issues, the timing of data sharing, data storage, security issues, sharing of data to third parties, intellectual property rights, plans for publication and authorship, etc. (106). Since we did not

plan to re-use and re-publish the data we requested in any form in our study, we prepared only NDA. The NDA is s an "agreement restricting the use of information by prohibiting a contracting party from divulging data." (121). We opted for the preparation of an NDA instead of a data-sharing agreement precisely because we did not plan to re-use and re-publish the received data in any way. By preparing an NDA, we wanted to assure the authors; we would keep their data confidential.

However, a few of the contacted authors requested us to sign an NDA as a precondition to share their data, and when we did sign it, the majority never responded back or shared their data.

Some authors wanted us to log into specific web platforms, with complicated procedures involved as a prerequisite for them to start considering our data request. However, we did not engage in those processes, as our prior experience and several other manuscripts indicate that such a decision process is often lengthy and ultimately with a negative outcome (65,98).

Several authors did not share their data with us with the explanation that they conducted qualitative studies. Due to fundamental differences in qualitative and quantitative studies, it has been reported that qualitative studies warrant specific considerations in the data-sharing movement (122). Unlike the set of numbers expected to be shared for quantitative studies, to enable statistical re-analysis, data in qualitative studies are usually collected via interviews, focus groups, direct observation and document review. These differences between quantitative and qualitative studies may have repercussions on the reproducibility of results. Reproducibility is defined as obtaining the consistent result by "using the same input data; computational steps, methods, and code; and conditions of analysis", thus, implying computational reproducibility (123).

This idea of computational verification, i.e. reproducibility, may not translate well to qualitative studies (122). For example, if a qualitative study was conducted via interviews, the question is what raw data, in that case, is – a recorded interview or a transcript, with or without field notes. It has been argued that interview transcripts, even when shared with detailed field notes, cannot adequately represent the interview and that it is questionable whether interview transcripts can genuinely be considered raw data (123). There are also hurdles with sharing data from interviews that may have been conducted in languages other than English (123). Indeed, in our study, some authors explained that they do not wish to

share their data from qualitative studies because interviews were conducted in Ukrainian and Russian languages. However, the language barrier should not be *a priori* reason for not sharing such data because individuals requesting data could have the necessary language proficiency or resources to secure translation.

It is anticipated that re-analysis of qualitative data should yield many of the significant themes identified initially, which may lead to questions about accurate reproducibility. However, for the sake of transparency and considering all hurdles associated with the re-analysis of data from qualitative studies, policies for sharing raw data from qualitative research can benefit the open data movement.

The availability of raw data, even on request, is considered as a safeguard of good research practices. However, the question is how to ensure that raw data are indeed shared. An editor of the journal Molecular Brain published an editorial in 2020 describing his effort to request raw data from manuscripts. Since 2017, he has requested raw data from 41 manuscripts. To his surprise, the authors of 21 (51%) of those 41 manuscripts decided to withdraw their manuscript without providing raw data. The editor rejected 19 out of 20 remaining manuscripts because of insufficient raw data. Thus, the editor hypothesized that either raw data did not exist at all, or at least in some portions. The editor concludes that journals should request raw data from the authors in order to verify possible data fabrication, increase research results' reproducibility and strengthen public trust in science (124).

It is possible that some of the authors that we have contacted did not respond to our request because they do not have raw data or because of problems with their raw data. Lack of transparent practices drives suspicion. Our studies show that we cannot rely on authors to provide raw data from published articles, even if they use DAS that promises that data will be available on request. Having a DAS is not enough. Editors as gatekeepers should start requesting raw data as the obligatory part of manuscript submission; this could likely be the only way to secure the accessibility and verifiability of raw data from published studies.

A limitation of the second study could be its inherent aim – our study was methodological, and we were interested in whether, in principle, the authors would respond to our data request, share their data and whether the data would be re-analysable. We did not intend to conduct a re-analysis of the data or to do follow-up studies on the raw data. It is possible that the corresponding authors could respond differently if our study aim was related to their data.

There is much emphasis on data sharing and open data currently in the research community. However, we have shown that there is a discrepancy between what authors say (write) and what authors do. Previous studies indicate that many authors express support for ideas of data sharing and open data; however, when it comes to sharing data, the authors may not behave in line with what they say. Our first study, partly published in the Journal of Clinical Epidemiology (98), has shown that only 3.6% of contacted trialists shared their raw data. Bergeris et al. have shown previously that only 5% of researchers conducting clinical trials were willing to share their individual patient data (IPD) and that trialists are not aware of the true meaning of IPD data sharing (125).

Clinical trials are used for creating clinical guidelines, which are the hallmark of contemporary patient care. Thus, it is unfortunate that clinical trial data cannot be accessed, verified and re-analyzed. A recent study by Ebrahim et al. explored re-analyses of clinical trial data and showed that 35% of those re-analyses could not confirm data reported in the original publication about a clinical trial (126). This highlights the importance of clinical trial data availability.

Currently, there are no mandatory requirements related to data sharing. However, data sharing statements are being widely adopted. ICMJE postulates that reports of clinical trials published from July 2018 onward need to provide a mandatory data sharing plan – but this applies only to clinical trials, not to other types of studies (95). Nevertheless, there is no guarantee that a data sharing plan will lead to actual data sharing.

Some publishers, such as Springer Nature, require authors of all types of studies to include a DAS in their manuscript. These initiatives are better, as they are not limited to a particular study design. However, we have shown that even though the majority of DAS that the authors use in BMC Springer Nature publications indicated that the data would be disclosed upon request, most of those authors were actually not willing to share their data.

The main scientific contribution of our study is our finding that authors cannot be trusted that they will indeed share their data, even if they wrote in their manuscript that they would do so. Thus, our findings can enable the creation of new guidelines and practices in the research community that would guide the availability of raw research data.

6 ABSTRACT

Background: This thesis consists of two cross-sectional studies that analyzed data sharing practices among biomedical researchers. The first study analysed the authors' willingness to share raw data from their randomized controlled trials (RCTs) in anesthesiology. The second study analysed researchers' compliance with their Data Availability Statement (DAS) from manuscripts published in open access journals with the mandatory DAS.

Methods: The first study included RCTs of anesthesiology interventions published from January 1, 2014, to December 31, 2016, in seven journals from the Journal Citation Reports category Anesthesiology belonging to Q1, that is, the highest-ranking 25% of journals in this category, according to the Clarivate Analytics' Journal Impact Factor distribution. We analyzed full-text articles and included only RCT's, and there were 619 manuscripts published in those journals in the selected time period. A de-identified email with a request for raw data was sent to corresponding authors. After accessing raw data sets, we checked whether data were available in a way that enabled reanalysis, i.e., published in a file that enabled data use and whether sufficient metadata (description of data and variables that would permit re-use) were included.

