NEURAL REGENERATION RESEARCH
www.nrronline.org

# The big data challenge – and how polypharmacology supports the translation from pre-clinical research into clinical use against neurodegenerative diseases and beyond

**Sven Marcel Stefan**[*,#], **Muhammad Rafehi**[*,#]



**Introductory comments:** The identification and validation of disease-modifying proteins are fundamental aspects in drug development. However, the multifactority of neurodegenerative diseases poses a real challenge for targeted therapies. Furthermore, the behavior of individually (over-)expressed target proteins *in vitro* is likely to differ from their actual functional behavior when embedded in cascades and pathways *in vivo*. Increased compartmentalization, e.g., in the brain, adds to the complexity.

More fundamental problems arise from the use of historical data acquired by others years or even decades before with, back then, different perspectives and assumptions. Researchers from different parts of the world of varying disciplines and educational backgrounds investigate different aspects of the same neurodegenerative disease using different techniques. Despite the unambiguous importance of data diversity, this decentralized and competing research gives rise to numerous obstacles that fundamentally impact the quality and quantity of shared heterogeneous scientific data that we would like to address in this perspective, and how we envision polypharmacology as a solution for many obstacles in the field of neurodegenerative diseases.

**The data bias:** ***experimental obstacles:*** The analysis of individual proteins is an important cornerstone of drug development. However, as no standardized procedures or language in any field of biotechnology, molecular pharmacology, or medicinal chemistry exist, experimental setups may differ in many assay parameters (Stefan et al., 2022). Greater complexity occurs in *in vivo* experiments, which are more commonly applied in neurodegeneration research (Möhle and Stefan et al., 2023; Wu et al., 2022). Here, data depends additionally on the disease model, treatment window, way of application, endpoints, or manner and quality of histological data to support hypotheses and (neuro-) pathological observations [e.g., amyloid-beta in Alzheimer's disease models (Wu et al., 2022) or muHTT in Huntington's disease models (Möhle and Stefam et al., 2023) in organ-specific tissues].

Most *in vivo* experiments are conducted in species other than humans. If not "humanized", all data generated is basically connected to the protein ortholog (and associated, potentially species-specific cascades and pathways) only. Surprisingly, in neurodegeneration research, even *in vitro* approaches are based on the use of non-human cell lines (e.g., cortical/striatal neurons or astrocytes) from animal disease models (Wu et al., 2022). Eventually, (species-specific) polymorphisms may challenge the overall outcome and interpretation of data (Matthaei et al., 2021) as may also the individual personality of the species.

***Journalistic obstacles – a researcher's perspective:*** Even if standardized assay procedures existed, results would vary due to the (in)voluntary personal input of the conducting researchers. Personal circumstances and the (in part) toxic work culture in science add to the pressure (Kucirkova, 2023). Language barriers may impede correct conveyance of scientific content and reproducibility by inaccurately described assay procedures. Author services exist, however, they require a fee that may be unaffordable for many groups.

For data analysis, sophisticated software exists, however, license restrictions may cause research groups to use outdated versions of programs or revert to less suitable alternatives that negatively impact the published outcome. Additionally, a thorough understanding of statistics is important, specifically in *in vivo* neurodegeneration research.

Regarding data interpretation, researchers are often enticed to explain *in vivo* effects by the relatively simple, single-targeted mode of action from previous *in vitro* experiments. However, as multi-target drugs are a large fraction of drugs passing clinical trials (Anighoro et al., 2014), which is particularly true for central nervous system drugs, the speculation about single-targeted modes-of-action also adds to the publication bias.

Finally, researchers' intentions are of high importance. A pressured researcher who desperately needs "good data" to attract funding will likely be more "optimistic" in data interpretation. Assay procedures may not entirely be described to actively prevent others from repeating experiments. Data falsification, fabrication, and plagiarism distort the "big picture" of published data. However, the awareness of such data through paper retraction and punishment is almost invisible until today (Hesselmann et al., 2017).

***Journalistic obstacles – a reviewer's perspective:*** One major assignment of journals is the evaluation of the goodness of data, which is acknowledged by the peer review process. The reviewers should be experts in that particular field who take their time to evaluate the goodness with utmost objectiveness. However, this system faces problems today: (i) as reviewers are themselves researchers under constant pressure to publish high-quality and -quantity, the willingness to review has decreased; (ii) in response, the journals consider reviewers whose research field may not suitably match; (iii) although security mechanisms exist (e.g., double-blind peer review), it is often still possible to identify authors from, for example, the research topic, funding statement, or the cited references, and a reviewer may not declare a conflict of interest and review the respective manuscript with personal intentions; or (iv) reviewers may be chosen by the editorial office to favor or discriminate against authors.

