Scientometrics (2025) 130:763–779
https://doi.org/10.1007/s11192-024-05227-5

The compliance to FAIR principles of shared data in addiction research

Andrea Sixto-Costoya[1,2] · **Antonia Ferrer-Sapena**[3] · **Rafael Aleixandre-Benavent**[2,4,5] · **Fernanda Peset**[3] · **Juan Carlos Valderrama-Zurián**[2,4] · **Luiza Petrosyan**[3]

Received: 8 March 2024 / Accepted: 20 December 2024 / Published online: 11 January 2025

Abstract
The aim of this study is to assess the scientific data sharing in the field of addictions by applying FAIR principles. These principles play an important role, as they guarantee a minimum of findability, accessibility, interoperability and reusability of the shared data. They are one of the main measures to improve the integrity and quality of research data. For this study, three automated tools were used: the Data Citation Index (DCI) to capture datasets on addictions; Bibliometricos, proprietary software for data retrieval; and the F-UJI tool for the FAIR evaluation of datasets. The datasets on the most common addiction topics, such as alcohol, cannabis, tobacco, cocaine, opioids and stimulants, were downloaded by the DCI (5967 DOIs) and parsed into a database for subsequent analysis. In terms of datasets characteristics, alcohol, tobacco and opioids were the most productive. After assessment by F-UJI, none of the addictions analyzed reached an average of 30% FAIR compliance since all of them were between 20% and 29%. When analyzing each principle, Findable was the best scored principle (in a range of 40%–59%), followed by Accessible, Interoperable and Reusable. The results of our study show, first, an increasing number of shared datasets over the years, especially from basic studies. In terms of quality, there are issues that remain to be resolved, especially in relation to interoperability and reusability principles. This emphasizes the important role of adequate data sharing procedures in ensuring that datasets are FAIR compliant and usable in addiction research.

✉ Luiza Petrosyan
luipet@inf.upv.es

[1] Grupo UISYS Departamento de Trabajo Social y Servicios Sociales, Facultad de Ciencias Sociales, Universitat de València, Valencia, Spain

[2] Unidad asociada al Instituto Interuniversitario de Investigación Avanzada Sobre Evaluación de la Ciencia y la Universidad (INAECU), UC3M-UAM3 Valencia, Spain

[3] Instituto Universitario de Matemática Pura y Aplicada – IUMPA, Universitat Politècnica de València, Valencia, Spain

[4] Departamento de Historia de la Ciencia y Documentación, Facultad de Medicina y Odontología, Universitat de València, Grupo UISYS, Valencia,, Spain

[5] Instituto de Gestión de la Innovación y del Conocimiento - Ingenio (CSIC-Universitat Politècnica de València), Valencia, Spain

Keywords Addictions · Data sharing · Open data · Repositories · FAIR principles

Introduction

In 2016, the so-called FAIR principles, which stand for "findable", "accessible", "interoperable" and "reusable", were published to bring order and consistency to the practice of sharing research data (Wilkinson et al., 2019). This practice, which is gaining momentum in the context of open science (OS) worldwide, is currently one of the main measures proposed to improve the integrity and quality of research. For this reason, many funders of competitive calls for projects promote this practice at the national and international levels (European Commission, 2016, 2022; National Institutes of Health, 2022; Ross et al., 2016). Additionally, it is present in the author guidelines, either as a requirement or as a recommendation, of numerous publishers from various disciplines (British Medical Journal, n.d.; Elsevier, 2019; PLOS One, 2019; Wiley, 2023). The functions of accountability in sharing research data and ability to reproduce experiments have already been amply pointed out (Pasquetto et al., 2017). Without compliance with these principles, especially reusable, the data quality can be so low that it is rendered useless by the difficulty understanding it. This is evident in other areas, such as data science, where the analysis of massive amounts of data is related to their quality. For Brenan Brennan (2017), "Data quality is commonly conceived as a multidimensional construct defined as “fitness for use”", which coincides with the two principles of interoperable and reusable, essential for the development of reliable generative artificial intelligences.

The area of addiction research is not new to the OS movement or data sharing practices (Scheibein et al., 2022). In the words of Munafò (2016), the scientific community studying addictions may even be well placed to embrace open science and open data. The collaborative culture that has historically existed in this area and the existence of regulations that predate and directly align with these movements, such as the requirement for preregistration of study protocols for clinical trials and data sharing for genetic studies, encourage this behaviour. Nevertheless, both papers find that the practice of data sharing is still more the exception than the norm in the addiction field.

In this line, some studies have examined how the practice of sharing data in addiction research has been carried out. In the case of the study by Vidal-Infer et al., (2019), they investigate the raw data in the supplementary material of journals indexed in the Journal Citation Report of Clarivate Analytics, finding that datasets in reusable formats were in the minority. Subsequently, Gorman (2020) continues by analysing the highest impact journals in the substance abuse category that have data sharing policies to determine whether they indicated where their data could be obtained and whether there was a formal data sharing statement. In another related publication, Gorman (2019) investigates the measures that journals in the addiction field take to improve the integrity of publications, finding that encouraging data sharing was the second most common measure, preceded by declaring conflicts of interest. In another work more focused on journal policies related to data sharing, Aleixandre-Benavent et al., (2014) find that most journals in the addiction field admit the possibility of storing data in thematic or institutional repositories, as well as accepting additional material and data reuse. Following the results of this study, most of them did not have specific instructions for authors.

Despite studies examining the practice of data sharing in addiction research, none have examined whether these data meet minimum quality standards. Therefore, the aim of this study is to analyse a large number of datasets dealing with the topic of addictions through automatic approaches to assess the compliance of the data with the FAIR principles. The results will provide insight into the degree of compliance to FAIR principles, the aspects that can be improved and the way forward in a context, both scientific and in professional practice, where the handling of large volumes of data is a matter of crucial importance.

Methods

The methodology used in this study consists of three stages (Fig. 1):

- Capturing datasets on addictions
- Downloading the records
- FAIR evaluation of datasets with the F-UJI tool

Capturing datasets on addictions

Six search equations were designed to retrieve datasets, one for each type of studied addiction: alcohol, cannabis, cocaine, opioids, stimulants and tobacco. The specific search equation for each substance can be found in the Supplementary Material.

The datasets were obtained through the Data Citation Index (DCI), a resource offered by the Clarivate company and included in the Web of Science (WoS) catalogue. The DCI provides access to many descriptive records on datasets obtained through its partnership with repositories and large-scale data and metadata providers. The DCI contains more than 80 types of data (e.g., images, sequence data, geoscientific information, and audiovisual), and all of them were included in our study. The strength of this resource is that it provides an overview that is very difficult to achieve when datasets or repositories are viewed in isolation (Clarivate Analytics, 2017).