The second study included manuscripts published during January 2019 in all open-access journals published by BioMed Central (BMC; part of Nature Springer). We analysed 3416 articles with DAS from 282 journals. We categorized types of DAS. We surveyed corresponding authors who wrote in their DAS that they would share the data on request or under certain circumstances. After accessing raw data sets, we checked whether data were available in a way that enabled reanalysis, i.e., published in a file that enabled data use and whether relevant metadata were included.

Results: The first study showed that out of 86 (13.89%) authors from 619 pooled who responded to our query for data sharing, only 24 (3.87%) provided the requested data. Only 24 (3.87%) of manuscripts contained DAS suggesting a willingness to share trial data; only one of those authors actually shared data. Statistically significant difference was found in the proportion of data sharing between studies with commercial and non-profit funding. Most of the contacted authors did not respond to our query at all; among the 62 authors who rejected to provide data, reasons were rarely provided. When reasons were provided, arguments

included issues regarding data ownership and patient privacy. Only one of the seven analyzed journals stimulated authors toward data sharing.

The second study showed that among 254 (14.17%) out of 1792 authors (52.46%) of 3419 DAS papers, who responded to our query for data sharing, only 123 (6.8%)* provided the requested data. Among 1792 manuscripts in which DAS indicated that authors are willing to share their data, 1669 (93.19%) authors either did not respond or refused to share their data with us, i.e., did not comply with their statement.

Conclusion: Willingness to share data among the authors of research articles is very low. To achieve widespread availability of de-identified research data, editors should request their publication instead of only encouraging authors to do so or to provide a DAS. Even when authors state in their manuscript that they will share data on request, they have the same compliance rate as those authors that do not have a DAS in their manuscripts, indicating that DAS has little relevance for ensuring data sharing.

* 123 = 6.8% of 1792 contacted, 48.03% of 254 responding

7 SAŽETAK

Praksa dijeljenja istraživačkih podataka među autorima biomedicinskih publikacija

Uvod. Ovaj se doktorski rad sastoji od dva presječna istraživanja čiji je cilj bio analizirati spremnost za dijeljenje podataka među biomedicinskim istraživačima. Prvo istraživanje analiziralo je spremnost autora za dijeljenje neobrađenih podataka iz randomiziranih kontroliranih pokusa (engl. *randomized controlled trial*, RCT) iz područja anesteziologije. Drugo istraživanje analiziralo je usklađenost istraživača s izjavom o pristupu podacima (DAS) iz svih časopisa s otvorenim pristupom objavljenim u Springer Nature BioMed Central.

Metode. Prvo istraživanje obuhvaćalo je RCT-ove o intervencijama objavljene od 1. siječnja 2014. do 31. prosinca 2016. u sedam časopisa iz kategorije Anesteziologija, Journal Citation Reports koji pripadaju Q1 rangu, prema distribuciji čimbenika utjecaja časopisa. Analizirali smo 619 RCT-ova. Autorima je poslan upit e-poštom sa zahtjevom za neobrađenim podacima. Nakon pristupa neobrađenim skupovima podataka provjerili smo jesu li podaci dostupni na način koji omogućava ponovnu analizu, odnosno, dijele li se u datoteci koja omogućuje upotrebu podataka i jesu li uključeni metapodaci (opis podataka i varijabli koji bi omogućili ponovnu upotrebu).

Drugo istraživanje obuhvatilo je rukopise objavljene tijekom siječnja 2019. u svim časopisima s otvorenim pristupom u izdanju BioMed Central (BMC; dio Nature Springer). Analizirali smo 3416 članaka s DAS-om iz 282 časopisa. Kategorizirali smo vrste DAS-a. Kontaktirali smo dopisne autore koji su u svom DAS-u napisali da će podatke dijeliti na zahtjev ili pod određenim okolnostima. Nakon pristupa neobrađenim skupovima podataka provjerili smo jesu li podaci dostupni na način koji omogućuje ponovnu analizu, tj. objavljeni u formatu koji omogućuje upotrebu podataka i jesu li uz neobrađene podatke uključeni relevantni metapodaci.

Rezultati. U prvom istraživanju 86 (13.89%) od 619 autora odgovorilo je na upit za dijeljenje podataka, a 24 (3.87%) ih je poslalo tražene podatke. Samo 24 (3.87%) rada sadržavalo je izjave koje sugeriraju spremnost na razmjenu podataka s pokusa; samo je jedan od tih autora podijelio podatke. Značajna statistička razlika je uočena u udjelu razmjene podataka između istraživanja s komercijalnim i neprofitnim financiranjem. Većina autora na naš upit nije uopće

odgovorila, a 62 autora koji su odbili dati podatke rijetko su naveli razloge. Kad su navedeni razlozi, uobičajene teme uključivale su probleme u vezi s vlasništvom podataka i privatnošću sudionika. Samo jedan od sedam analiziranih časopisa u uputama potiče autore na dijeljenje podataka.

Drugo istraživanje pokazalo je da od 254 (14.17%) između 1792 autora koji su odgovorili na naš upit za dijeljenje podataka, samo 123 (6.8%) pružilo tražene podatke. Od 1792 članka gdje je u DAS-u bila navedena spremnost za dijeljenje podataka, 1669 (93.19%) autora nisu odgovorili na upit ili su odbili podijeliti podatke, odnosno nisu se pridržavali vlastite izjave.

Zaključak. Spremnost za dijeljenje podataka među autorima znanstvenih članaka vrlo je skromna. Kako bi se omogućila dostupnost neobrađenih podataka prikupljenih tijekom istraživanja, časopisi bi trebali zatražiti njihovo objavljivanje zajedno s člankom, a ne samo poticati autore na dijeljenje podataka i pisanje izjava o dijeljenju podataka u radu. Čak i kada autori u svom članku izjave da će dijeliti podatke na zahtjev, imaju istu stopu dijeljenja podataka s onima koji nemaju DAS u svojim rukopisima, što ukazuje da DAS nije veoma relevantan za dijeljenje podataka.

8 REFERENCES

1. European Commission. Towards a thriving data-driven economy. European Commission. 2014.

2. Izdebski K. Transparency and open data principles: Why they are important and how they increase public participation and tackle corruption. U. S. Department of State. 2015.

3. Boulton G. The open data imperative. Insights. 2014;27(2):133–8.

4. Greenaway F. Science international: A history of the international council of scientific unions. Cambridge University Press; 1996. p. 33-5.

