



THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e. g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

- This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
- A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
- This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
- The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
- When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Curating and summarising data from animal research on neurodevelopment at scale



THE UNIVERSITY
of EDINBURGH

Emma Wilson

PhD Thesis

The University of Edinburgh

2025

Declaration

I declare that this thesis has been composed by myself and that the work has not been submitted for any other degree of professional qualification. Except where stated otherwise or referenced, the work presented is my own. I have not used generative AI in any part of this work, as I believe it inhibits the learning process.

The work presented in Chapter 2 was previously published in *Brain and Neuroscience Advances* as “Genetically modified animals as models of neurodevelopmental conditions: a review of systematic review reporting quality” by Emma Wilson¹, Gillian Currie, Malcolm Robert Macleod², Peter Kind², and Emily Sena². Author contributions are listed in the publication using contributor role taxonomy (CRediT) and detailed at the beginning of the chapter.

In Chapter 3, Peter Kind and Fiona Brown provided subject matter and library science knowledge respectively while developing search strategies; Malcolm Macleod, Sean Smith, Francesca Tinsdeall, and Ian Simpson provided guidance on how best to validate the classifier; Tamsin Baxter, Sarah Bendova, Sarah Giachetti, Chloe Henley, Nawon Kim, Łucja Kostrzewa, Malcolm Macleod, Jessica Pierce, Fiona Ramage, and Eleni Tsoukala provided second and reconciled screening decisions for the dataset used to train the classifier; and Kaitlyn Hair helped me conceptualise how to validate NER approaches. Sena Smith, Kaitlyn Hair, and Maria Economou contributed some of the functions used to retrieve data and produce visualisations used in Chapter 4. And finally, Lucy Pritchard, Chloe Henley, and Francesca Tinsdeall provided second and reconciled screening decisions for my systematic review presented in Chapter 5.

Emma Wilson, 28th March 2025

¹ Thesis author

² Thesis supervisors

Acknowledgements

I would first like to thank my funder, the Simons Initiative for the Developing Brain (part of the Simons Foundation Autism Research Initiative) for making my work possible.

Thank you to Emily Sena, Peter Kind, and Malcolm Macleod for being wonderful supervisors throughout my PhD journey. Your unwavering support for me has helped me grow and find my own confidence. And thank you to the other members of my thesis committee, Emma Wood and Ian Simpson, for being there for me during my PhD, being supportive, and taking the time to answer all my questions.

Thank you to Gillian Currie, Emily Sena, Malcolm Macleod, and Peter Kind for contributing to my first ever Registered Report (Chapter 2).

Thank you to Tamsin Baxter, Sarah Bendova, Sarah Giachetti, Chloe Henley, Nawon Kim, Łucja Kostrzewa, Malcolm Macleod, Jessica Pierce, Fiona Ramage, and Eleni Tsoukala for providing screening decisions to use as training and validation data for the machine learning classifier (Chapter 3). Thank you to Peter Kind and Fiona Brown for helping develop search strategies (Chapter 3).

Thank you to the SOLES Team – Kaitlyn Hair, Sean Smith, Alexandra Bannach-Brown, Maria Economou, Fiona Ramage, Francesca Tinsdeall, Charis Wong, Dasha Gramenitskaya, and Tamires Martins – for collaborating on making systematic online living evidence summaries the best they can be. Special thank you to Sean, Kaitlyn, and Maria for all your contributions to the SOLES code (Chapter 4). Thank you also to the ROpenSci team who maintain many of the R packages used throughout this thesis.

Thank you to Lucy Pritchard, Chloe Henley, and Francesca Tinsdeall for providing screening data for my systematic review (Chapter 5).

Thank you to Brittany Prokop for giving me a tour of the animal house and taking the time to chat to me about reproducibility and open research in animal studies.

Thank you to the University of Edinburgh Open Research Team – Kerry Miller and Nel Coleman – for providing amazing support throughout my PhD and my internship. You always made me feel welcome and I have learned so much from both of you.