Downloading the records

The WoS bibliographic information from the six searches mentioned above was imported for each retrieved record: (1) author/creator; (2) title of the dataset; (3) year of data publication (collected records published from 1900 to the present); (4) data source (repository where the data are deposited or the organization in charge of making the

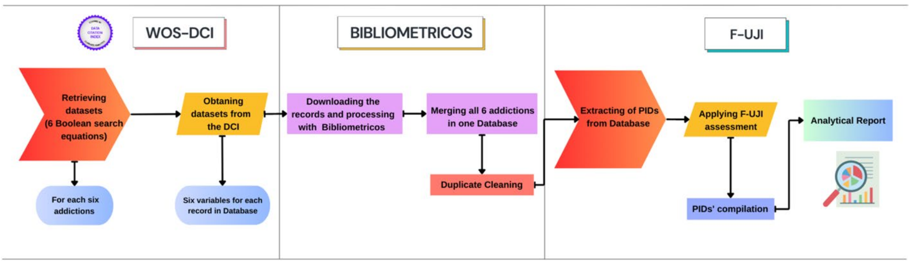

Fig. 1 Assessment stages and process. The WOS-DCI tool is used for capturing datasets, BIBLIOMETRICOS for downloading records, and F-UJI for the FAIR assessment stage

Table 1 Classification of addiction by type of unique identifier

Addiction	DOI	URL
Alcohol	1203	897
Cannabis	44	23
Cocaine	170	540
Opioids	106	74
Stimulants	24	4
Tobacco	2302	3622

data available); (5) Web of Science subject category of the data source (i.e., Genetics & Heredity; Biochemistry & Molecular Biology or Biodiversity Conservation); and 6) persistent identifier (PID) (e.g., digital object identifier (DOI), unique URL, a databank accession number or other permanent identifier). The year of publication, the data source and the Web of Science subject category are considered for further analysis.

The total recovered records (datasets) for each substance were n = 2100 for alcohol, n = 67 for cannabis, n = 710 for cocaine, n = 180 for opioids, n = 28 for stimulants and n = 5924 for tobacco. These records were downloaded in.txt format and processed with our Bibliometricos software.[1] Six databases were created, one for each type of addiction. To make it more efficient, all six were merged into a single database, creating a new variable called "type of addiction" to differentiate the datasets once the records were merged.

Once this information was parsed and organized, it was necessary to know the unique identifier of the total records in each substance. This identifier allowed us to move to the next step of the methodology. For this purpose, the dataset was divided into records that have a DOI and records that have other types of identifiers (such as URL). In addition, the results can be observed in Table 1.

FAIR evaluation of datasets with the F-UJI tool

The FAIR assessment for these datasets was performed with the F-UJI tool that evaluates research data objects, which is basically a REST API using OpenAPI Specification[2] from a remote server, published under an open-source MIT licence. It is based on aggregated metadata, including metadata embedded in the landing page and metadata retrieved from a PID. F-UJI considers the following schemes to be PIDs: handle, persistent uniform resource locator, archival resource key, permanent identifier for web applications and DOI (Sun et al., 2022).

The utilization of F-UJI in assessing datasets involves an examination of numerous cross-domain metadata standards, including those put forth by Dublin Core, DCAT, DataCite, and Schema.org, as expounded upon by Devaraju et al., (2022). The outcomes of such evaluations yield diverse scores pertaining to the metadata of data and datasets, with 16 metrics distributed across four principles: findability (5 metrics), accessibility (3 metrics),

[1] Bibliométricos: proprietary software dedicated to retrieving records from various resources such as WoS and parsing them in a relational database for further analysis.

[2] *OpenAPI Specification (OAS)* defines a standard, programming language-agnostic interface description for HTTP APIs.

interoperability (3 metrics), and reusability (5 metrics). They delve into the aspects of each of these metrics in more detail.

Devaraju et al., (2022) states that, the metrics are based on the set of indicators proposed by the RDA FAIR Data Maturity Model Working Group (FAIR Data Maturity Model Working Group, 2020), and the specification of metrics follows the template modified by Wilkinson (Wilkinson et al., 2018). The approach of the tool is hierarchical, whereby each FAIR principle is associated with specific metrics, and for each metric, one or more practical tests are defined. Each test is based on community best practices and standards (Table 2).

1. The online web application offers an intuitive summary of the FAIR aspects of the evaluated PID manually by inserting only one PID individually (https://www.f-uji.net/index.php?action=test)
2. The tool can be installed locally and run API calls in a loop for a list of PIDs. In this way, F-UJI can compile many datasets automatically (the script is published by the FAIRsFAIR project and can be found on GitHub).

Table 2 The complete set of metrics used by F-UJI (Devaraju et al., 2022)

Number	FAIR principle	Metric Identifier	Description name
1	F	FsF-F1-01D	Data is assigned a globally unique identifier
2	F	FsF-F1-02D	Data is assigned a persistent identifier
3	F	FsF-F2-01 M	Metadata includes descriptive core elements (creator, title, data identifier, publisher, publication date, summary and keywords) to support data findability
4	F	FsF-F3-01 M	Metadata includes the identifier of the data it describes
5	F	FsF-F4-01 M	Metadata is offered in such a way that it can be retrieved by machines
6	A	FsF-A1-01 M	Metadata contains access level and access conditions of the data
7	A	FsF-A1-02 M	Metadata is accessible through a standardized communication protocol
8	A	FsF-A1-03D	Data is accessible through a standardized communication protocol
9	I	FsF-I1-01 M	Metadata is represented using a formal knowledge representation language
10	I	FsF-I2-01 M	Metadata uses semantic resources
11	I	FsF-I3-01 M	Metadata includes links between the data and its related entities
12	R	FsF-R1-01MD	Metadata specifies the content of the data
13	R	FsF-R1.1-01 M	Metadata includes license information under which data can be reused
14	R	FsF-R1.2-01M	Metadata includes provenance information about data creation or generation
15	R	FsF-R1.3-01M	Metadata follows a standard recommended by the target research community of the data
16	R	FsF-R1.3-02D	Data is available in a file format recommended by the target research community