***Journalistic obstacles – a journal's perspective:*** The vast majority of journals are owned by publishers with commercial interests that compete with other journals for publicity, reputation, and impact, which is associated with "best", state-of-the-art, and ground-breaking research. To ensure scientific quality, many journals define scientific standards that go along with critical, field-specific aspects that need to be met before publication of an article. However, these standards must not be confounded with general scientific standards, which do not exist, resulting in (i) contradicting experimental requirements; (ii) unconsidered, but actually required standards. Both aspects are selectors for "preferential" data.

The strong demand for journals for originality is understandable. However, the confirmation of published data by other groups increases the overall confidence of the data generated (and potentially used later on). Particularly in *in vivo* neurodegeneration research, statistical significance is harder to obtain. The strong discouragement of redundancy by journals as well as the widespread lack of interest in negative data are major impediments to the trustworthiness of publicly available data.

***The big data generation, storage, extraction, and usage problems:*** The list of obstacles in data generation is very long and the individual errors add up to a distorted picture that can barely be corrected afterwards, as the original parameters of generation are unknown to the public readership.

In light of technical advancement, it became easier to generate more data *in vitro* or *in silico* in shorter time frames (e.g., proteomics; Halder and Drummond, 2024). This fact is in principle favorable, as more valuable data can be generated saving precious resources. However, "big blocks" of more or less homogeneous data supersede the current pool of historical and heterogeneous data compiled over decades. The homogeneity of new data conveys a feeling of confidence but threatens the overall data diversity.

The next obstacle is how data is presented and made accessible to the public. The journals' web pages hinder large-scale searches for key terms to gather published knowledge. Repositories like PubMed or Google Scholar and the use of standardized medical subject headings (MeSH) may help to condense the desired information. However, MeSH and keywords are solely at librarians' and authors' discretion and searches still require manual collection, interpretation, and curation of data – processes that are prone to human errors, distorting the resulting "big picture" from the very start. Big databases exist which provide large datasets (e.g., PubChem). However, these databases work in principle on a one target-one compound basis, meaning they associate one molecule of interest with one particular target of interest only. Smaller web pages with interconnected data emerged recently, but these are at a very early stage (tiny amounts of data stored and searchable), mostly unknown to the public, and thus, not used on a broad scale. The format in which data should be stored is undefined, and even false data will inevitably be stored forever, contributing to "data pollution".

Through trained algorithms (e.g., machine learning, neural networks, artificial intelligence, etc.), ultra-large datasets can be analyzed, interpreted, and novel, ultra-large amounts of data can be generated. Journals favor publications including these techniques, which led not only (i) to the development and evolution of these techniques; but also (ii) to discrimination of other publications with similar or even greater importance. Trained algorithms and computer-aided data extraction and analyses are of great support to handle vastly growing, heterogeneous, and in large parts noisy data. However, particularly artificial intelligence is also a threat as the way data is extracted, analyzed, and generated remains a black box. Thus, (i) novel data could be completely incorrect; or (ii) data could intentionally be falsified on a large scale. The creation of smart algorithms entirely depends on the skills and intentions of the programmers and the (also noisy) input data, and thus, its use strongly affects general scientific credibility and public acceptance.

Nevertheless, it should also be acknowledged that computational workflows have been demonstrated to correctly predict outcomes by

the use of heterogeneous and noisy data – proving that the "data barrier" can indeed be overcome by thorough curation, interpretation, and evaluation of big data (Namasivayam et al., 2022). In summary, big data generation, storage, extraction, and usage determine the applicability domain of the very same data itself.

***Compromised and prevented data:*** Recently, an article claimed that "diversity of workforce", particularly of "preferential" researchers, negatively impacted scientific output. The article has meanwhile been retracted, however, it has caused strong indignation in the scientific communities. Although data heterogeneity indeed poses an obstacle in data evaluation as stated above, and different people will inevitably produce different, sometimes inconsistent data, the widespread discrimination of minorities based on their cultural, religious, racial, social, marital, familial, health, political or any other kind of "status" leads to a bad work environment and negative impact on the quality of data output ("compromised data"), adding to the data bias. Moreover, the systematic exclusion of these people and disrespect of current challenges in gender equality, inclusion, diversity, and discrimination will essentially prevent the generation of potentially very good data. This "prevented data" fails to rectify historical data, and thus, indirectly contributes to the data bias.