Thank you to the Edinburgh CAMARADES Team – Emily Sena, Malcolm Macleod, Gillian Currie, Kaitlyn Hair, Sean Smith, Fiona Ramage, Francesca Tinsdeall, Chris Sena, Olena Maksym, Clara Sanchez-Izquierdo Lozano, and Dasha Gramenitskaya. You are all such warm, friendly, and supportive people to work with. And thank you to the wider CAMARADES members for providing such a wonderful systematic review community.

I want to give a special thank you to the members of West Lothian Writers for all the writing and grammar tips you've given me over the years and for encouraging me to be bold with my writing.

And finally thank you to my Mum and Dad for always believing in me and supporting me no matter what I do.

Abstract

Neurodevelopmental conditions (NDCs) are a heterogeneous group of lifelong disabilities affecting brain development. While many NDCs arise from a complex combination of genetic and environmental factors, some NDCs are monogenic; animals with targeted alterations in homologous genes can be used as models to provide a foundational knowledge of the biological mechanisms underlying brain development and avenues for testing potential therapeutics. Complexities in animal research, including construct validity, statistical errors, heterogeneity in experimental design, and threats to internal validity mean, it is important to consider and critique all available evidence. Systematic review is an established evidence synthesis method used to identify, curate, and summarise published research. Systematic reviews of animal research can identify gaps in knowledge, explain heterogeneity in findings, work towards the Replacement, Reduction and Refinement of animals in research (the 3Rs), and be used to make evidence-based decisions. However, the rapid rate of research publication and breadth of research scope limits the feasibility of conducting reviews. Additionally, narrow review topics only give a snapshot of available research. Systematic maps are an alternative approach to evidence synthesis, summarising information at a higher level, allowing broader overviews of research landscapes, and offering a starting point for more in-depth reviews. However, to enable summarisation of large research areas, mapping approaches require streamlined workflows enabling information from various sources to be combined. Additionally – by making use of automation approaches such as code scripts, existing data accessed via application programming interface (API), text mining, and machine learning – we may be able to increase the feasibility of producing systematic maps. In this thesis, I investigate how evidence synthesis methods and automation approaches can be combined to create a systematic map and make the most use out of research using animals to study genetic NDCs. First, I reviewed previously conducted

systematic reviews in this research area to understand their aims and methods, including assessing their reporting quality using a draft extension of the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) guidelines (PRISMA-Pre). Secondly, I developed and validated automation approaches to enable fast identification genetic NDC animal research, including comprehensive search strategies, a machine learning classifier to identify relevant publications, and named entity recognition to extract key details relating to animal models. Thirdly, I used these automation approaches as the foundation for a continually updating systematic map of research in this area, integrating data from additional scientific data sources, including OpenAlex, and using existing software to extract further data including data and code availability statements and reporting of measures to reduce the risk of bias in animal experiments. Finally, I used my systematic map as the starting point for an in-depth systematic review and validated the systematic map as a comprehensive resource by comparing it to manual systematic searching and screening. Through my review, I also identified key action points for authors of animal studies on how experiments could be reported to improve transparency and better enable their use in evidence synthesis. Overall, my work has contributed to the understanding of how evidence synthesis and automation methods can be integrated to summarise evidence and maximise effective use of data from animal experiments to develop our understanding of how genetics may impact neurodevelopment.