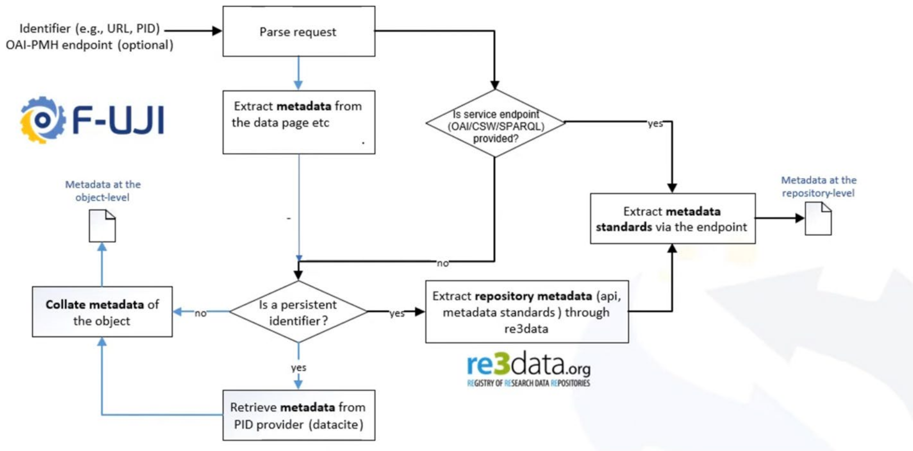

Fig. 2 The high-level flow of each Fuji assessment summarized (Huber, 2021)

In our case, having many datasets to compile (5967 PIDs), we adopted the second approach to assessment.

F-UJI is taking a PID or URL input parameter to perform the assessment (Fig. 2). It then attempts to identify the metadata provided. Additionally, it employs other techniques, such as content negotiation or following type links, to collect supplementary metadata, which may be presented in alternative formats by the provider. The collated metadata is then employed in the subsequent assessment stages, with the resulting scores predefined in the architecture of the tool based on the aforementioned document.

The tool gives us the JavaScript Object Notation (JSON) files for each PID with an assessment result that includes scores, practical tests, inputs/outputs and assessment contexts for each of the 16 metrics (Fig. 3).

The final stage of the evaluation of our datasets is to visualize the scores after compiling all PIDs. By creating a report using a computational notebook (.ipynb document) provided by the FAIRsFAIR team (*GitHub—Pangaea-Data-Publisher/Fuji: FAIRsFAIR Research Data Object Assessment Service*, n.d.), we obtain an overall analysis and visualization of all PID responses that were assessed using F-UJI.

During the analysis, we focused on two specific aspects of these reports: the overall FAIR scores and the scores obtained for each principle.

This methodology has been successfully applied in other works analysing other areas, such as agriculture (Petrosyan et al., 2023).

Results

Data obtained from the DCI

Regarding the data obtained from the DCI, an analysis of the year of publication, data source (repositories) and WoS category was carried out.

Regarding the year of publication, the merged database of the six addictions was considered for this analysis. Accordingly, removing duplicates was necessary because it was

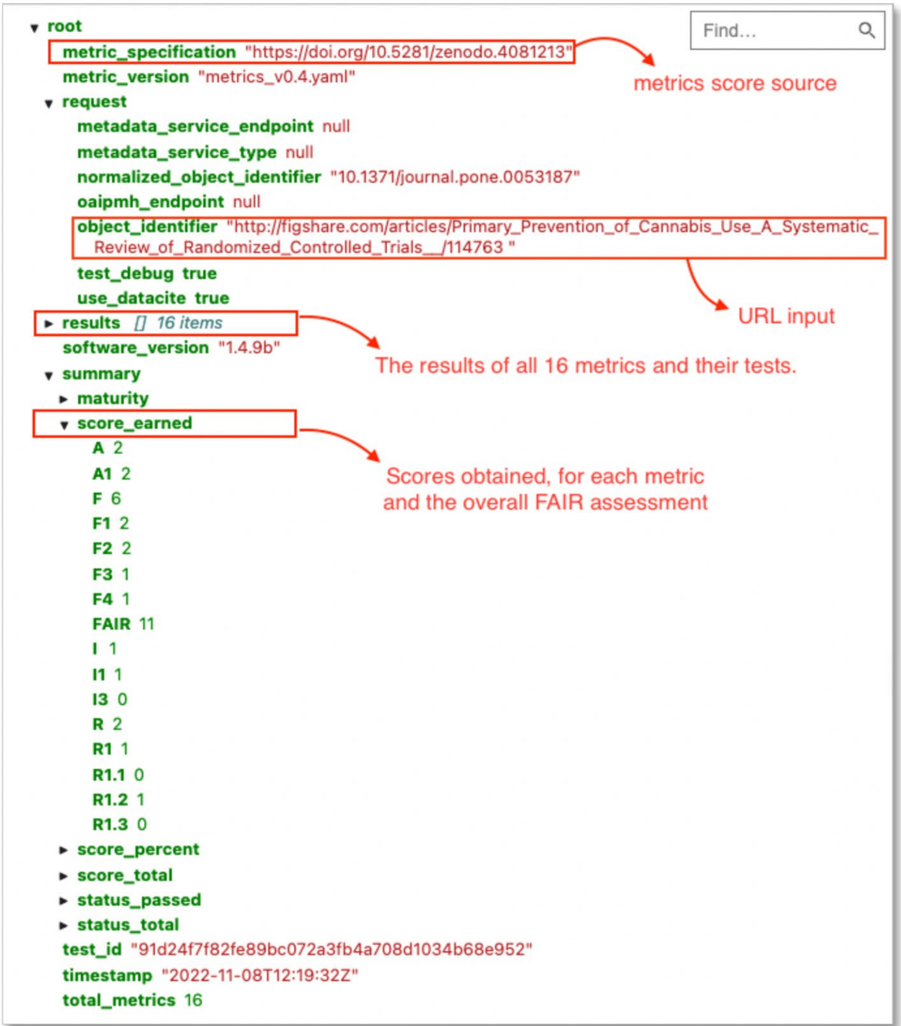

Fig. 3 The following example illustrates the JSON report generated following the evaluation of the tool. It displays the metadata collected from the input URL in the "request" section and the scores obtained for each metric

possible that one dataset could be classified as including more than one addiction. The result without duplication was 7902 datasets. Over this quantity, we calculated the frequency per year, which showed sustained growth from the first years to the 2010s onwards, where stabilization was observed. The most productive years were 2014, 2015 and 2016, accounting for 40% of the total.

In terms of data sources (repositories), forty repositories were found for all the addictions studied. At the general level, the results showed that among the repositories with more than 10 datasets, both generic and thematic repositories on genetics were predominant, representing close to 100% of the total.

