frontiers | Frontiers in Drug Discovery

TYPE Mini Review
PUBLISHED 12 July 2023
DOI 10.3389/fddsv.2023.1226727

OPEN ACCESS

EDITED BY
Bruno Villoutreix,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

REVIEWED BY
Rodolpho C. Braga,
InsilicAll, Brazil
Mahesh Koirala,
Akttyva Therapeutics, Inc., United States

*CORRESPONDENCE
Wei Gu,
✉ wei.gu@elixir-luxembourg.org

RECEIVED 22 May 2023
ACCEPTED 30 June 2023
PUBLISHED 12 July 2023

CITATION
Gadiya Y, Ioannidis V, Henderson D, Gribbon P, Rocca-Serra P, Satagopam V, Sansone S-A and Gu W (2023), FAIR data management: what does it mean for drug discovery?
Front. Drug Discov. 3:1226727.
doi: 10.3389/fddsv.2023.1226727

FAIR data management: what does it mean for drug discovery?

Yojana Gadiya[1,2,3], Vassilios Ioannidis[4,5], David Henderson[6], Philip Gribbon[1,2], Philippe Rocca-Serra[7,8], Venkata Satagopam[9], Susanna-Assunta Sansone[7] and Wei Gu[9,10]*

[1]Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Hamburg, Germany, [2]Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Frankfurt, Germany, [3]Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, Bonn, Germany, [4]Vital-IT Group, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland, [5]UNIRIS, University of Lausanne, Lausanne, Switzerland, [6]Bayer AG, Business Development & Licensing & OI, Pharmaceuticals, Berlin, Germany, [7]Oxford e-Research Centre, Department of Engineering Science, University of Oxford, Oxford, United Kingdom, [8]AstraZeneca, Data Office, Data Science and AI Unit R&D, Cambridge, United Kingdom, [9]Luxembourg Centre for Systems Biomedicine, ELIXIR Luxembourg, University of Luxembourg, Esch-sur-Alzette, Luxembourg, [10]Luxembourg National Data Service, Esch-sur-Alzette, Luxembourg

The drug discovery community faces high costs in bringing safe and effective medicines to market, in part due to the rising volume and complexity of data which must be generated during the research and development process. Fully utilising these expensively created experimental and computational data resources has become a key aim of scientists due to the clear imperative to leverage the power of artificial intelligence (AI) and machine learning-based analyses to solve the complex problems inherent in drug discovery. In turn, AI methods heavily rely on the quantity, quality, consistency, and scope of underlying training data. While pre-existing preclinical and clinical data cannot fully replace the need for *de novo* data generation in a project, having access to relevant historical data represents a valuable asset, as its reuse can reduce the need to perform similar experiments, therefore avoiding a "reinventing the wheel" scenario. Unfortunately, most suitable data resources are often archived within institutes, companies, or individual research groups and hence unavailable to the wider community. Hence, enabling the data to be Findable, Accessible, Interoperable, and Reusable (FAIR) is crucial for the wider community of drug discovery and development scientists to learn from the work performed and utilise the findings to enhance comprehension of their own research outcomes. In this mini-review, we elucidate the utility of FAIR data management across the drug discovery pipeline and assess the impact such FAIR data has made on the drug development process.