Plain language summary

Neurodevelopmental conditions (NDCs) are a diverse group of lifelong disabilities, causing differences in how peoples' brains develop and function. We don't yet understand how most NDCs are caused – although it's likely a mixture of DNA and environment. However, some NDCs, like Fragile X syndrome, are known to have a single gene cause, making them easier to study. Many animals, including rodents, share much of their DNA with humans. Using animals that share genetic alterations like those causing the human conditions, researchers can investigate how the brain develops. However, animal research is highly complex, and differences in the animals used, how experiments are done, and how experiments are analysed, can influence findings. Therefore, it is important to consider all available research, and assess how differences between experiments could impact findings, to make sure the evidence used is trustworthy. Systematic reviews are a well-established research method used to identify, organise, and summarise all available published research. Systematic reviews of animal research can be used to explain potential reasons for differences in results between different experiments, identify gaps in evidence, improve animal welfare, and be used to make evidence-based decisions. In-depth reviews with specific research questions can take months or even years to produce. This means it's not feasible to use systematic review to summarise evidence from research areas where hundreds of new experiments are published each year. Systematic maps are an alternative approach to identifying and summarising evidence from published research that aim instead to give broad overviews of large research areas. Systematic maps make use of digital research methods like computer software and code to help find and summarise information from research quicker. These digital methods cannot fully replace expert systematic reviewers, but can help with simpler tasks such as searching for research and identifying if research publications are relevant or not. However, systematic reviews can use the broad overviews of research

provided by systematic maps as a starting point for conducting more in-depth systematic reviews. In this thesis, I investigate how systematic review methods and digital methods can be used together to summarise and make the most out of research using animals to study genetic NDCs. First, I reviewed previously conducted systematic reviews in this research area to understand their aims and methods, including assessing how transparently they reported the methods they used. Secondly, I developed digital methods to identify genetic NDC animal research, including methods to find and retrieve research publications, software to filter out research publications that might have been picked up by mistake (like clinical research), and software to pick out information about the animals used in experiments (like animal species, what genes are altered, and what tests are used). Using these approaches, I was able to create a database of relevant research that can be updated when new research is published. Thirdly, I explored pre-existing software to build on my database and turn it into a systematic map, including adding information on whether research is published free-to-read, and if researchers report using methods to reduce biases when conducting their experiments. Finally, I used my systematic map as the starting point for an in-depth systematic review. By comparing my systematic map against manual methods of finding relevant research, I was able to test how effective it was as a resource. In my review, I found that generally publications did not contain enough detail on how experiments were conducted to have confidence in their results. Using this finding, I outlined important steps animal researchers can take to improve how they report what they did in their experiments. Overall, my work has contributed to the understanding of how systematic review and digital methods can be integrated to summarise evidence and maximise effective use of data from animal experiments to develop our understanding of how genetics may impact neurodevelopment.

List of publications

Eleftheriou, C., Giachetti, S., Hickson, R., Kamnioti-Dumont, L., Templar, R., Aaltonen, A., Tsoukala, E., Nawon, K., Fryer-Petridis, L., Henley, C., Erdem, C., **Wilson, E.**, Maio, B., Ye, J., Pierce, J., Mazur, K., Landa, L., Petrovic, N., Bendova, S., Woods, H., Rizzi, M., Salazar Sánchez, V., Anstey, N., Asiminas, A., Basu, S., Booker, S., Harris, A., Heyes, S., Jackson, A., Crocker-Buque, A., McMahon, A. C., Till, S. M., Wijetunge, L. S., Wyllie, D. J. A., Abbott, C. M., O’Leary, T., Kind, P. C. (2025). Better statistical reporting does not lead to statistical rigour: lessons from two decades of pseudoreplication in mouse-model studies of neurological disorders. *Molecular autism*. 16(1), 30. <https://doi.org/10.1186/s13229-025-00663-3>

Zieliński, T., Romanowski, A., **Wilson, E.**, Anderson F., Gountouna, E. Mimault, M., Meynert, A., Wallace, E. W. J. (2025). Good Enough Practices in Scientific Computing: A Learning Module for Researchers. *Journal of Open Source Education*. 8(87), 279. <https://doi.org/10.21105/jose.00279>

Kalijusto, H. K., **Wilson, E.**, Fletcher-Watson, S. (2025). Do influential articles on the genetics of autism show evidence of engagement with the autistic community? *American journal of medical genetics part B: neuropsychiatric genetics*. 198(6), 46-61. <https://doi.org/10.1002/ajmg.b.33030>