**The translation problem – Why are so many drug candidates unsuccessful?** The historical data on shortlisted (pre-)clinical candidates is disillusioning. In Huntington's disease, for example, hundreds of small molecules that showed promising results *in vitro* have failed *in vivo* (Wu et al., 2022). The reasons could be (i) incorrect/incomplete assumptions deduced from biased data; (ii) a discrepancy between the setups of *in vitro* and *in vivo* experiments, in which the first do not mirror the physiological reality of the latter (Stefan, 2019); and (iii) false emphasis on single-targeted approaches in a multifactorial concert of sophisticated feedback mechanisms of (redundant) cascades and pathways.

**Polypharmacology – One solution to multiple problems:** Large-scale, poly-targeted *in vitro* assessment of drug candidates, even at an early stage in the drug development pipeline, would tremendously boost our understanding of the network of targets they address *in vivo* and additionally add valuable, new information to data space. Polypharmacology will extend opportunity space for the druggability of yet undruggable, orphan targets embedded in (redundant) cascades, pathways, and networks in neurodegeneration and beyond (Stefan et al., 2020, 2023). In addition, the intentional engagement of multiple targets as a therapeutic strategy emerged over the last two decades, which has special implications in neural regeneration and neurodegenerative diseases (Al-Ali et al., 2016). Considering the multi-targeted central nervous system drugs approved on health markets (e.g., neuroleptics or antidepressants), polypharmacology seems suitable to tackle (yet untreatable) neurodegenerative diseases. A wide acceptance of polypharmacology as a valid strategy including multiple-track approaches and diversity-based data generation will project its positive impact toward the current obstacles of biased, big, compromised, and prevented data, creating a supportive, inclusive, and open-minded research environment.

**Concluding remarks:** The largest part of this perspective has been dedicated to the big data challenge and the multifactority of publicly available data upon which all assumptions and knowledge of neurodegenerative diseases relies on. Polypharmacology is a new strategy to gain more, diverse data to complement the "big picture" of health and disease in both humans and other species. We suggest a change in research culture and politics to overcome information barriers and propose the following aspects to be widely implemented in global research groups:

(i) Redundant data. Originality is important, but cross-validation by independently repeated (alternative) experiments and confirmation (or refutation) of existing results is vital as it increases the overall confidence of the respective data and rectifies historical data. Journals could implement such reports in a novel format (e.g., "data validation" or "data correction"), which could tackle the problem of "data pollution" by simply incorrect data that otherwise will be stored forever without correction or opposition.

(ii) Negative data. Data that does not prove a hypothesis is widely rejected, which causes one of the largest biases there are. However, particularly computational models and their applicability domain rely on negative references (Namasivayam et al., 2022). Allowing negative (and redundant) data to be published could create a counter-weight to the today easily produced (digital) "big data" that supersedes historical data.

(iii) Diverse data. Concentrating the focus of limited funds on specific aspects of diseases is important. However, it will inevitably lead to a narrow view of the "big picture". Journals should encourage additional and supplementary data even if it may not be in line with the golden thread of the main publication. Reviewers should not criticize such data as being "too much" or "too different", as it may become an important puzzle piece in future science. Furthermore, diverse data is the prerequisite for drug (and target) repurposing strategies.

(iv) Promoted data. The exclusion of minorities and people with personal constraints from scientific participation adds to the "compromised data" and "prevented data" biases. Gaining these people in scientific communities as a positive workforce by support adapted to their individual needs will ultimately promote the generation of additional high-quality data that may rectify historical data.

(v) Joint data. Not only poly-targeted data within one group is important, but also between groups. Assessment of the entire proteome is yet impossible as (i) over 98% of the disease-modifying proteome cannot be targeted to this date; (ii) establishing and maintenance of diverse protocols to various targets is very costly and requires advanced laboratory logistics; and (iii) trained personnel embedded in these logistics will be hard to retain as their great diversity of skills will attract other groups and drive their career. Implementing a diverse (and redundant) research culture in international collaboration with interdisciplinary expertise is vital and needs to be globally supported without objections, addressing not only the biological activity of compounds, but also associated physicochemistry, which is particularly important in neurodegeneration research (Namasivayam and Stefan et al., 2022).

Open-mindedness toward redundant, negative, diverse, promoted, and joint data in combination with historical data could generate novel annotations of drugs with various biological effects and targets that could be harnessed to cure neurodegenerative (and other) diseases with real clinical breakthroughs.