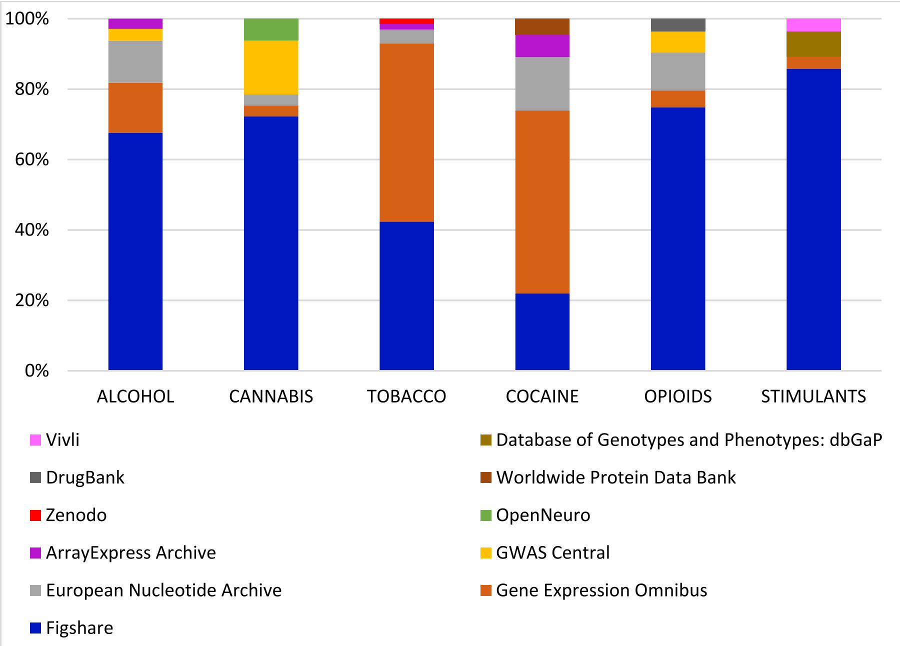

Fig. 4 Percentage of datasets deposited in the top 5 repositories per addiction

When a differentiation between types of addiction was performed, the top 5 repositories for each type of addiction were considered. As shown in Fig. 4, three repositories contain data on all the addictions: Figshare, Gene Expression Omnibus and European Nucleotide Archive. Notably, specific repositories on gene expression profiles or DNA and RNA sequences were more common for cocaine and tobacco use than for use of other substances, with greater representation in generalist repositories such as Figshare. Nevertheless, the high presence of thematic repositories for all addictions is remarkable, since all the most frequent repositories used, except Figshare, focused on a specific area. Most of the thematic repositories are designed to deposit and store data on basic sciences, with Vivli (clinical research data) and DrugBank (drug and drug target information) being the exceptions. This result is complemented by the data shown in Fig. 5, where the classification per top 5 WoS source category[3] is observed. Here, the most frequent categories after multidisciplinary sciences are basic research, such as genetics and heredity or biochemistry and molecular biology, with other disciplines such as neurosciences or medicine, research and experimental occurring less often.

[3] Web of Science Core Collection Overview—http://webofscience.help.clarivate.com/en-us/Content/wos-core-collection/wos-core-collection.htm?Highlight=category

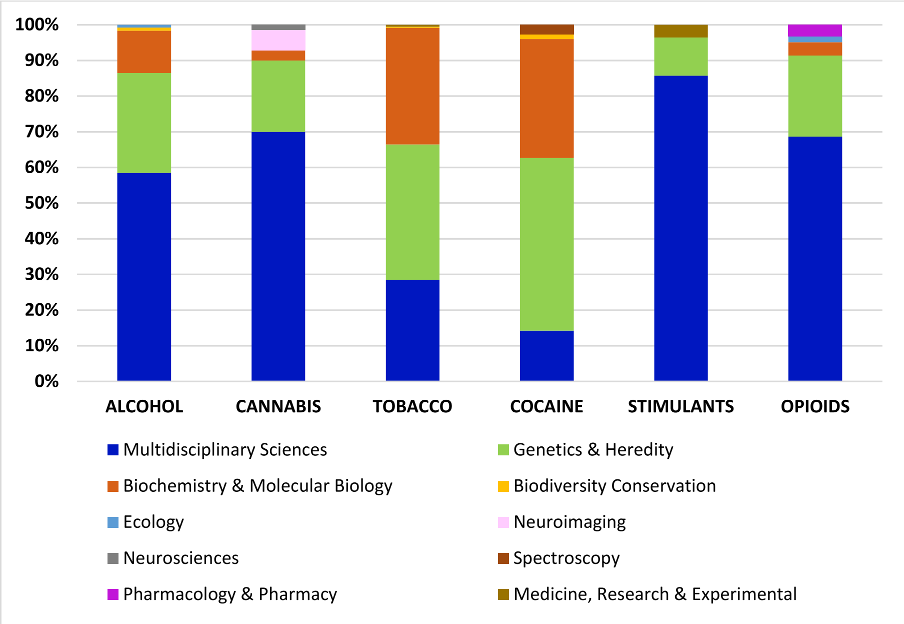

Fig. 5 Percentage of datasets by WoS thematic Category differentiated by each addiction

FUJI analysis of PID (DOI and URL)

The FUJI analysis results were divided into DOI object results and URL object results. The division by type of object (DOI and URL) was performed at the stage of downloading from the DCI, as it was one of the most significant differences within our PID database.

As we mentioned in the methodology, we considered a) the total FAIR level score obtained by each addiction and b) the differentiated score according to the FAIR principle.

DOI object results

None of the addictions analysed reached an average of 30% FAIR compliance, since all were between approximately 20–29%. Within this range, cannabis addiction scores higher in general FAIR terms, followed by cocaine.

Following the analysis by each principle (Fig. 6), the trend is similar for all types of addictions. Findability had the highest score, followed by accessibility, interoperability and reusability. All types of addiction scored above 50% in findability (with the highest score for cannabis, at 59.3%), but none of the remaining principles was able to reach this percentage. Again, cannabis addiction scored the highest in all principles. Stimulant addiction scored quite poorly for interoperability and reusability compared to other forms of addiction.