5. Our milestones [Internet]. U.S. National library of medicine. 2011 [cited 2021 Jan 12]. Available from: https://wayback.archive-it.org/org-350/20170628152022/https://apps.nlm.nih.gov/175/milestones.cfm

6. National Research Council. On the full and open exchange of scientific data. Washington, DC: National Academies Press; 1995. p. 27-9.

7. Merton RK. The sociology of science, theoretical and empirical investigations. Chicago: University of Chicago Press; 1973. p. 267-78.

8. Wilkinson MD, Dumontier M, Aalbersberg IJJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci data. 2016;3(1):160018.

9. Ayris P, Bernal I, Cavalli V, Dorch B, Frey J, Hallik M, et al. LIBER open science roadmap. 2018;4–16.

10. Krahe MA, Wolski M, Mickan S, Toohey J, Scuffham P, Reilly S. Developing a strategy to improve data sharing in health research: A mixed-methods study to identify barriers and facilitators. Heal Inf Manag J. 2020;

11. Tennison J. Data sharing is not open data [Internet]. governmentcomputing.com. 2014 [cited 2021 Jan 12]. Available from: https://www.governmentcomputing.com/central-

government/features/featuredata-sharing-is-not-open-data-4198712

12. History of IMO | World Meteorological Organization [Internet]. [cited 2021 Jan 7]. Available from: https://public.wmo.int/en/about-us/who-we-are/history-IMO

13. Kalkman S, Mostert M, Udo-Beauvisage N, Van Delden JJ, Van Thiel GJ. Responsible data sharing in a big data-driven translational research platform: lessons learned. BMC Med Inform Decis Mak. 2019;19(1):283.

14. Rodgers W, Nolte M. Solving problems of disclosure risk in an academic setting: using a combination of restricted data and restricted access methods. J Empir Res Hum Res Ethics. 2006;1(3):85–97.

15. GISAID - History [Internet]. [cited 2021 Jan 7]. Available from: https://www.gisaid.org/about-us/history/

16. Jones L, Cooke E, Hertzel D, Ormerod D, Patterson F, Lorimer E. Data sharing between public bodies. 2013;25–48.

17. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. Towards a data sharing culture: recommendations for leadership from academic health centers. PLoS Med. 2008;5(9):e183.

18. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science (80-). 2001;291(5507):1304–51.

19. Wojick D. By Birdie: A new national academy report on data sharing [Internet]. The Scholarly Kitchen. 2012 [cited 2021 Jan 7]. Available from: https://scholarlykitchen.sspnet.org/2012/12/19/by-birdie-a-new-national-academy-report-on-data-sharing/

20. Shaheen NA, Manezhi B, Thomas A, Alkelya M. Reducing defects in the datasets of clinical research studies: Conformance with data quality metrics. BMC Med Res Methodol. 2019;19(1):98.

21. Williams S. Data sharing may lead to some embarrassment but will ultimately improve scientific transparency and accuracy. [Internet]. The London school of economics and

political science. 2014 [cited 2021 Jan 7]. Available from: https://blogs.lse.ac.uk/impactofsocialsciences/2014/05/29/data-sharing-exciting-but-scary/

22. Leshner AI. Accountability and transparency. Science. 2009;324(5925):313.

23. Thursby M, Thursby J, Haeussler C, Lin J. Do academic scientists share information with their colleagues? Not necessarily [Internet]. Voxeu.org. 2009 [cited 2021 Jan 11]. Available from: https://voxeu.org/article/why-don-t-academic-scientists-share-information-their-colleagues

24. Wallis JC, Rolando E, Borgman CL. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One. 2013;8(7):67332.

25. Hanson B, Sugden A, Alberts B. Making data maximally available. Science (80-). 2011;331(6018):649–51.

26. Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, et al. Retracted: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. Lancet. 1998;351(9103):637–41.

27. Macchiarini scandal is a valuable lesson for the Karolinska Institute. Nature. 2016;537:137.

28. Marcus A, Oransky I. How Yoshitaka Fujii, the biggest fabricator in science, got caught [Internet]. Nautil.us. 2015 [cited 2021 Jan 11]. Available from: http://nautil.us/issue/24/error/how-the-biggest-fabricator-in-science-got-caught

29. Fraud case stuns anesthesiologists [Internet]. Anesthesiology News. 2009 [cited 2021 Jan 11]. Available from: https://www.anesthesiologynews.com/PRN-/Article/04-09/Fraud-Case-Stuns-Anesthesiologists/12868

30. NIH's definition of a clinical trial [Internet]. NIH grants and funding. 2017 [cited 2021 Jan 11]. Available from: https://grants.nih.gov/policy/clinical-trials/definition.htm

31. Clinical trials [Internet]. World Health Organisation. [cited 2021 Jan 11]. Available

from: https://www.who.int/health-topics/clinical-trials/#tab=tab_1

32. Prindle WD. Holy Bible: 21st Century King James Version. In: The book of Daniel, Ch 1. Gary, South Dakota: Deuel Enterprises; 1994.

33. Esterman A. The fascinating history of clinical trials [Internet]. The Conversation. 2020 [cited 2021 Jan 11]. Available from: https://theconversation.com/the-fascinating-history-of-clinical-trials-139666

34. Drucker CB. Ambroise Paré and the birth of the gentle art of surgery. Yale J Biol Med. 2008;81(4):199–202.

35. Lind J. The nature of the symptoms, explained and deduced from the foregoing theory and dissections. In: A Treatise of the Scurvy, in Three Parts. Cambridge: Cambridge University Press; 2014. p. 317–40.

36. Hull G. Caleb Hillier Parry 1755-1822: a notable provincial physician. J R Soc Med. 1998;91:335–8.

37. Flint A. A contribution toward the natural history of articular rheumatism; consisting of a report of thirteen cases treated solely with paliative measures. Am J Med Sci. 1863;46:17–36.

38. Fletcher W. Rice and beri-beri: preliminary report on an experiment conducted in the Kuala Lumpur insane asylum. Lancet. 1894;1:1776–9.

39. Hill AB. Memories of the British streptomycin trial in tuberculosis. The first randomized clinical trial. Control Clin Trials. 1990;11(2):77–9.

40. Sackett DL, Rosenberg WMC, Gray JAM, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn’t. 1996. BMJ. 1996;312:71–2.

41. Minkow D. The evidence based medicine pyramid! [Internet]. Students 4 Best Evidence. 2014 [cited 2021 Jan 13]. Available from: https://s4be.cochrane.org/blog/2014/04/29/the-evidence-based-medicine-pyramid/

42. Gueyffier F, Cucherat M. The limitations of observation studies for decision making

regarding drugs efficacy and safety. Therapie. 2019;74(2):181–5.