Rudkin, G., Warrington, K. L., **Wilson, E.**, Iley, B., Zisk, A. H. (2025) Curricula May Not Apply: “Abnormal” Exceptions Speak Back to the Curriculum. *Teaching of Psychology*. <https://doi.org/10.1177/00986283251328986>

Deeb, H., Creasey, S., Lucini de Ugarte, D., Strevens, G., Usman, T., Wong, H. Y., Kutzer M. A. M., **Wilson, E.**, Zieliński, T., Miller, A. J. (2025). The Rise of Open Data Practices Among Bioscientists at the University of Edinburgh. *PLoS One*. 20(7), e0328065. <https://doi.org/10.1371/journal.pone.0328065>

Wilson, E., Currie, G., Macleod, M., Kind, P., & Sena, E. S. (2024). Genetically modified animals as models of neurodevelopmental conditions: A review of systematic review reporting quality. *Brain and neuroscience advances*, 8, 23982128241287279. <https://doi.org/10.1177/23982128241287279>

Hair, K., **Wilson, E.**, Maksym, O., Macleod, M. R., & Sena, E. S. (2024). A Systematic Online Living Evidence Summary of experimental Alzheimer's disease research. *Journal of neuroscience methods*, 409, 110209. <https://doi.org/10.1016/j.jneumeth.2024.110209>

Hirst, T. C., **Wilson, E.**, Browne, D., & Sena, E. S. (2024). A machine learning-assisted systematic review of preclinical glioma modeling: Is practice changing with the times?. *Neuro-oncology advances*, 6(1), vdae193. <https://doi.org/10.1093/noajnl/vdae193>

Hair, K., **Wilson, E.**, Wong, C., Tsang, A., Macleod, M., & Bannach-Brown, A. (2023). Systematic online living evidence summaries: emerging tools to accelerate evidence synthesis. *Clinical science (London, England : 1979)*, 137(10), 773–784. <https://doi.org/10.1042/CS20220494>

Brody, S., Loree, S., Sampson, M., Mensinkai, S., Coffman, J., Mueller, M. H., Askin, N., Hamill, C., **Wilson, E.**, McAteer, M. B., Staines, H., & Best Practices for Searching During Public Health Emergencies Working Group (2023). Searching for evidence in public health emergencies: a white paper of best practices. *Journal of the Medical Library Association : JMLA*, 111(1-2), 566–578. <https://doi.org/10.5195/jmla.2023.1530>

Wilson, E., Ramage, F. J.³, Wever, K. E., Sena, E. S., Macleod, M. R., & Currie, G. L. (2023). Designing, conducting, and reporting reproducible animal experiments. *The Journal of endocrinology*, 258(1), e220330. <https://doi.org/10.1530/JOE-22-0330>

³ Emma Wilson and Fiona Ramage contributed equally and share joint first authorship

Wilson, E., Cruz, F., Maclean, D., Ghanawi, J., McCann, S. K., Brennan, P. M., Liao, J., Sena, E. S., & Macleod, M. (2023). Screening for *in vitro* systematic reviews: a comparison of screening methods and training of a machine learning classifier. *Clinical science (London, England : 1979)*, 137(2), 181–193.

<https://doi.org/10.1042/CS20220594>

Sander, T., Ghanawi, J., **Wilson, E.**, Muhammad, S., Macleod, M., & Kahlert, U. D. (2022). Meta-analysis on reporting practices as a source of heterogeneity in *in vitro* cancer research. *BMJ open science*, 6(1), e100272. <https://doi.org/10.1136/bmjos-2021-100272>

List of presentations

Systematic Online Living Evidence Summaries: a computational workflow for evidence synthesis. Edinburgh Digital Research Conference, Edinburgh, UK. January 2025. Talk. Available at: <https://osf.io/sa4dq>