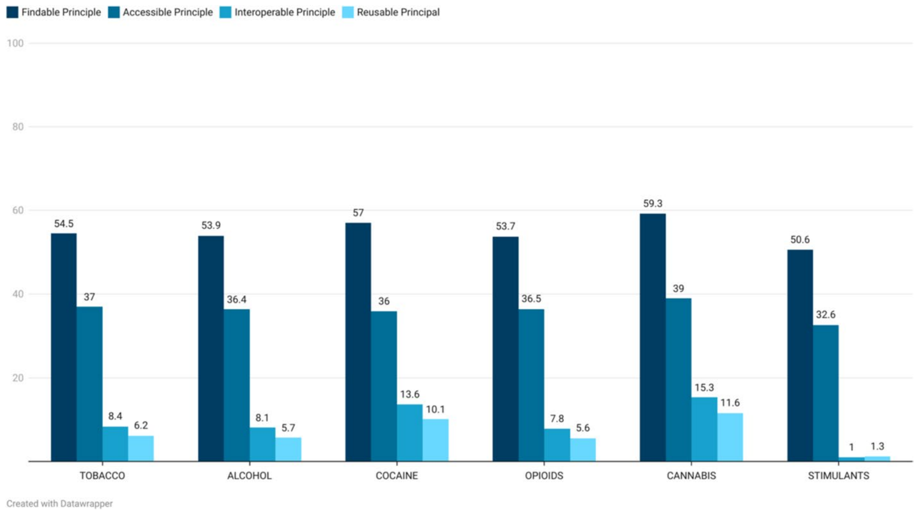

Fig. 6 FAIR scores per principle for DOIs

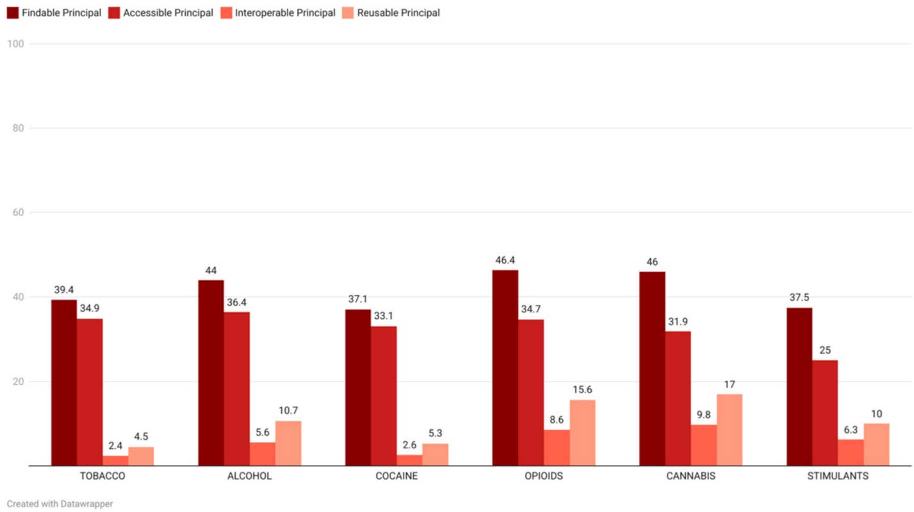

Fig. 7 FAIR scores per principle for URL

URL object results

Regarding the analysis of URL identifiers, the results of the six addictions ranged from 17 to 22%, having cannabis and opioids the highest scores.

Observing the differences among the principles (Fig. 7), a distinct tendency from DOIs can be detected here. In the case of URLs, findability and accessibility also scored higher than interoperability and reusability (even though the results were much lower than for the DOI group). Although none of the addictions achieved 50% compliance for any of the principles, the principle of interoperability was highlighted due to its low score in all cases, as it did not reach 10% compliance. The interoperability and reusability scores for stimulants (6.3% & 10%) were interesting, as they were higher than in the DOI group (1% & 1.3%).

Finally, despite the overall low mean scores for both the DOI and URL results, importantly, a dispersion was detected in some cases between the best and worst scoring datasets. This means that there are datasets that obtained good or even very good FAIRness scores but were hidden in the average due to the majority obtaining low scores. That is, in the DOI group, some datasets scored in the range of 70–85% for FAIRness, and in the URL group, they scored in the range of 50–60% (this can be observed in the supplementary figures available on Figshare, which were obtained directly from the report).

Discussion

The results of this study allow us to approximate the quantity, characteristics and quality of published datasets in the area of addictions research. This was possible due to the use of the DCI, the resource of Clarivate Analytics that collects datasets from a wide range of proven solvent repositories. Moreover, in a second step, the level of FAIRness of those collected datasets was checked through the F-UJI tool of the FAIRsFAIR project. The overview obtained is very useful to understand the current state of the practice of data sharing in a context that increasingly requires the participation of different scientific actors not only in the practice of data sharing itself, but also in the need to share quality data.

Regarding the quantity and characteristics of the datasets, it is important to highlight that there are large differences between addictions in terms of the volume of datasets. As we can observe, tobacco addiction had the most published datasets, followed by alcohol and opioid. This is due to tobacco and alcohol being the most consumed substances currently and historically (Substance Abuse and Mental Health Services Administration (SAMHSA), 2022). As for the opioids, there are examples in the literature of the usefulness of sharing data in the fight against the epidemic opioid's crisis (Lin et al., 2019; Stopka et al., 2021). Another proof of this relationship is the use of open data portals to obtain raw data, for example the NYC Open Data to find alcohol-related complaints in NYC (Ransome et al., 2019); an Open Data Kit Software form creating questionnaires for collecting data on tobacco availability around education centres (Leimapokpam & Swasticharan, 2017); the World Bank Open Data for studying the impact of public health policies on alcohol-related liver disease (Díaz et al., 2021); or the use of government open data sources to obtain sociodemographic indicators for studying the socioeconomic determinants of opioid prescribing in England (Schifanella et al., 2020).

As for the repositories, our study showed us that at least in the topic of addiction research, the datasets were mostly deposited in multidisciplinary and genetic repositories. The case of cocaine research stands out, as it does not present the largest number of published articles, but it does present the largest number of datasets among the group of illicit drugs. This could be due to the large amount of basic research related to cocaine (the most frequent WoS category in cocaine was genetics & inheritance, and "Gene Expression Omnibus" was the most used repository). For this reason, cocaine research is more related to data sharing practices than others (Haeusermann et al., 2018; Pereira et al., 2014; Vassilakopoulou & Aanestad, 2019).

Regarding the quality, this leads us to the results of the second part of our study, that is the degree of compliance of the FAIR principles by the shared datasets. In this regard, the analysis using the F-UJI tool offers different interpretations. First, although the registration of a digital data object can be accomplished through either a persistent or nonpersistent identifier, it is worth noting that the use of a nonpersistent identifier, such as a URL, may result in limited automated evaluation capabilities (Devaraju & Huber, 2021). Thus, the usefulness of persistent identifiers is demonstrated at least at the level of analysis of compliance with FAIR principles through F-UJI.