43. El-Gilany A-H. What is case series? Asploro J Biomed Clin Case Reports. 2018;1(1):10–5.

44. Dunn PM. Perinatal lessons from the past: Sir Norman Gregg, ChM, MC, of Sydney (1892-1966) and rubella embryopathy. Arch Dis Child Fetal Neonatal Ed. 2007;92(6):F513–4.

45. Pneumocystis Pneumonia - Los Angeles. Morb Mortal Wkly Rep. 1981;30(21):250–2.

46. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual - New York City and California. Morb Mortal Wkly Rep. 1981;30(25):305–8.

47. Thomas J. The most famous case report ever [Internet]. mcdreeamiemusings.com. 2019 [cited 2021 Jan 13]. Available from: https://mcdreeamiemusings.com/blog/2019/2/15/the-most-famous-case-report-ever

48. Paneth N, Susser E, Susser M. Origins and early development of the case-control study: Part 1, Early evolution. Soz Praventivmed. 2002;47(5):282–8.

49. Lane-Claypon JE. A further report on cancer of the breast, with special reference to its associated antecedent conditions. Reports Public Heal Med Subj No 32 Minist Heal London Publ by His Majesty's Station Off. 1926;1–135.

50. The Roman imperial legion | UNRV.com [Internet]. [cited 2021 Jan 13]. Available from: https://www.unrv.com/military/legion.php

51. Song JW, Chung KC. Observational studies: Cohort and case-control studies. Plast Reconstr Surg. 2010;126(6):2234–42.

52. Barrett D, Noble H. What are cohort studies? Evid Based Nurs. 2019;22(4):95–6.

53. Desai VS, Camp CL, Krych AJ. What Is the hierarchy of clinical evidence? Basic methods handbook for clinical orthopaedic research. Berlin, Heidelberg: Springer; 2019. p. 11-22.

54. Hill AB, Doll R. Smoking and carcinoma of the lung preliminary report. Br Med J.

1950;2(4682):739–48.

55. Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: the Framingham Study. Am J Public Health. 1951;41(3):279–81.

56. Viera AJ, Bangdiwala SI. Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. Fam Med. 2007;39(2):132–7.

57. Murad MH, Asi N, Alsawas M, Alahdab F. New evidence pyramid. Evid Based Med. 2016;21(4):125–7.

58. Chalmers TC, Smith H, Blackburn B, Silverman B, Schroeder B, Reitman D, et al. A method for assessing the quality of a randomized control trial. Control Clin Trials. 1981;2(1):31–49.

59. Koenig F, Slattery J, Groves T, Lang T, Benjamini Y, Day S, et al. Sharing clinical trial data on patient level: opportunities and challenges. Biom J. 2015;57(1):8–26.

60. Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, et al. Increasing value and reducing waste: addressing inaccessible research. Lancet. 2014;383(9913):257–66.

61. Dallmeier-Tiessen S, Darby R, Gitmans K, Lambert S, Matthews B, Mele S, et al. Enabling sharing and reuse of scientific data. New Rev Inf Netw. 2014;19(1):16–43.

62. Selph SS, Ginsburg AD, Chou R. Impact of contacting study authors to obtain additional data for systematic reviews: diagnostic accuracy studies for hepatic fibrosis. Syst Rev. 2014;3(1):107.

63. Doshi P, Jefferson T, del Mar C. The imperative to share clinical study reports: Recommendations from the Tamiflu experience. PLoS Med. 2012;9(4):e1001201.

64. Doshi P, Jefferson T. Open data 5 years on: a case series of 12 freedom of information requests for regulatory data to the European Medicines Agency. Trials. 2016;17(1):78.

65. Puljak L, Marin A, Vrdoljak D, Markotic F, Utrobicic A, Tugwell P. Celecoxib for osteoarthritis. Cochrane Database Syst Rev. 2017;5:CD009865.

66. Ross JS, Lehman R, Gross CP. The importance of clinical trial data sharing: toward more open science. Circ Cardiovasc Qual Outcomes. 2012;5(2):238–40.

67. Tenopir C, Allard S, Douglass K, Aydinoglu AU, Wu L, Read E, et al. Data sharing by scientists: practices and perceptions. PLoS One. 2011;6(6):e21101.

68. Hrynaszkiewicz I, Norton ML, Vickers AJ, Altman DG. Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials. 2010;11(1):9.

69. ICMJE. Recommendations for the conduct, reporting, editing, and publication of scholarly work in medical journals [Internet]. 2019. p. 1–19. Available from: http://www.icmje.org/icmje-recommendations.pdf

70. IOM. Sharing clinical trial data: maximizing benefits, minimizing risk. Washington, DC: The National Academies Press; 2015.

71. World Medical Association. World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA. 2013;310(20):2191–4.

72. Ohmann C, Banzi R, Canham S, Battaglia S, Matei M, Ariyo C, et al. Sharing and reuse of individual participant data from clinical trials: principles and recommendations. BMJ Open. 2017;7(12):e018647.

73. Krleža-Jerić K, Gabelica M, Banzi R, Krnić-Martinić M, Pulido B, Mahmić-Kaknjo M, et al. IMPACT observatory: Tracking the evolution of clinical trial data sharing and research integrity. Biochem Medica. 2016;26(3):308–17.

74. Taichman DB, Backus J, Baethge C, Bauchner H, de Leeuw PW, Drazen JM, et al. Sharing clinical trial data: proposal from the ICMJE. JAMA. 2016;315(5):384–6.

75. Thorogood A, Knoppers BM. Can research ethics committees enable clinical trial data sharing? Ethics, Med Public Heal. 2017;3(1):56–63.

76. Banzi R, Canham S, Kuchinke W, Krleza-Jeric K, Demotes-Mainard J, Ohmann C. Evaluation of repositories for sharing individual-participant data from clinical studies.

Trials. 2019;20(1):169.

77. Ross JS, Waldstreicher J, Bamford S, Berlin JA, Childers K, Desai NR, et al. Overview and experience of the YODA project with clinical trial data sharing after 5 years. Sci data. 2018;5(1):180268.

78. Navar AM, Pencina MJ, Rymer JA, Louzao DM, Peterson ED. Use of open access platforms for clinical trial data. JAMA. 2016;315(12):1283–4.

79. Le Noury J, Nardo JM, Healy D, Jureidini J, Raven M, Tufanaru C, et al. Restoring Study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. BMJ. 2015;351:h4320.