Finding the evidence in the animal research haystack. Association for Interdisciplinary Meta-Research and Open Science (AIMOS) Conference, Canberra, Australia. November, 2024. Invited Talk. Available at: <https://osf.io/2wrez>

Designing, conducting, and reporting reproducible animal experiments (journal club). Leicester ReproducibiliTea, Leicester, UK. November 2024. Invited Talk. Available at: <https://osf.io/s6yxq>

Finding the evidence in the animal research haystack. Edinburgh Neuroscience Afternoon, Edinburgh, UK. October 2024. Talk. Available at: <https://osf.io/6npf9>

Developing a real-time summary of animal research modelling genetic neurodevelopmental conditions. Simons Initiative for the Developing Brain (SIDB) Retreat, Edinburgh, UK. September 2024. Invited Talk. Available at: <https://osf.io/hgjzs>

Reproducibility in Neuroscience. TIER2 Reproducibility Workshop, Vienna, Austria. June 2024. Invited Talk. Available at: <https://osf.io/w7pnk>

Practical reasons to embrace data sharing. FENS Forum, Vienna, Austria. June 2024. Invited Talk. Available at: <https://osf.io/6ctzr>

Reproducibility in biological research. EASTBIO DTP Annual Symposium, St Andrews, UK. June 2024. Invited Panel Speaker.

Open research training: who, what, how, why, when, and where. Edinburgh Open Research Conference, Edinburgh, UK. May 2024. Poster. Available at: <https://osf.io/952rp>

The Open Science Framework. Edinburgh Neuroscience Day, Edinburgh, UK. April 2024. Invited Talk. Available at: <https://osf.io/mw38k>

OSF: a researcher's perspective. Digital Research in Action, College of Medicine and Veterinary Medicine, University of Edinburgh, Edinburgh, UK. November 2023.

Invited Talk. Available at: <https://osf.io/x7kfy>

Living evidence synthesis of animal neurodevelopmental research. Research Café, University of Edinburgh, Edinburgh, UK. November 2023. Invited Talk. Available at: <https://osf.io/cshd4>

OSF: a researcher's perspective. Introduction to Open Science Framework, Edinburgh University Library, Edinburgh, UK. August 2023. Invited Talk. Available at: <https://osf.io/zp4af/>

Research skills for Nuffield Research Placements. Nuffield Research Placements Induction for Scottish Students, SSERC, Online. May 2023. Invited Talk. Available at: <https://osf.io/95ady>

Managing your systematic review references with R. Evidence Synthesis and Meta-Analysis in R (ESMAR), online. March 2023. Talk. Available at: <https://osf.io/wf4a6>

Investigating how we summarise evidence. It Takes All Kinds of Minds (ITAKOM), Edinburgh, UK. March 2023. Poster. Available at: <https://osf.io/fdeg9>

Developing a real-time summary of animal research modelling neurodevelopmental conditions. Edinburgh Neuroscience Day, Edinburgh, UK. March 2023. Talk. Available at: <https://osf.io/g56dw>

Systematic Online Living Evidence Summaries (SOLES). National Stroke Session, online. November 2022. Invited talk. Available at: <https://osf.io/ws85x>

Developing a living evidence summary of *in vivo* NDC research. Simons Initiative for the Developing Brain (SIDB) Retreat, Edinburgh, UK. September 2022. Poster. Available at: <https://osf.io/mfdpy>