In terms of the level of compliance for six addictions and their identifier groups, they all had similar score patterns, with findability and accessibility being the most compliant and interoperability and reusability being the least compliant. This pattern is similar to that obtained in the study by Petrosyan et al., (2023) in the area of agriculture, where the worst scores were also obtained on the principles of interoperability and reusability.

Although no other examples of FAIR principles compliance in addiction research were found, several recent studies highlight both the relevance of considering FAIR principles in health data management and the difficulties that remain in doing so (Bhatia et al., 2020; Holub et al., 2018). These difficulties, which are in line with the results of our study, have to do with the lack of key elements for compliance with the FAIR principles but also with the lack of a culture of sharing among both institutions that manage health data, such as hospitals or health departments, and professionals, such as clinicians or researchers. As Bhatia et al., (2020) mentioned, there is a need to modernize the way in which researchers use, share, find and cite datasets in the research community as a whole. Their study focused on creating a catalogue that addressed the deficiencies in the Emergency Medicine databases included in the Healthcare Cost and Utilization Project (HCUP) using the FAIR principles as a guide. These deficiencies are related to the absence of searchable tools, the lack of unique identifiers, the ambiguity of what counts as a dataset, and finally, the need to standardize terminology. In another study by Holub et al., (2018), some challenges of reusing biological material in medical research were raised. These challenges were related to quality aspects, incentives, privacy-respecting approaches and the lack of common guidelines for cloud architectures that can deal with the enormous growth of medical data. From their point of view, life sciences still suffer from fragmentation and substantial reproducibility issues. These concerns were especially evident during the COVID-19 pandemic, when the FAIRification of healthcare data was a necessity. Following processes such as the one proposed by Queralt-Rosinach et al., (2022), they argued that using existing standardized terminology can turn patient data into machine-readable objects for secondary uses. The health sector's interest in compliance with the FAIR principles also manifests itself in the search for solutions that make compliance with the principles compatible with personal data management and data protection (Delgado & Llorente, 2022; Govarts et al., 2022). In this sense, Govarts et al., (2022) pointed out that all the new technical developments in the European Union under the Open Science umbrella, such as federated data management and analysis systems, machine learning, advanced search software, harmonized ontologies, and data quality standards, need a balance between FAIR data management and compliance with personal data protection safeguards.

Other examples of how FAIR principles have influenced health data were shown through initiatives such as OpenPVSignal, by which a model was developed with the aim of supporting the semantic enrichment and rigorous communication of pharmacovigilance (PV) signal information to 1) publish signal information according to the FAIR principles and 2) exploit automatic reasoning capabilities upon the interlinked PV signal data (Natsiavas et al., 2018). Additionally, Project Tycho, which is a repository designed to improve FAIR compliance with global health data (epidemiological data), discovered the necessity of developing community-accepted standard representations for these data (Van Panhuis et al., 2018), similar to other specialized schemes. Moreover, the initiative proposed by Leung and Dumontier (2019) from the clinical field also defends the need for a new generation of clinical practice guidelines (CPGs) in the health system that considers the FAIR principles, offering a basic checklist for developing those FAIR-CPGs.

Limitations

This study has some limitations that need to be discussed. First, the Data Citation Index was used to locate the datasets in the area of addictions. This index, which belongs to the WoS catalogue of the company Clarivate Analytics, provides access to information on 449 checked and validated repositories as of 2023. This assures us of the quality and relevance of the repositories but does not cover all repositories, so it is possible that some relevant repositories may not have been captured. Another limitation is the F-UJI tool itself, which, despite being the most complete of those that have been tested by the authors, is not perfect and is in continuous improvement, in line with the changing subjects it analyses, the sharing of research data and compliance with the FAIR principles.

Conclusions

The main conclusions of our work revolve around two facts. First, a significant effort to share datasets has been observed in the field of addiction research, demonstrated by the growing number of datasets over the years. This could mean that the increase in policies, indications or even mandates for data sharing are having an effect at least in terms of the number of datasets shared. Institutions such as the European Commission have developed scientific open data policies that include not only the obligation of correct treatment of the data generated in their funded projects (including compliance with the FAIR principles), but also the indication to the Member States to include similar policies in the calls for local projects (European Commission, 2024; SPARC Europe, 2021). Similarly, the National Institutes of Health are involved in the development and implementation of policies to ensure access to the data generated in their funded projects (National Institutes of Health, 2024).

Second, the results of our study showed the issues that remain to be resolved, especially in relation to the FAIR principles of interoperability and reuse. As has been seen in other studies, there are initiatives that, seeing the need to share data in accordance with the FAIR principles, are working to find useful methodologies and systems. In this line, our study aims to contribute to this, offering analyses and methodologies that should help in the collective search for solutions in the field of addictions, which effectively calls for solutions in the context of data sharing. Compliance with the FAIR principles is crucial not only to make the practice of data sharing useful to researchers in their daily tasks but also to make these data (and the accompanying metadata) machine-readable, both for use in research and in "datametrics" studies such as this one. Several relevant statements and initiatives such as the recent CoARA—Coalition for Advancing Research (Coalition for Advancing Research Assessment (CoARA), 2022) have insisted on Open Science and Open Data as a paradigm and FAIR principles as a model to follow. The now classic article by (Wilkinson et al., 2016) on FAIR principles and how to achieve them, including the use of repositories and indications of what a good deposit would look like, is still very useful for researchers and institutions. The reason for the poor results obtained on the Reusable and Interoperable principles in our studied sample seems to be related to the lack of operational knowledge once the deposit in a chosen repository has been made. In this line, institutions and funding agencies have a duty not only to promote and require open research data, but also to provide training and support to researchers from different disciplines to do it correctly.

Future directions

A new research direction is proposed that is under development, which will apply machine learning algorithms to results from F-UJI, in order to classify and identify FAIR metrics more accurately after evaluation by the tool. This approach will facilitate the effective and rapid detection of metrics that fail during the evaluation of such a large number of datasets, as in our case, particularly those related to the principles of interoperability and reusability during the assessment. Such approach would help the identification of gaps and bring these problems to the attention of both the repositories that publish them and the scientific community in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11192-024-05227-5.