80. Keller MB, Ryan ND, Strober M, Klein RG, Kutcher SP, Birmaher B, et al. Efficacy of paroxetine in the treatment of adolescent major depression: a randomized, controlled trial. J Am Acad Child Adolesc Psychiatry. 2001;40(7):762–72.

81. Vaduganathan M, Nagarur A, Qamar A, Patel RB, Navar AM, Peterson ED, et al. Availability and use of shared data from cardiometabolic clinical trials. Circulation. 2018;137(9):938–47.

82. Miller J, Ross JS, Wilenzick M, Mello MM. Sharing of clinical trial data and results reporting practices among large pharmaceutical companies: cross sectional descriptive study and pilot of a tool to improve company practices. BMJ. 2019;366:l4217.

83. MHRA. Investigation into Glaxosmithkline/Seroxat [Internet]. 2008 [cited 2021 Feb 16]. p. 1–13. Available from: https://webarchive.nationalarchives.gov.uk/20141206221413/http://www.mhra.gov.uk/home/groups/es-policy/documents/websiteresources/con014155.pdf

84. Weng C, Friedman C, Rommel CA, Hurdle JF. A two-site survey of medical center personnel's willingness to share clinical data for research: implications for reproducible health NLP research. BMC Med Inform Decis Mak. 2019;19(Suppl 3):70.

85. Savage CJ, Vickers AJ. Empirical study of data sharing by authors publishing in PLoS journals. PLoS One. 2009;4(9):e7078.

86. Weitzman ER, Kelemen S, Kaci L, Mandl KD. Willingness to share personal health record data for care improvement and public health: a survey of experienced personal health record users. BMC Med Inform Decis Mak. 2012;12(1):39.

87. Weitzman ER, Kaci L, Mandl KD. Sharing medical data for health research: the early personal health record experience. J Med Internet Res. 2010;12(2):e14.

88. Stuart D, Baynes G, Hrynaszkiewicz I, Allin K, Penny D, Lucraft M, et al. Whitepaper: Practical challenges for researchers in data sharing [Internet]. SpringerNature.com. Springer Nature; 2018 [cited 2021 Feb 16]. Available from: /articles/journal_contribution/Whitepaper_Practical_challenges_for_researchers_in_data_sharing/5975011/1

89. Springer Nature. Data Availability Statements [Internet]. [cited 2021 Feb 24]. Available from: https://www.springernature.com/gp/authors/research-data-policy/data-availability-statements/12330880

90. What is a data availability statement? [Internet]. [cited 2021 Feb 24]. Available from: https://authorservices.taylorandfrancis.com/data-sharing-policies/data-availability-statements/#

91. PLoS One. Data Availability [Internet]. [cited 2021 Feb 24]. Available from: https://journals.plos.org/plosone/s/data-availability

92. Silva L, Bloom T, Ganley E, Winker M. PLoS' new data policy: public access to data - EveryONE [Internet]. 2014 [cited 2021 Feb 16]. Available from: https://everyone.plos.org/2014/02/24/plos-new-data-policy-public-access-data-2/

93. Federer LM, Belter CW, Joubert DJ, Livinski A, Lu Y-L, Snyders LN, et al. Data sharing in PLoS ONE: An analysis of Data Availability Statements. PLoS One. 2018;13(5):e0194768.

94. CHORUS. Publisher data availability policies index [Internet]. [cited 2021 Feb 24]. Available from: https://www.chorusaccess.org/resources/chorus-for-publishers/publisher-data-availability-policies-index/

95. ICMJE recommendations. Clinical trials. 2. Data sharing [Internet]. [cited 2021 Sep

26]. Available from: http://www.icmje.org/recommendations/browse/publishing-and-editorial-issues/clinical-trial-registration.html

96. Huh S. Recent trends in medical journals' data sharing policies and statements of data availability. Arch Plast Surg. 2019;46(6):493–7.

97. Enago Academy. How important are data availability statements (DAS)? [Internet]. 2019 [cited 2021 Feb 24]. Available from: https://www.enago.com/academy/how-important-are-data-availability-statements-das/

98. Gabelica M, Cavar J, Puljak L. Authors of trials from high-ranking anesthesiology journals were not willing to share raw data. J Clin Epidemiol. 2019;109:111–6.

99. Naudet F, Sakarovitch C, Janiaud P, Cristea I, Fanelli D, Moher D, et al. Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine. BMJ. 2018;360:k400.

100. Taichman DB, Sahni P, Pinborg A, Peiperl L, Laine C, James A, et al. Data sharing statements for clinical trials. BMJ. 2017;357(1):j2372.

101. Rathi V, Dzara K, Gross CP, Hrynaszkiewicz I, Joffe S, Krumholz HM, et al. Sharing of clinical trial data among trialists: a cross sectional survey. BMJ. 2012;345:e7570.

102. Rathi VK, Strait KM, Gross CP, Hrynaszkiewicz I, Joffe S, Krumholz HM, et al. Predictors of clinical trial data sharing: exploratory analysis of a cross-sectional survey. Trials. 2014;15(1):384.

103. Polanin JR, Terzian M. A data-sharing agreement helps to increase researchers' willingness to share primary data: results from a randomized controlled trial. J Clin Epidemiol. 2019;106:60–9.

104. Spence OM, Onwuchekwa Uba R, Shin S, Doshi P. Patient consent to publication and data sharing in industry and NIH-funded clinical trials. Trials. 2018;19(1):1–5.

105. Wolfe N, Gøtzsche PC, Bero L. Strategies for obtaining unpublished drug trial data: a qualitative interview study. Syst Rev. 2013;2(1):31.

106. Ventresca M, Schünemann HJ, Macbeth F, Clarke M, Thabane L, Griffiths G, et al. Obtaining and managing data sets for individual participant data meta-analysis: scoping review and practical guide. BMC Med Res Methodol. 2020;20(1):113.

107. Doshi P, Goodman SN, Ioannidis JPA. Raw data from clinical trials: within reach? Trends Pharmacol Sci. 2013;34(12):645–7.

108. Gøtzsche PC. Why we need easy access to all data from all clinical trials and how to accomplish it. Trials. 2011;12(1):249.

109. Ross JS, Krumholz HM. Ushering in a new era of open science through data sharing: the wall must come down. JAMA. 2013;309(13):1355–6.

110. Fecher B, Friesike S, Hebing M. What drives academic data sharing? PLoS One. 2015;10(2):e0118053.

111. Perrino T, Howe G, Sperling A, Beardslee W, Sandler I, Shern D, et al. Advancing science through collaborative data sharing and synthesis. Perspect Psychol Sci. 2013;8(4):433–44.