KEYWORDS

drug discovery, FAIR principles, data management, data sharing, machine learning

Introduction

Ensuring effective exploitation of experimental and computational data resources is a major issue within the drug discovery community, which faces rising costs in bringing safe and effective medicines to market. As part of the search for new medicines, large amounts of data are generated in order to support decision-making on the efficacy, safety, and developability of a potential new drug as it progresses along the discovery pipeline. These new data are generated on a daily basis as a part of *in silico*, laboratory, or clinical studies, and the high cost incurred directly impacts the overall capacity of the pharmaceutical and biotech industries to bring treatments to the clinic. The average cost of research and development (R&D) to bring a new drug to market is estimated to be around 900 million to 2.8 billion dollars (Wouters et al., 2020; Simoens and Huys, 2021). Research expenditure is eventually transferred to the price of treatments and represents a significant part of healthcare spending. To add a further burden, in recent years, the volume and complexity of data generated by scientists involved in research and development have increased exponentially, creating what has been termed a "Big Data" challenge. This has followed the increased adoption of large-scale automated experimentation methods. For example, it is routine to sequence cancer patients' tumour biopsies to identify which specific genetic mutations are associated with their individual tissue malignancies. As part of drug research efforts, these same tumour-derived tissues can then be analysed using powerful high-resolution imaging microscopes to help identify prototype drugs which kill the tumour cells and have the potential to be further developed into new medicines. The challenge scientists now must face in the light of economic constraints is to make the data which has been expensively generated within their studies reusable so that the entire community has the chance to learn from the work performed and, ideally, apply the results to understand the results of their own studies better. It is far more cost-effective to reuse well-validated results from a trusted database rather than repeat the same experimental study again. This situation has led to the previously "un-exciting" process of data management becoming increasingly important in drug discovery, as it directly supports the use of artificial intelligence (AI) and machine learning (ML) based analyses. Such advanced analyses are highly dependent on the quality, consistency, and scope of the training data upon which predictive models are built. In situations where effective data management and quality assessments are not prioritised, then there is a risk of low-quality, poorly controlled or out-of-scope training data emerging, which in the worst case can lead to a counter-productive "garbage-in garbage-out" scenario.

The costs associated with data generation are distributed across the pre-clinical and clinical stages of drug discovery. In the preclinical stage, complex and diverse data are generated, mainly on cellular or *in-vivo* models, to establish the development and toxicity profile of potential drug candidates. In clinical stages, where the major costs of a development programme are incurred, drug candidates are tested for safety and then efficacy in humans, resulting in large amounts of electronic health record-type data. These clinical trial data may be simple numerical results, for example, the level of a diagnostic marker in a blood sample, or highly complex data, which require additional analysis tools such as a low-dose CT image of a patient's lung. Although existing preclinical and clinical data cannot fully replace the need to generate new data in clinical trials, especially when developing a new drug that has not been tested in the clinic before, they are very valuable as they can help to reduce the need to perform redundant research. An additional potential strategy is the usage of "virtual clinical cohorts", created based on information in electronic health records (Tan et al., 2021). Electronically assembled cohorts can act as placebo or control arms in both Phase 2 and 3 trials (wherein the drug is administered to a larger diseased population and observed for long-term effects) creating a situation where all trial participants have the chance to benefit from the therapeutic, as well as reducing the total number of individuals involved. At this point, it is important to highlight that up to 90% of the cost of bringing a drug to market is incurred when conducting clinical trials. In most cases, these cannot be replaced by accessing existing data because the drug being developed is novel and has not been in the clinic previously, rather, the existing data can enable directed decision-making for novel drugs (for, e.g., drugs with active scaffolds). Nevertheless, it has been estimated that the availability of high-quality data could reduce the capitalised R&D costs by about 200 million dollars for each new drug brought to the clinic (Simoens and Huys, 2021). On the other hand, it has been estimated that a high quality data platform in neurology could bring more efficient research and development of new drugs with an annual value of 2.8 billion dollars (https://www.mckinsey.com/industries/life-sciences/our-insights/better-data-for-better-therapies-the-case-for-building-health-data-platforms).

Despite the value represented by large data resources, many are often archived within institutes, companies, or individual research groups and hence effectively unavailable to the wider community. As a consequence, they are in practice "invisible" to the wider community and in some cases even divisions within the same company. This leads to the need for data to be Findable, Accessible, Interoperable, and Reusable (FAIR) (Wilkinson et al., 2016). Each FAIR aspect can be tackled individually. Associating standardised metadata (i.e., information that describes the data) to globally unique and persistent identifiers can then readily ensure the findability of the data it describes. Data needs to be accessible and should be made available via repositories (which are storage spaces for researchers to deposit data sets associated with their research) with a clearly-defined access protocol potentially integrating an authentication and authorisation procedure to control access. Overall FAIR data should be "as open as possible and as close as necessary" (Collins et al., 2018): "open" in order to foster the reusability, or, if relevant, "closed" to safeguard the privacy of the information. This is very important for commercial organisations seeking to generate intellectual property, as they can protect their data and control its sharing for instance during a patent deposition or for collaborations (van Vlijmen, 2020). Similarly, it is important to protect sensitive personal data, such as patients' medical records and to ensure compliance with data protection regulations. Then is the interoperability factor, which involves adopting standards using consistent models, formats, dictionaries (ontologies) and vocabularies for the terms and documentation of the data, including the methods used to generate the data. Several standards exist with their applicability to the Life Sciences (https://fairsharing.org/search?fairsharingRegistry=Standard).
Failure to ensure data are interoperable can lead to extensive time and resource expenditure since additional curation must occur before data can be used. Finally, information about the restrictions defined in consent, local and international laws and rules, or user licences for the data collected ensures that a firm legal framework exists to support the eventual reuse of the data by others. Academic and industry research groups have acknowledged the need to drive reusability and have adopted changes to working practices, for example, collaborating with scientific journals to implement better documentation and deposition of research data in public repositories (McNutt, 2014; van Vlijmen, 2020). Furthermore, pharmaceutical industries have adopted data