List of acronyms

3ie	International Initiative for Impact Evaluation
3Rs	Replacement, reduction and refinement of animals in research
ADHD	Attention deficit hyperactivity disorder
ADNP	Activity-dependent neuroprotector homeobox
AI	Artificial intelligence
API	Application programming interface
ARRIVE	Animal research: reporting <i>in vivo</i> experiments
ASD	Autism spectrum disorder
AutDB	Autism database
ASySD	Automated systematic review deduplicator
BERT	Bidirectional encoder representations from transformers
C2	Calcium/lipid-binding
CACNA2D3	Calcium channel, voltage-dependent, alpha 2/delta subunit 3
CAMARADES	Collaborative Approach to Meta Analysis and Review of Animal Data from Experimental Studies
CDKL5	Cyclin-dependent kinase-like 5
CEE	Collaboration or Environmental Evidence
CHD8	Chromodomain helicase DNA binding protein 8
CMA	Chromosomal microarray analysis
CONSORT	Consolidated standards of reporting trials

CRedit	Contributor role taxonomy
CSV	Comma separated value
DOI	Digital object identifier
EEG	Electroencephalogram
FAIR	Findable, Accessible, Interoperable, Reusable
FMR1	Fragile X messenger ribonucleoprotein 1
FMRP	Fragile X messenger ribonucleoprotein
FN	False negative
FP	False positive
FXS	Fragile X syndrome
GAP	GTPase-activation protein
GPS	Global positioning system
GRADE	Grading of recommendations assessment, development, and evaluation
HARKing	Hypothesising after results are known
ID	Intellectual disability
KCNMA1	Potassium large conductance calcium-activated channel, subfamily M, alpha member 1
MCC	Matthew's correlation coefficient
MD	Mean difference
MECP2	Methyl CpG binding protein 2
MeSH	Medical subject headings
ML	Machine learning

mTOR	Mechanistic target of rapamycin
NCBI	National Center for Biotechnology Information
NDC	Neurodevelopmental conditions
NER	Named entity recognition
NIH	National Institutes of Health
NLP	Natural language processing
NRXN1	Neurexin 1
OCD	Obsessive compulsive disorder
ODDPub	Open data detection in publications
ORCID	Open Contributor Roles Identifier
OSF	Open Science Framework
PCI-RR	Peer Community in Registered Reports
PBM	PDZ-binding motif
PDF	Portable document format
PICO	Population, intervention, control, outcome
PMC	PubMed Central
PMCID	PubMed Central identifier
PMID	PubMed identifier
PREPARE	Planning Research and Experimental Procedures on Animals: Recommendations for Excellence
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
PROSPERO	International prospective register of systematic reviews
PSD	Postsynaptic density

PSD95	Postsynaptic density protein 95
PTEN	Phosphatase and tensin homolog
RCT	Randomised controlled trial
REML	Restricted maximum likelihood
ROR	Research Organization Registry
RRID	Research resource identifier
RTT	Rett syndrome
SCN2A	Sodium Voltage-Gated Channel Alpha Subunit 2
SFARI	Simons Foundation Autism Research Initiative
SHANK3	SH3 and multiple ankyrin repeat domains 3
SOLES	Systematic online living evidence summaries
SQL	Structured query language
SVM	Support vector machine
SWD	Spike wave discharge
SWiM	Synthesis without meta-analysis
SYNGAP1	Synaptic Ras GTPase-activating protein 1
SYRCLE	Systematic Review Center for Experimentation with Laboratory Animals
SyRF	Systematic Review Facility
TBR1	T-box, brain, 1
TN	True negative
TP	True positive
UBE3A	Ubiquitin-protein ligase E3A

UID	Unique identifier
XML	Extensible markup language

Table of contents

Declaration.....	iii
Acknowledgements	v
Abstract	vii
Plain language summary	ix
List of publications	xi
List of presentations.....	xv
List of acronyms	xvii
Table of contents.....	xxiii
Table of figures	xxviii
Table of tables	xxix
Table of equations.....	xxxi
Chapter 1: Introduction.....	1
1.1 Context	1
1.2 Neurodevelopmental conditions	1
1.2.1 Co-occurring conditions	2
1.2.2 Altered sensory reactivity.....	3
1.2.3 Pain and distress.....	3
1.3 Genetics.....	4
1.3.1 Rare monogenic NDCs.....	4
1.3.2 Identifying genes associated with NDCs	4
1.3.3 Common molecular pathways.....	6
1.4 Researching genetic NDCs	6
1.4.1 Developing animal models.....	7
1.4.2 Research priorities	8
1.4.3 Translation of animal research to the clinic	8
1.5 Reproducibility and replicability in animal research	10
1.6 Factors influencing reproducibility.....	11
1.6.1 Model suitability.....	12
1.6.2 Sample size and pseudoreplication	14