112. Bierer BE, Crosas M, Pierce HH. Data authorship as an incentive to data sharing. N Engl J Med. 2017;376(17):1684–7.

113. Sydes MR, Ashby D. Data authorship as an incentive to data sharing. N Engl J Med. 2017;377(4):402.

114. Devriendt T, Shabani M, Borry P. Data sharing in biomedical sciences: A systematic review of incentives. Biopreserv Biobank. 2021;19(3):219–27.

115. Veroniki AA, Straus SE, Ashoor H, Stewart LA, Clarke M, Tricco AC. Contacting authors to retrieve individual patient data: study protocol for a randomized controlled trial. Trials. 2016;17(1):138.

116. Veroniki AA, Ashoor HM, Le SPC, Rios P, Stewart LA, Clarke M, et al. Retrieval of individual patient data depended on study characteristics: a randomized controlled trial. J Clin Epidemiol. 2019;113:176–88.

117. Rockhold F, Nisen P, Freeman A. Data sharing at a crossroads. N Engl J Med. 2016;375(12):1115–7.

118. Lo B. Sharing clinical trial data. JAMA. 2015;313(8):793–4.

119. Tucker K, Branson J, Dilleen M, Hollis S, Loughlin P, Nixon MJ, et al. Protecting patient privacy when sharing patient-level data from clinical trials. BMC Med Res Methodol. 2016;16(S1):77.

120. Tudur Smith C, Hopkins C, Sydes MR, Woolfall K, Clarke M, Murray G, et al. How should individual participant data (IPD) from publicly funded clinical trials be shared? BMC Med. 2015;13(1):298.

121. Beyer GW, Redden KR, Beyer MM. Modern dictionary for the legal profession. Vol. 31. W.S. Hein & Co; 1993. p. 31-1297.

122. Tsai AC, Kohrt BA, Matthews LT, Betancourt TS, Lee JK, Papachristos A V., et al. Promises and pitfalls of data sharing in qualitative research. Soc Sci Med. 2016;169:191–8.

123. National Academies of Sciences Engineering and Medicine. Reproducibility and replicability in science. Washington, D.C.: National Academies Press; 2019. p. 39-54.

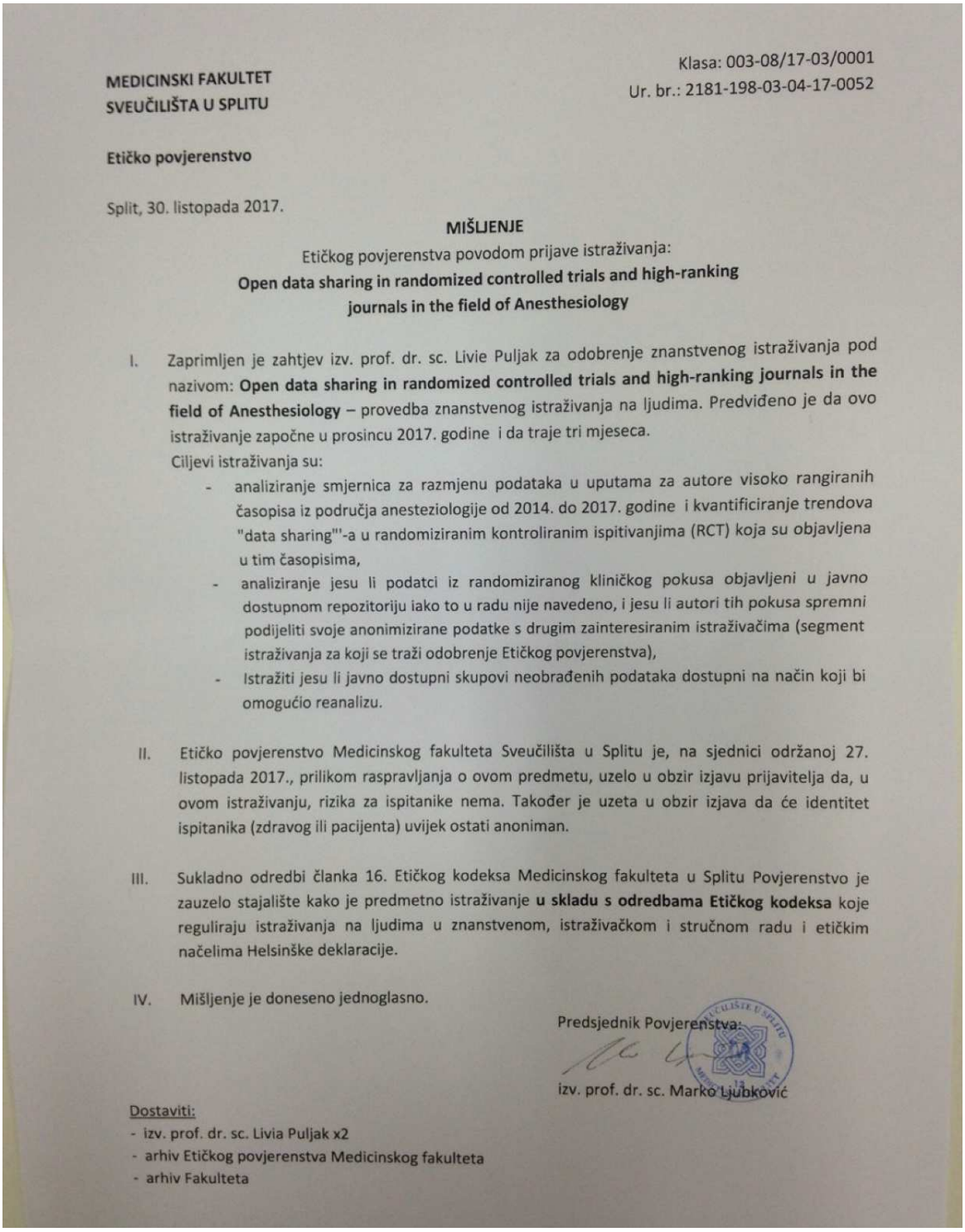
124. Miyakawa T. No raw data, no science: another possible source of the reproducibility crisis. Mol Brain. 2020;13(1):24.

125. Bergeris A, Tse T, Zarin DA. Trialists' intent to share individual participant data as disclosed at ClinicalTrials.gov. JAMA. 2018;319(4):406–8.

126. Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ, et al. Reanalyses of randomized clinical trial data. JAMA. 2014;312(10):1024–32.

9 APPENDICES

9.1 Appendix 1. Scanned approval of the study protocol by the Ethics Committee of the University of Split School of Medicine for Study 1



MEDICINSKI FAKULTET
SVEUČILIŠTA U SPLITU

Klasa: 003-08/17-03/0001
Ur. br.: 2181-198-03-04-17-0052

Etičko povjerenstvo

Split, 30. listopada 2017.