Acknowledgements The authors would like to thank the grant "Stable methodologies to evaluate and measure quality, interoperability, blockchain and reuse of open data in the agricultural field" funded by MCIN/AEI/https://doi.org/10.13039/501100011033; the predoctoral grant (PRE2020-092585) of LP; the postdoctoral grant (MS21-020) of ASC; and the funding from the Generalitat Valenciana (Spain) through the PROMETEO 2024 CIPROM/2023/32 grant.

Author contributions Luiza Petrosyan; Andrea Sixto-Costoya: conception and design of study, acquisition of data, analysis and/or interpretation of data, writing-original draft. Juan Carlos Valderrama-Zurián; Rafael Aleixandre-Benavent; Antonia Ferrer-Sapena; Fernanda Peset: conception and design of study, writing-review and editing.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Data availability The raw data used in this study can be found through the following link: https://figshare.com/articles/figure/_strong_The_quality_of_datasets_in_the_area_of_addictions_REPORTS_strong_/23716287

Declarations

Conflict of interest The authors do not have any conflict of interest to declare.

References

Aleixandre-Benavent, R., Vidal-Infer, A., Alonso Arroyo, A., Valderrama Zurián, J. C., Bueno Cañigral, F., & Ferrer Sapena, A. (2014). Public availability of published research data in substance abuse journals. *International Journal of Drug Policy, 25*(6), 1143–1146. https://doi.org/10.1016/j.drugpo.2014.07.007

Clarivate Analytics. (2017). Recommended practices to promote scholarly data citation and tracking: The role of data citation index. https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/2019/08/Crv_WOS_Whitepaper_DCI_web.pdf

Bhatia, K., Tanch, J., Chen, E. S., & Sarkar, I. N. (2020). Applying fair principles to improve data searchability of emergency department datasets: A case study for HCUP-SEDD. *Methods of Information in Medicine, 59*(1), 48–56. https://doi.org/10.1055/s-0040-1712510

Brennan, R. (2017). Challenges for value-driven semantic data quality management. *Proceedings 19th of the International Conference on Enterprise Information Systems*. https://doi.org/10.5220/0006387803850392

Coalition for Advancing Research Assessment (CoARA). (2022). Agreement on reforming research assessment (Issue July). https://coara.eu/app/uploads/2022/09/2022_07_19_rra_agreement_final.pdf

Delgado, J., & Llorente, S. (2022). Fair aspects of a health information protection and management system. *Methods of Information in Medicine, 61*(6), E172–E182. https://doi.org/10.1055/s-0042-1758765

Devaraju, A., & Huber, R. (2021). An automated solution for measuring the progress toward FAIR research data. *Patterns, 2*(11), 100370.

Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., & White, A. (2022). FAIRsFAIR data object assessment metrics (0.5). Zenodo. https://doi.org/10.5281/zenodo.6461229

Díaz, L. A., Idalsoaga, F., Fuentes-López, E., Márquez-Lomas, A., Ramírez, C. A., Roblero, J. P., Araujo, R. C., Higuera-de-la-Tijera, F., Toro, L. G., Pazmiño, G., Montes, P., Hernandez, N., Mendizabal, M., Corsi, O., Ferreccio, C., Lazo, M., Brahmania, M., Singal, A. K., Bataller, R., & Arab, J. P. (2021). Impact of public health policies on alcohol-associated liver disease in Latin America: An ecological multinational study. *Hepatology, 74*(5), 2478–2490. https://doi.org/10.1002/hep.32016

Elsevier. (2019). Research data. https://www.elsevier.com/about/policies/research-data

SPARC Europe. (2021). An analysis of open science policies in Europe, v7 (April). https://zenodo.org/record/4725817

European Commission. (2016). H2020 programme. guidelines on FAIR data management in horizon 2020. July 2016.

European Commission. (2022). Horizon Europe programme guide (February). https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

European Commission. (2024). Horizon Europe (HORIZON). https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/programme-guide_horizon_en.pdf

FAIR data maturity model working group. (2020). FAIR data maturity model. Specification and guidelines (1.0). Zenodo. https://doi.org/10.15497/rda00050

GitHub - pangaea-data-publisher/fuji: FAIRsFAIR Research Data Object Assessment Service. (n.d.). Retrieved July 7, 2022, from https://github.com/pangaea-data-publisher/fuji

Gorman, D. M. (2019). Use of publication procedures to improve research integrity by addiction journals. *Addiction, 114*(8), 1478–1486. https://doi.org/10.1111/add.14604

Gorman, D. M. (2020). availability of research data in high-Impact addiction journals with data sharing policies. *Science and Engineering. Ethics., 26*(3), 1625–1632. https://doi.org/10.1007/s11948-020-00203-7

Govarts, E., Gilles, L., Bopp, S., Holub, P., Matalonga, L., Vermeulen, R., Vrijheid, M., Beltran, S., Hartlev, M., Jones, S., Rodriguez, L., Standaert, M. A., Swertz, M. A., Theunis, J., Trier, X., Vogel, N., Van Espen, K., Remy, S., & Schoeters, G. (2022). Position paper on management of personal data in environment and health research in Europe environ. *Environment International, 165*, 107334. https://doi.org/10.1016/j.envint.2022.107334

Haeusermann, T., Fadda, M., Blasimme, A., Greshake, B., Vayena, E., Haeusermann, T., Fadda, M., Blasimme, A., & Greshake, B. (2018). Genes wide open: Data sharing and the social gradient of genomic privacy. *AJOB Empirical Bioethics, 9*(4), 207–221. https://doi.org/10.1080/23294515.2018.1550123

Holub, P., Kohlmayer, F., Prasser, F., Michaela, T., Mayrhofer, I. S., Martin, G. M., Casati, S., Koumakis, L., Wutte, A., Kozera, Ł, Strapagiel, D., Anton, G., Zanetti, G., Sezerman, O. U., Mendy, M., Valík, D., Lavitrano, M., Dagher, G., Zatloukal, K., … Litton, J.-E. (2018). Enhancing reuse of data and biological material in medical research: From FAIR to FAIR-health. *Biopreservation and Biobanking, 16*(2), 97–105. https://doi.org/10.1089/bio.2017.0110

Huber, R. (2021). FAIRsFAIR data object assessment metrics and F-UJI automated FAIR data assessment tool. Virtual SciDataCon. https://www.youtube.com/watch?v=Tiamzs4f_Ac

British Medical Journal. (n.d.). Article types and preparation. https://www.bmj.com/about-bmj/resources-authors/article-types

Leimapokpam, R. Y., & Swasticharan, R. G. (2017). Compliance of specific provisions of tobacco control law around educational institutions in Delhi, India. *International Journal of Preventive Medicine, 8*, 1–4. https://doi.org/10.4103/ijpvm.IJPVM