standards aligned with FAIR principles to strengthen cross-collaborations with academic and industry partners in the research years. Roche and AstraZeneca have provided a holistic overview of their FAIRification pipelines alongside their downstream impact (Harrow et al., 2022). Despite these efforts, there's still a considerable need to regularly improve the state of FAIR data (Begley and Ioannidis, 2015; Baker, 2016). This simply indicates that FAIR is a journey and needs to be re-visited at specific time points during data evolution to ensure the data follows a FAIR path as addressed by Harrow et al. (2022).

In the following part of this mini-review, we will illustrate with examples the application of FAIR data at various stages within the drug discovery pipeline, starting from the preclinical through to the clinical stages. Beyond these applications, FAIR data is a valuable resource supporting research across multiple scientific and non-scientific fields.

Preclinical applicability of FAIR data

As mentioned, large efforts have been initiated to organise and structure data commonly used in research and development. These involve the establishment of large-scale open-source repositories such as UniProt (UniProt Consortium, 2023) which reports data related to the proteins potentially involved in disease processes, ChEMBL (Gaulton et al., 2012) which includes results on drug-like compounds which are investigated in the early discovery phase, and SureChEMBL (Papadatos et al., 2016) which covers patent-related data. Such repositories serve two main functions within the FAIR context: first, the formalisation of a structure for storing domain-specific information, and second, the open source feature of the repositories allow researchers across the globe to store, access, and interpret the underlying data. As machine-readable and interpretable resources, the data stored in these repositories can become training data for advanced machine algorithms such as artificial intelligence (AI). A compelling example of the impact of data reuse is provided by AlphaFold, an AI model developed by DeepMind (Jumper et al., 2021). The model can predict protein 3-D organisation, thus expanding the repertoire of knowledge from the existing "known" protein structures (which had been solved experimentally) to now include previously "unknown" protein structures. In the drug discovery field, such predictive models play a role in identifying protein-protein and drug-protein interactions that contribute to our understanding of how drugs act at a molecular level. An important aspect of such modelling systems is that they allow computational assessment of the binding efficiency of a molecule to a protein of interest for which an experimentally derived 3-D structure is not available. This can save costs when identifying new compounds which bind proteins and also creates new ways to help understand how the function of the protein can be modulated to change a disease process in a beneficial way. The model owes its success to the presence of open-access and FAIR data repositories and infrastructures. AlphaFold has been trained on data available in UniProt for sequence-based similarity and Protein Data Bank (PDB) for computation of the 3D structure of the model (Berman et al., 2000). Without such repositories supported by machine-interpretable data formats, the training and building of a groundbreaking AI model such as AlphaFold would not have been possible.