MIŠLJENJE

Etičkog povjerenstva povodom prijave istraživanja:
Open data sharing in randomized controlled trials and high-ranking journals in the field of Anesthesiology

I. Zaprimljen je zahtjev izv. prof. dr. sc. Livie Puljak za odobrenje znanstvenog istraživanja pod nazivom: **Open data sharing in randomized controlled trials and high-ranking journals in the field of Anesthesiology** – provedba znanstvenog istraživanja na ljudima. Predviđeno je da ovo istraživanje započne u prosincu 2017. godine i da traje tri mjeseca.
Ciljevi istraživanja su:
- analiziranje smjernica za razmjenu podataka u uputama za autore visoko rangiranih časopisa iz područja anesteziologije od 2014. do 2017. godine i kvantificiranje trendova "data sharing"'-a u randomiziranim kontroliranim ispitivanjima (RCT) koja su objavljena u tim časopisima,
- analiziranje jesu li podatci iz randomiziranog kliničkog pokusa objavljeni u javno dostupnom repozitoriju iako to u radu nije navedeno, i jesu li autori tih pokusa spremni podijeliti svoje anonimizirane podatke s drugim zainteresiranim istraživačima (segment istraživanja za koji se traži odobrenje Etičkog povjerenstva),
- Istražiti jesu li javno dostupni skupovi neobrađenih podataka dostupni na način koji bi omogućio reanalizu.

II. Etičko povjerenstvo Medicinskog fakulteta Sveučilišta u Splitu je, na sjednici održanoj 27. listopada 2017., prilikom raspravljanja o ovom predmetu, uzelo u obzir izjavu prijavitelja da, u ovom istraživanju, rizika za ispitanike nema. Također je uzeta u obzir izjava da će identitet ispitanika (zdravog ili pacijenta) uvijek ostati anoniman.

III. Sukladno odredbi članka 16. Etičkog kodeksa Medicinskog fakulteta u Splitu Povjerenstvo je zauzelo stajalište kako je predmetno istraživanje **u skladu s odredbama Etičkog kodeksa** koje reguliraju istraživanja na ljudima u znanstvenom, istraživačkom i stručnom radu i etičkim načelima Helsinške deklaracije.

IV. Mišljenje je doneseno jednoglasno.

Predsjednik Povjerenstva:

izv. prof. dr. sc. Marko Ljubković

Dostaviti:
- izv. prof. dr. sc. Livia Puljak x2
- arhiv Etičkog povjerenstva Medicinskog fakulteta
- arhiv Fakulteta

9.2 Appendix 2. Personalized emails to each potential participant in Study 1

Dear **<title and name of the corresponding author**

My name is Mirko Gabelica, MD, and I am a researcher affiliated with the University of Split School of Medicine in Split, Croatia. I am conducting a study entitled Open data sharing in randomized controlled trials in the field of Anesthesiology. Your manuscript **<manuscript title>** was selected for this study.

Could I please receive a copy of a raw data set from that trial, or alternatively, can you direct me to a repository where the raw data set is publicly available? Our team is interested in re-examining RCT raw data sets. This analysis will be de-identified, and in the presentation of our results, we will not disclose any details about individual RCT and author characteristics. All data and communication will be treated with strict confidence. Only summary results will be presented. All raw data sets will be deleted after being examined.

We hope that our findings will contribute to the open data sharing movement in biomedical, clinical trials.

Our study was approved by the University of Split School of Medicine Ethics Committee (approval number, Klasa: 003-08/17-03/0001, Ur.br: 2181-198-03-04-17-0052). If you have any questions or concerns that you would like to address with the principal investigator of this study, the contact is:

Prof. Livia Puljak, MD, PhD

University of Split School of Medicine, Šoltanska 2, 21000 Split, Croatia

Phone: +385-21-557-807

email: livia.puljak@mefst.hr

Your response to this email will be considered as consent to participate in this study

I am looking forward to hearing from you.

Kind regards, Mirko Gabelica, MD

9.3 Appendix 3. Personalized emails to each potential participant in Study 2

Dear **<title and name of the corresponding author>**

My name is Mirko Gabelica, MD, and I am a researcher affiliated with the University of Split School of Medicine in Split, Croatia. Together with Prof. Livia Puljak, MD, PhD from Cochrane Croatia, I am conducting a study entitled Data sharing practices among authors of recent studies published in BioMed Central. Your manuscript was selected for this study.

In your manuscript, **<manuscript title>**, you have indicated in section Availability of data and materials that data collected within your study are available on request.

I am kindly asking you to please share with our team a copy of the raw data set used in Your study. Our team is interested in re-examining those raw data sets, whether or not they are adequate for reanalysis. This analysis will be de-identified, and in the presentation of our results, we will not disclose any details about author characteristics. All data and communication will be treated with strict confidence. Only the summary results will be presented. All raw data sets will be deleted after being examined.

We hope that our findings will contribute to the open data sharing movement in biomedical research.

Our study was approved by the University of Split School of Medicine Ethics Committee (approval number, Klasa: 003-08/17-03/0001, Ur.br: 2181-198-03-04-17-0052). If you have any questions or concerns that you would like to address with the principal investigator of this study, the contact is:

Prof. Livia Puljak, MD, PhD
Cochrane Croatia
e-mail: livia.puljak@unicath.hr

Mirko Gabelica, MD
University Hospital
Centre Split, Croatia
mgabelica@kbsplit.hr

9.4 Appendix 4. Non-disclosure agreement

NON-DISCLOSURE AGREEMENT

THIS AGREEMENT [**the Agreement**] is entered into on this [insert number of the day] day of [insert month and year] by and between:

1. Researcher(s) sharing raw data from their research study [Insert the name(s) of the lead researcher (s)], affiliated with the [insert the Legal Address of the Entity] hereinafter referred to as [**the Discloser**] and

2. Researchers requesting raw data from the research study of the Discloser, namely: Mirko Gabelica, MD, affiliated with the University of Split Hospital (registered ENT surgeon) and University of Split School of Medicine (PhD student) in Split, Croatia, and Prof. Livia Puljak, MD, PhD, Full Professor and Head of the Center for Evidence-Based Medicine and Healthcare at the Catholic University of Croatia in Zagreb, Croatia, hereinafter referred to as [**the Recipients**].

Purpose of the Agreement

The Agreement is used to define confidentiality aspects of the raw data from the research study of the Discloser, which were requested by the Recipients from the Discloser for the purpose of the study titled "Data sharing practices among authors of recent studies published in BioMed Central", which was approved by the Ethics Committee of the University of Split School of Medicine.