*This work was supported by the Walter Benjamin and Research Grant Programs of the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG, Germany; #446812474, #504079349 [PANABC]) (to SMS); the DFG (#437446827) and the Research Program of the University Medical Center Göttingen (to MR).*

**Sven Marcel Stefan**[*,#]
**Muhammad Rafehi**[*,#]
Drug Development and Chemical Biology, Lübeck Institute of Experimental Dermatology (LIED), University of Lübeck and University Medical Center Schleswig-Holstein, Lübeck, Germany; Department of Pathology, Section of Neuropathology, Translational Neurodegeneration Research and Neuropathology Lab, University of Oslo and Oslo University Hospital, Oslo, Norway; School of Medical Sciences, Faculty of Medicine and Health, University of Sydney, Sydney, NSW, Australia (Stefan SM)
Institute of Clinical Pharmacology, University Medical Center Göttingen, Göttingen, Germany; Department of Medical Education, Augsburg University Medicine, Augsburg, Germany (Rafehi M)
***Correspondence to:** Sven Marcel Stefan, PhD, svenmarcel.stefan@uksh.de; Muhammad Rafehi, PhD, muhammad.rafehi@med.uni-goettingen.de.
https://orcid.org/0000-0002-2048-8598 (Sven Marcel Stefan)
https://orcid.org/0000-0002-4314-4800 (Muhammad Rafehi)
#Both authors contributed equally to this work.
**Date of submission:** July 31, 2023
**Date of decision:** September 6, 2023
**Date of acceptance:** September 23, 2023
**Date of web publication:** November 8, 2023

**https://doi.org/10.4103/1673-5374.387984**
**How to cite this article:** *Stefan SM, Rafehi M (2024) The big data challenge – and how polypharmacology supports the translation from pre-clinical research into clinical use against neurodegenerative diseases and beyond. Neural Regen Res 19(8):1647-1648.*


## References

Al-Ali H, Bixby JL, Lemmon VP (2016) Exploiting kinase polypharmacology for nerve regeneration. Neural Regen Res 11:71-72.
Anighoro A, Bajorath J, Rastelli G (2014) Polypharmacology: challenges and opportunities in drug discovery. J Med Chem 57:7874-7887.
Halder A, Drummond E (2024) Strategies for translating proteomics discoveries into drug discovery for dementia. Neural Regen Res 19:132-139.
Hesselmann F, Graf V, Schmidt M, Reinhart M (2017) The visibility of scientific misconduct: a review of the literature on retracted journal articles. Curr Sociol 65:814-845.
Kucirkova NI (2023) Academia's culture of overwork almost broke me, so I'm working to undo it. Nature 614:9.
Matthaei J, Brockmoller J, Steimer W, Pischa K, Leucht S, Kullmann M, Jensen O, Ouethy T, Tzvetkov MV, Rafehi M (2021) Effects of genetic polymorphism in CYP2D6, CYP2C19, and the organic cation transporter OCT1 on amitriptyline pharmacokinetics in healthy volunteers and depressive disorder patients. Front Pharmacol 12:688950.
Möhle L, Stefan K, Bascuñana P, Brackhan M, Brüning T, Eiriz I, El Menuawy A, van Genderen S, Santos-García I, Górska AM, Villa M, Wu J, Stefan SM, Pahnke J (2023) ABC transporter C1 prevents dimethyl fumarate from targeting Alzheimer's disease. Biology (Basel) 12:932.
Namasivayam V, Stefan K, Silbermann K, Pahnke J, Wiese M, Stefan SM (2022a) Structural feature-driven pattern analysis for multitarget modulator landscapes. Bioinformatics 38:1385-1392.
Namasivayam V, Stefan K, Gorecki L, Korabecny J, Soukup O, Jansson PJ, Pahnke J, Stefan SM (2022b) Physicochemistry shapes bioactivity landscape of pan-ABC transporter modulators: anchor point for innovative Alzheimer's disease therapeutics. Int J Biol Macromol 217:775-791.
Stefan K, Wen Leck LY, Namasivayam V, Bascuñana P, Huang MLH, Riss PJ, Pahnke J, Jansson PJ, Stefan SM (2020) Vesicular ATP-binding cassette transporters in human disease: relevant aspects of their organization for future drug development. Future Drug Discov 2:FDD51.
Stefan SM (2019) Multi-target ABC transporter modulators: what next and where to go? Future Med Chem 11:2353-2358.
Stefan SM, Jansson PJ, Pahnke J, Namasivayam V (2022) A curated binary pattern multitarget dataset of focused ATP-binding cassette transporter inhibitors. Sci Data 9:446.
Stefan SM, Rafehi M (2023) Medicinal polypharmacology – exploration and exploitation of the polypharmacolome in modern drug development. Drug Dev Res doi: 10.1002/ddr.22125.
Wu J, Mohle L, Bruning T, Eiriz I, Rafehi M, Stefan K, Stefan SM, Pahnke J (2022) A novel Huntington's disease assessment platform to support future drug discovery and development. Int J Mol Sci 23:14763.

*C-Editors: Zhao M, Sun Y, Qiu Y; T-Editor: Jia Y*