It is, unfortunately, the case that only a limited subset of data in the drug discovery field is FAIR and efforts to mobilise the community to implement FAIR-compliant systems need to be initiated (Wise et al., 2019). One prominent effort leading the way in bringing FAIR into practice is the IMI Innovative Medicines Initiative (IMI) FAIRplus project (https://fairplus-project.eu/). FAIRplus was established with the aim to generate reproducible workflows for data FAIRification in the life science field and promoting the FAIR principles among academic and industrial researchers. One project, focussed on reducing drug-associated toxicology, is a useful example of how FAIR data can be leveraged to enable automated downstream tasks. For each potential compound, toxicity data associated with specific chemical structural features can be identified and act as a guide when designing novel compounds with fewer or less acute safety issues. Acknowledging the importance of effectively reusing toxicology data, the project IMI eTOX (http://www.etoxproject.eu/) was established. Within eTOX, a database of preclinical toxicity data from participating pharmaceutical companies was created. After the completion of the project, the FAIR pipelines built by IMI FAIRplus for eTOX were provided to the IMI eTRANSAFE project for further reuse (Custers et al., 2021). Similarly, the IMI CARE project was initiated in response to the COVID-19 pandemic, and as part of the project, ~5,500 FDA-approved drugs and clinical candidates were screened *in vitro* for anti-SARS-CoV-2 activity. Therefore, IMI FAIRplus project assisted in disseminating these data into the ChEMBL public repository (Custers et al., 2022). While these data did not lead to the discovery of an eligible compound for further development to treat COVID-19, they are still very valuable information for informing community-wide COVID-19 drug development efforts. The eTRANSAFE (https://etransafe.eu/) project also developed predictive models for translational clinical research. A common tool, known as FLAME, was published in the project, which reused the bioactivity data within ChEMBL and assisted in activity prediction, specifically toxicity, for compound libraries of interest (Pastor et al., 2021). A key advantage of the tool is its ability to be repurposed for datasets not available in public repositories, such as in-house pharmaceutical company databases (Steger-Hartmann et al., 2018; Sanz et al., 2023). Thus, researchers can re-use the tool for proprietary data by simply harmonising the data format for in-house generated bioassay data to a ChEMBL-compliant format.

FAIR data in clinical studies

During the latter clinical phases of drug development, testing of candidate drugs in patients is done to assess the drug's efficacy for the intended indication. Furthermore, an investigational drug's short- and long-term effects are measured to confirm the safety and tolerability profile of the drug. A recently proposed alternative approach to the design of clinical trials involves generating synthetic patients in the form of virtual cohorts. Such virtual cohorts can represent the diverse human population that differs across ethnicity, anatomy, genetics, environmental, and lifestyle factors, and can be constructed using access to standardised, anonymised FAIR clinical data. Of particular utility is the potential to replace control cohort participants in trials, patients who normally receive a placebo or comparator drug treatments (Azizi et al., 2021). This diverse population representation allows for two significant advantages: first, the ability to evaluate virtually large patient groups irrespective of geographic location or condition; second, it is relatively cost-efficient since analyses are computational in nature.

Two fundamental ingredients are needed to generate useful synthetic data that can mimic the features of a real dataset: advanced algorithms/methods and access to high-quality clinical data and healthcare records. Many ML-based methods have been derived for the method aspect, acknowledging the interest of the drug discovery industry in synthetic patient generators. Models such as Synthea (Walonoski et al., 2018) and SASC (Khorchani et al., 2022) leverage statistical rules defined on real-world healthcare data to generate the synthetic patient cohort. On the other hand, deep neural network-based models like autoencoder-based VAMBN (Gootjes-Dreesbach et al., 2020) or an agent-based simulation model (Popper et al., 2021) have accelerated the field with virtual patient simulation being closer to the real patient. With respect to the data ingredient, resources have been built towards different types of data related to biomedical research. The clinicaltrials.gov is a large open-access database for clinical trial data. The European Health Data & Evidence Network (EHDEN, www.ehden.eu) has built a federated network to enable FAIRness of electronic health record data. A broader list of synthetic data resources has been summarised in the FAIR Cookbook (https://w3id.org/faircookbook/FCB069). Overall, there are ongoing efforts to improve and automate the process of cohort generation, given the benefits which can be accrued in terms of flexibility to share virtual clinical data, lower costs, and reduced data privacy needs relative to real-world clinical data. In summary, it is essential to note that although synthetic data is closed aligned with FAIR principles (given its seamless data sharing and reuse without infringing on privacy), the importance of this data is mainly in building ML/AI algorithms that can mimic real-world scenarios. Consortiums like Common Infrastructure for National Cohorts in Europe, Canada, and Africa (CINECA, https://www.cineca-project.eu/) have aligned their mission in this direction.