Non-disclosure obligations by the Recipients

Hereby, the Recipients confirm that their research intention is solely to examine data sharing practices of authors in the field of biomedicine, to study whether those authors have de-identified raw data available for sharing and whether the shared data would allow reanalysis.

The Recipients **will not** conduct any new analyses, or publish further studies based on the shared de-identified raw dataset, or share the de-identified raw data of the Discloser with anyone else outside the research team of the Recipients (Dr. Gabelica and Prof. Puljak).

The Recipients agree to:

- keep all the research information shared confidentially by not discussing or sharing the research information or shared data in any form or format (e.g., disks, tapes, transcripts) with anyone other than the Discloser.
- keep all research information in any form or format (e.g., digital files, disks, tapes, transcripts, etc.) secure while it is in possession of the Recipients.
- return all research information received in a physical form or format (e.g., disks, tapes, transcripts) to the Discloser when the Recipients have completed their study.
- to destroy all research information and raw data received from the Discloser in any form or format regarding this research project that is not returnable to the researcher (s) (e.g., information stored on computer hard drive) five years after completion of the Recipients' study.

The Discloser:

_____	_____	_____
(Print Name)	(Signature)	(Date)

Mirko Gabelica, on behalf of both Recipients:

_____	_____	_____
(Print Name)	(Signature)	(Date)

9.5 Appendix 5. Scanned approval of the study protocol by the Ethics Committee of the University of Split School of Medicine for Study 2

MEDICINSKI FAKULTET
SVEUČILIŠTA U SPLITU
Etičko povjerenstvo
Split, 22. svibnja 2019.

Klasa: : 003-08/19-03/0003
Ur. br.: 2181-198-03-04-19-0043

MIŠLJENJE

Etičkog povjerenstva povodom prijave istraživanja:
Compliance with Data Availability Statement included in a research manuscript

I. Zaprimljen je zahtjev Mirka Gabelice, dr. med. studenta poslijediplomskog studija TRIBE, za odobrenje znanstvenog istraživanja pod nazivom: **Compliance with Data Availability Statement included in a research manuscript** – provedba znanstvenog istraživanja na ljudima. Predviđeno je da ovo istraživanje započne u svibnju 2019. g. te da traje 3 mjeseca. Glavni ciljevi ovog istraživanja su: 1. analizirati jesu li autori spremni podijeliti podatke iz objavljenog rada u skladu s izjavom o dostupnosti podataka te 2. istražiti jesu li podijeljeni neobrađeni podatci pripremljeni na način koji bi omogućio ponovnu analizu (sekundarna analiza primarnih podataka).

II. Etičko povjerenstvo Medicinskog fakulteta Sveučilišta u Splitu je, prilikom raspravljanja o ovom predmetu, uzelo u obzir izjavu prijavitelja da rizika za ispitanike nema. Također je uzeta u obzir izjava da će identitet ispitanika (zdravog ili pacijenta) uvijek ostati anoniman.

III. Sukladno odredbi članka 16. Etičkog kodeksa Medicinskog fakulteta u Splitu Povjerenstvo je zauzelo stajalište kako je predmetno istraživanje **u skladu s odredbama Etičkog kodeksa** koje reguliraju istraživanja na ljudima u znanstvenom, istraživačkom i stručnom radu i etičkim načelima Helsinške deklaracije.

IV. Mišljenje je doneseno jednoglasno.

Predsjednik Povjerenstva:
prof. dr. sc. Marko Ljubković

SVEUČILIŠTE U SPLITU MEDICINSKI FAKULTET 12

Dostaviti:
- Mirko Gabelica, dr. med.
- arhiv Etičkog povjerenstva Medicinskog fakulteta
- arhiv Fakulteta

10 Curriculum vitae

Mirko Gabelica

Date and place of birth:

November 17, 1983, Split, Croatia

Address:

Alkarska 160, 21000, Split, Croatia

Email: mgabelica@kbsplit.hr, gabelica@gmail.com

Education:

2002 – 2010 Medical Doctor, University of Split, School of Medicine, Split, Croatia

2015 – 2021 PhD Program Translational Research in Biomedicine (TRIBE), University of Split School of Medicine, Split, Croatia

Work experience:

2018 - present. Attending specialist at Department for ear nose and throat disorders with head and neck surgery residency, University Hospital Centre Split

2019 - present. Otorhinolaryngology sub-specialization, head and neck plastic and reconstructive surgery

2013 - 2018. Otorhinolaryngology with head and neck surgery residency at Department for ear nose and throat disorders with head and neck surgery residency, University Hospital Centre Split

2011 - 2013. An emergency medical doctor at Department for emergency medicine Split-Dalmatia County, an emergency medical doctor at Helicopter Emergency Service at Department for emergency medicine Split-Dalmatia County

Publications:

1. Gabelica M, Sapunar D, Marušić M, Puljak L. The ideal repository for hosting data from clinical trials: blueprint using business process management F1000Research. 2021;10:23.

2. Gabelica M, Tafra R, Martinić MK, Kontić M, Markić J, Kovačević T, Čulo Čagalj I, Ninčević Ž. Feather foreign body caused periparotid and peritonsillar abscess in a 9-month-old girl, Auris Nasus Larynx. 2021;48(5):1023-25.

3. Paladin I, Martinić MK, Gabelica M, Puljak L. Nasopharyngeal perforation after blunt neck trauma during epileptic seizure, J. Oral and Maxillofac. Surg. 2020;78(10):1812.e1-1812.e4.

4. Gabelica M, Čavar J, Puljak L. Authors of trials from high-ranking anesthesiology journals were not willing to share raw data, J Clin Epidemiol. 2019;109:111-6.

5. Kontić M, Čolović Z, Paladin I, Gabelica M, Barić A, Pešutić-Pisac V, Association between EGFR expression and clinical outcome of laryngeal HPV squamous cell carcinoma, Acta Otolaryngol. 2019;139(10):913-7.

6. Ninčević Ž, Ninčević J, Gabelica M, Šundov Ž, Puljiz Ž, Gabelica M. Severe glyphosate - surfactant herbicide poisoning; successful treatment-case report. MOJAMT. 2017;4(1):202–4.

7. Krleža-Jerić K, Gabelica M, Banzi R, Krnić-Martinić M, Pulido B, Mamić Kahnjo M, Reveiz L, Šimić J, Utrobičić A, Hrgović I. IMPACT Observatory: tracking the evolution of clinical trial data sharing and research integrity, Biochem Med. (Zagreb) 2016;26(3):308-317.