Leung, T. I., & Dumontier, M. (2019). FAIR principles for clinical practice guidelines in a learning health system. *Studies in Health Technology and Informatics, 264*, 1690–1691. https://doi.org/10.3233/SHTI190599

Lin, H. C., Wang, Z., Simoni-Wastila, L., Boyd, C., & Buu, A. (2019). Interstate data sharing of prescription drug monitoring programs and associated opioid prescriptions among patients with non-cancer chronic pain. *Preventive Medicine, 118*, 59–65. https://doi.org/10.1016/j.ypmed.2018.10.011

Munafò, M. R. (2016). Opening up addiction science. *Addiction, 111*(3), 387–388. https://doi.org/10.1111/add.13147

National Institutes of Health. (2022). Data management & sharing policy overview. https://sharing.nih.gov/data-management-and-sharing-policy/about-data-management-and-sharing-policy/data-management-and-sharing-policy-overview#before

National Institutes of Health. (2024). NIH. Scientific data sharing. https://sharing.nih.gov/accessing-data

Natsiavas, P., Boyce, R. D., Jaulent, M. C., & Koutkias, V. (2018). OpenPVSignal: Advancing information search, sharing and reuse on pharmacovigilance signals via fair principles and semantic web technologies. *Frontiers in Pharmacology*. https://doi.org/10.3389/fphar.2018.00609

PLOS One. (2019). Data Availability. https://journals.plos.org/plosone/s/data-availability

Pasquetto, I. V., Randles, B. M., & Borgman, C. L. (2017). On the reuse of scientific data. *Data Science Journal, 16*, 1–9. https://doi.org/10.5334/dsj-2017-008

Pereira, S., Gibbs, R. A., & McGuire, A. L. (2014). Open access data sharing in genomic research. *Genes, 5*(3), 739–747. https://doi.org/10.3390/genes5030739

Petrosyan, L., Aleixandre-Benavent, R., Peset, F., Valderrama-Zurián, J. C., Ferrer-Sapena, A., & Sixto-Costoya, A. (2023). FAIR degree assessment in agriculture datasets using the F-UJI tool. *Ecological Informatics*. https://doi.org/10.1016/j.ecoinf.2023.102126

Queralt-Rosinach, N., Kaliyaperumal, R., Bernabé, C. H., Long, Q., Joosten, S. A., van der Wijk, H. J., Flikkenschild, E. L. A., Burger, K., Jacobsen, A., Mons, B., & Roos, M. (2022). Applying the FAIR principles to data in a hospital: Challenges and opportunities in a pandemic. *Journal of Biomedical Semantics, 13*(1), 1–19. https://doi.org/10.1186/s13326-022-00263-7

Ransome, Y., Luan, H., Shi, X., Duncan, D. T., & Subramanian, S. V. (2019). Alcohol outlet density and area-level heavy drinking are independent risk factors for higher alcohol-related complaints. *Journal of Urban Health, 96*(6), 889–901. https://doi.org/10.1007/s11524-018-00327-z

Ross, J. S., Ritchie, J. D., Finn, E., Desai, N. R., Lehman, R. L., Krumholz, H. M., & Gross, C. P. (2016). Data sharing through an NIH central database repository: A cross-sectional survey of BioLINCC users. *British Medical Journal Open, 6*(9), e012769. https://doi.org/10.1136/bmjopen-2016-012769

Scheibein, F., Donnelly, W., & Wells, J. S. (2022). Assessing open science and citizen science in addictions and substance use research: A scoping review. *International Journal of Drug Policy, 100*, 103505. https://doi.org/10.1016/j.drugpo.2021.103505

Schifanella, R., Delle Vedove, D., Salomone, A., Bajardi, P., & Paolotti, D. (2020). Spatial heterogeneity and socioeconomic determinants of opioid prescribing in England between 2015 and 2018. *BMC Medicine*. https://doi.org/10.1186/s12916-020-01575-0

Stopka, T. J., Jacque, E., Kelley, J., Emond, L., Vigroux, K., & Palacios, W. R. (2021). Examining the spatial risk environment tied to the opioid crisis through a unique public health, EMS, and academic research collaborative: Lowell, Massachusetts, 2008–2018. *Preventive Medicine Reports*. https://doi.org/10.1016/j.pmedr.2021.101591

Substance Abuse and Mental Health Services Administration (SAMHSA). (2022). Key substance use and mental health indicators in the United States: Results from the 2018 national survey on drug use and health. https://www.samhsa.gov/data/sites/default/files/reports/rpt39443/2021NSDUHFFRRev010323.pdf

Sun, C., Emonet, V., & Dumontier, M. (2022). A comprehensive comparison of automated FAIRness evaluation tools. *CEUR Workshop Proceedings, 3127*, 44–53.

Van Panhuis, W. G., Cross, A., & Burke, D. S. (2018). Project Tycho 2.0: A repository to improve the integration and reuse of data for global population health. *Journal of the American Medical Informatics Association, 25*, 1608–1617. https://doi.org/10.1093/jamia/ocy123

Vassilakopoulou, P., & Aanestad, M. (2019). Communal data work: Data sharing and re-use in clinical genetics. *Health Informatics Journal, 25*(3), 511–525. https://doi.org/10.1177/1460458219833117

Vidal-Infer, A., Aleixandre-Benavent, R., Lucas-Domínguez, R., & Sixto-Costoya, A. (2019). The availability of raw data in substance abuse scientific journals. *Journal of Substance Use*. https://doi.org/10.1080/14659891.2018.1489905

Wiley. (2023). Wiley's data sharing policies. https://authorservices.wiley.com/author-resources/Journal-Authors/open-access/data-sharing-citation/data-sharing-policy.html

Wilkinson, M. D., Dumontier, M., Aalbersberg, J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., & Mons, B. (2016). Comment: The FAIR guiding principles for scientific data management and stewardship. *Scientific Data, 3*, 1–9. https://doi.org/10.1038/sdata.2016.18

Wilkinson, M. D., Dumontier, M., Sansone, S. A., da Silva, B., Santos, L. O., Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M. E., & Schultes, E. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. *Scientific Data, 6*(1), 1–12. https://doi.org/10.1038/s41597-019-0184-5

Wilkinson, M. D., Sansone, S. A., Schultes, E., Doorn, P., Da Silva Santos, L. O. B., & Dumontier, M. (2018). Comment: A design framework and exemplar metrics for FAIRness. *Scientific Data, 5*, 7–10. https://doi.org/10.1038/sdata.2018.118