FAIR data and drug repurposing

In the scenarios discussed above, we have examined the role played by the analysis of FAIR data in the classical drug discovery process, in which the goal is the identification of new drug candidates for the disease in question. Equally, however, re-use of data can be applied in the search among existing marketed drugs for new therapeutic purposes. This approach is referred to as “drug repurposing” or “repositioning” and is of particular interest in the search for treatment for rare diseases, where the very small number of patients hampers the conduct of clinical trials (Whicher et al., 2018; Pushpakom et al., 2019). The identification of repurposed drugs is supported by resources such as the Drug Repurposing Hub (https://clue.io/repurposing) that comprehensively aggregate pre-clinical and clinical data to assist in decision-making (Corsello et al., 2017). Furthermore, the open resources with curated data generated during pre-clinical and clinical drug discovery pipelines like Open Targets (Koscielny et al., 2017), SureCHEMBL (Papadatos et al., 2016), PubChem (Kim et al., 2016), allow for tools such as Swiss Target Prediction (Gfeller et al., 2014), COVID-19 Pharamcome (Schultz et al., 2021), Patent EnrichMent Tool PEMT (Gadiya et al., 2023a), and others to access, extract, evaluate, and predict patterns in the underlying data. The COVID-19 Pharamcome based approach by Schultz et al. (2021) enabled the integration of existing data (both literature and experimental) on Sars-CoV-2 allowing for the identification of synergistic drug combinations like remdesivir-thioguanosine and nelfinavir-raloxifene. On the other hand, the applicability of the PEMT tool by Gadiya et al. (2023B) focused on retrospective analysis of patent documents to identify the reasoning behind existing drug repurposing cases like Cleave Biosciences’s CB-5083, from cancer to rare diseases, for its target specificity. Both these approaches emphasise the significance of adopting a legacy data perspective to inform future decisions in drug discovery. Furthermore, these endeavours have garnered recognition from European communities resulting in the launch of drug repurposing initiatives such as REPO4EU (https://repo4.eu/) and REMEDI4ALL (https://remedi4all.org/).

Discussion

There is an urgent need to lower the costs and accelerate the process of drug discovery. To help achieve these necessary improvements, access to FAIR data can make a major contribution to community-wide learning of lessons from past failures and successes. FAIR data can also support ML predictions based on well-curated findings from past experiences. The increased adoption of ML methods also drives the further adoption of FAIR principles. FAIR data management involves ensuring that data is easily located, accessible to all who need it (and by machines/automated access and analyses), structured in a way that allows it to be used with other data, and accompanied by sufficient metadata to make it understandable and interpretable. The implementation of FAIR principles in data management comes with an initial cost but has the potential to significantly accelerate scientific discovery by enabling the effective use of data across a range of domains and disciplines. Given the benefits of following the FAIR data principles, it is clear that the effort of making data FAIR is considerable. Any attempt to implement FAIR should be carefully planned, and its benefits should be evaluated prior to starting. Obstacles, as well as potential solutions or strategies to overcome them, have been reviewed in recent works (Gu et al., 2021; Alharbi et al., 2022). Through the journey of making data FAIR, maintaining a close watch on FAIR pipelines’ reusability is always encouraged by FAIR Doers. This has led to the establishment of practical recipes on how to implement FAIR in practices such as the FAIR Cookbook (https://faircookbook.elixir-europe.org, Rocca-Serra et al., 2023) created by biopharmaceutical and academic professionals and guidance on data management practices, such as the RDMKit (https://rdmkit.elixir-europe.org/); both community-driven resources welcome contributions from knowledgeable individuals to share examples and showcase resources that help researchers in their FAIR journey.

Author contributions

YG, VI, DH, PG, and WG contributed to the conception, refinement of the framework and wrote the manuscript. PR-S, VS, and S-AS, along with the other authors, critically revised the paper for intellectual content and approved the final version of the manuscript. All authors contributed to the article and approved the submitted version.

Funding

The authors received funding from the Innovative Medicines Initiative Joint Undertaking (FAIRplus project, grant agreement number IMI 802750).

Acknowledgments

The authors thank reviewers for their comments and suggestions on the manuscript.

Conflict of interest

The authors PR-S, VPS, S-AS, and WG declared that they were editorial board members of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision. Author DH was employed by the Company Bayer AG. Author PR-S was employed by the company AstraZeneca, Data Office, Data Science and AI Unit R&D.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Alharbi, E., Gadiya, Y., Henderson, D., Zaliani, A., Delfin-Rossaro, A., Cambon-Thomsen, A., et al. (2022). Selection of data sets for FAIRification in drug discovery and development: Which, why, and how? *Drug Discov. today* 27, 2080–2085. doi:10.1016/j.drudis.2022.05.010

Alharbi, E., Skeva, R., Juty, N., Jay, C., and Goble, C. (2021). Exploring the current practices, costs and benefits of FAIR implementation in pharmaceutical research and development: A qualitative interview study. *Data Intell.* 3 (4), 507–527. doi:10.1162/dint_a_00109

Azizi, Z., Zheng, C., Mosquera, L., Pilote, L., and El Emam, K.GOING-FWD Collaborators (2021). Can synthetic data be a proxy for real clinical trial data? A validation study. *BMJ open* 11 (4), e043497. doi:10.1136/bmjopen-2020-043497

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature* 533 (7604), 452–454. doi:10.1038/533452a

Begley, C. G., and Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation Res.* 116 (1), 116–126. doi:10.1161/CIRCRESAHA.114.303819

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., et al. (2000). The protein data bank. *Nucleic acids Res.* 28 (1), 235–242. doi:10.1093/nar/28.1.235

Collins, S., Genova, F., Harrower, N., Hodson, S., Jones, S., Laaksonen, L., et al. (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. Available at: https://op.europa.eu/s/yBY0.

Corsello, S. M., Bittker, J. A., Liu, Z., Gould, J., McCarren, P., Hirschman, J. E., et al. (2017). The drug repurposing Hub: A next-generation drug library and information resource. *Nat. Med.* 23 (4), 405–408. doi:10.1038/nm.4306

Custers, Ilse, Boutsma, Erwin, Boiten, Jan-Willem, Duyndam, Alexander, Xu, Fuqi, Juty, Nick, et al. (2022). FAIRplus use case IMI CARE: Quick-response COVID-19 effort opens FAIR data on ~5,500 compounds. *Zenodo*. doi:10.5281/zenodo.7441699

Custers, I., Weitenberg, E., Duyndam, A., Boiten, J. W., Pérez, S., XèniaWillighagen, E., et al. (2021). FAIRplus: eTOX case study - opening up toxicology data about candidate drugs. *Zenodo*. doi:10.5281/zenodo.5786675

Gadiya, Y., Gribbon, P., Hofmann-Apitius, M., and Zaliani, A. (2023b). Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artif. Intell. Life Sci.* 3, 100069. doi:10.1016/j.ailsci.2023.100069

Gadiya, Y., Zaliani, A., Gribbon, P., and Hofmann-Apitius, M. (2023a). Pemt: A patent enrichment tool for drug discovery. *Bioinformatics* 39 (1), btac716. doi:10.1093/bioinformatics/btac716

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., et al. (2012). ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic acids Res.* 40 (D1), D1100–D1107. doi:10.1093/nar/gkr777

Gfeller, D., Grosdidier, A., Wirth, M., Daina, A., Michielin, O., and Zoete, V. (2014). SwissTargetPrediction: A web server for target prediction of bioactive small molecules. *Nucleic acids Res.* 42 (W1), W32–W38. doi:10.1093/nar/gku293

Gootjes-Dreesbach, L., Sood, M., Sahay, A., Hofmann-Apitius, M., and Fröhlich, H. (2020). Variational autoencoder modular Bayesian networks for simulation of heterogeneous clinical study data. *Front. big Data* 3, 16. doi:10.3389/fdata.2020.00016

Gu, W., Hasan, S., Rocca-Serra, P., and Satagopam, V. P. (2021). Road to effective data curation for translational research. *Drug Discov. Today* 26 (3), 626–630. doi:10.1016/j.drudis.2020.12.007

Harrow, I., Balakrishnan, R., McGinty, H. K., Plasterer, T., and Romacker, M. (2022). Maximizing data value for biopharma through FAIR and quality implementation: FAIR plus Q. *Drug Discov. Today* 27 (5), 1441–1447. doi:10.1016/j.drudis.2022.01.006

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* 596 (7873), 583–589. doi:10.1038/s41586-021-03819-2

Khorchani, T., Gadiya, Y., Witt, G., Lanzillotta, D., Claussen, C., and Zaliani, A. (2022). Sasc: A simple approach to synthetic cohorts for generating longitudinal observational patient cohorts from COVID-19 clinical data. *Patterns* 3 (4), 100453. doi:10.1016/j.patter.2022.100453

Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., et al. (2016). PubChem substance and compound databases. *Nucleic acids Res.* 44 (D1), D1202–D1213. doi:10.1093/nar/gkv951

Koscielny, G., An, P., Carvalho-Silva, D., Cham, J. A., Fumis, L., Gasparyan, R., et al. (2017). Open targets: A platform for therapeutic target identification and validation. *Nucleic acids Res.* 45 (D1), D985–D994. doi:10.1093/nar/gkw1055

McNutt, M. (2014). Journals unite for reproducibility. *Science* 346 (6210), 679. doi:10.1126/science.aaa1724

Papadatos, G., Davies, M., Dedman, N., Chambers, J., Gaulton, A., Siddle, J., et al. (2016). SureChEMBL: A large-scale, chemically annotated patent document database. *Nucleic acids Res.* 44 (D1), D1220–D1228. doi:10.1093/nar/gkv1253

Pastor, M., Gómez-Tamayo, J. C., and Sanz, F. (2021). Flame: An open source framework for model development, hosting, and usage in production environments. *J. Cheminformatics* 13 (1), 31–15. doi:10.1186/s13321-021-00509-z

Popper, N., Zechmeister, M., Brunmeir, D., Rippinger, C., Weibrecht, N., Urach, C., et al. (2021). Synthetic reproduction and augmentation of COVID-19 case reporting data by agent-based simulation. *Data Sci. J.* 20 (1), 16. doi:10.5334/dsj-2021-016

Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., et al. (2019). Drug repurposing: Progress, challenges and recommendations. *Nat. Rev. Drug Discov.* 18 (1), 41–58. doi:10.1038/nrd.2018.168

Rocca-Serra, P., Gu, W., Ioannidis, V., Abbassi-Daloii, T., Capella-Gutierrez, S., Chandramouliswaran, I., et al. (2023). The FAIR Cookbook - the essential resource for and by FAIR doers. *Sci. data* 10, 292. doi:10.1038/s41597-023-02166-3

Sanz, F., Pognan, F., Steger-Hartmann, T., Díaz, C., Asakura, S., Amberg, A., et al. (2023). eTRANSAFE: data science to empower translational safety assessment. *Nat. Rev. Drug Discov*. doi:10.1038/d41573-023-00099-5

Schultz, B., Zaliani, A., Ebeling, C., Reinshagen, J., Bojkova, D., Lage-Rupprecht, V., et al. (2021). A method for the rational selection of drug repurposing candidates from multimodal knowledge harmonization. *Sci. Rep.* 11 (1), 11049. doi:10.1038/s41598-021-90296-2

Simoens, S., and Huys, I. (2021). R&D costs of new medicines: A landscape analysis. *Front. Med.* 8, 760762. doi:10.3389/fmed.2021.760762

Steger-Hartmann, T., and Pognan, F. (2018). Improving the safety assessment of chemicals and drug candidates by the integration of bioinformatics and chemoinformatics data. *Basic & Clin. Pharmacol. Toxicol.* 123, 29–36. doi:10.1111/bcpt.12956

Tan, K., Bryan, J., Segal, B., Bellomo, L., Nussbaum, N., Tucker, M., et al. (2022). Emulating control arms for cancer clinical trials using external cohorts created from electronic health record-derived real-world data. *Clin. Pharmacol. Ther.* 111 (1), 168–178. doi:10.1002/cpt.2351

The UniProt Consortium (2023). UniProt: The universal protein knowledgebase in 2023. *Nucleic Acids Res.* 51 (D1), D523–D531. doi:10.1093/nar/gkac1052

van Vlijmen, H., Mons, A., Waalkens, A., Franke, W., Baak, A., Ruiter, G., et al. (2020). The need of industry to go FAIR. *Data Intell.* 2 (1-2), 276–284. doi:10.1162/dint_a_00050

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., et al. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *J. Am. Med. Inf. Assoc.* 25 (3), 230–238. doi:10.1093/jamia/ocx079

Whicher, D., Philbin, S., and Aronson, N. (2018). An overview of the impact of rare disease characteristics on research methodology. *Orphanet J. rare Dis.* 13 (1), 14–12. doi:10.1186/s13023-017-0755-5

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Sci. data* 3 (1), 160018–160019. doi:10.1038/sdata.2016.18

Wise, J., de Barron, A. G., Splendiani, A., Balali-Mood, B., Vasant, D., Little, E., et al. (2019). Implementation and relevance of FAIR data principles in biopharmaceutical R&D. *Drug Discov. today* 24 (4), 933–938. doi:10.1016/j.drudis.2019.01.008

Wouters, O. J., McKee, M., and Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009-2018. *Jama* 323 (9), 844–853. doi:10.1001/jama.2020.1166