1.6.3 Heterogeneity and generalisability of results.....	15
1.6.4 Internal validity, risk of bias, and transparent reporting.....	16
1.7 Research improvement.....	18
1.8 Evidence synthesis.....	20
1.8.1 Systematic review	20
1.8.2 The benefits of systematic reviews of animal research.....	21
1.8.3 Barriers and limitations to reviewing animal research.....	22
1.8.4 Automating the systematic review process	23
1.8.5 Systematic maps	24
1.9 Aims and objectives.....	25
Chapter 2: Review of systematic review reporting quality.....	27
2.1 Context	27
2.2 Introduction.....	27
2.2.1 Aims.....	29
2.3 Methods.....	29
2.3.1 Protocol registration	29
2.3.2 Bibliographic search	30
2.3.3 Screening	30
2.3.4 Inclusion and exclusion criteria	31
2.3.5 Data extraction.....	32
2.3.6 PROSPERO search	33
2.3.7 Data synthesis	33
2.4 Results	34
2.4.1 Systematic search results.....	34
2.4.2 Characteristics of included reviews	34
2.4.3 Evaluation of reporting quality	39
2.4.4 Identification of ongoing reviews via PROSPERO	44
2.5 Discussion	47
2.5.1 Comparison with results from Hunniford et al. (2021).....	51
2.5.2 Strengths and limitations	53
2.6 Conclusion.....	54
Chapter 3: Systematic approach to identifying research publications	55
3.1 Context	55

3.2 Introduction	55
3.2.1 Aims.....	59
3.3 Methods.....	60
3.3.1 Protocol registration.....	60
3.3.2 Inclusion and exclusion criteria	60
3.3.3 AutDB animal model resource.....	61
3.3.4 Systematic searching	61
3.3.5 Screening publications for relevance	62
3.3.6 Extracting animal model features from texts.....	65
3.4 Results	68
3.4.1 Performance of systematic search	68
3.4.2 Performance of binary classifier for screening	68
3.4.3 Performance of NER for extracting animal model features.....	71
3.5 Discussion	83
3.5.1 Comparison with other approaches	84
3.5.2 Strengths and limitations	85
3.6 Conclusion.....	90
Chapter 4: Automated workflow to summarise research publications.....	91
4.1 Context	91
4.2 Introduction	91
4.2.1 Aims.....	94
4.3 Methods.....	94
4.3.1 Protocol registration.....	94
4.3.2 Use of previous data.....	95
4.3.3 Data storage	95
4.3.4 Bibliographic search	95
4.3.5 Removal of duplicate publications	96
4.3.6 Screening using a trained machine learning classifier	96
4.3.7 Full text document retrieval.....	96
4.3.8 Tagging model, animal, and experimental test details.....	97
4.3.9 Using metadata from OpenAlex	98
4.3.10 Assessing data and code sharing	98
4.3.11 Assessing risk of bias reporting.....	99

4.3.12 Web application	99
4.4 Results	99
4.4.1 The systematic map dataset	100
4.4.2 Model, animal, and experimental test detail tags	101
4.4.3 OpenAlex metadata	102
4.4.4 Data and code sharing practices	111
4.4.5 Risk of bias reporting	113
4.4.6 Web application	115
4.5 Discussion	117
4.5.1 Comparison with other approaches	118
4.5.2 Strengths and limitations	119
4.6 Conclusion	122
Chapter 5: Systematic review of heterogeneity in <i>Syngap1</i> rodent models	123
5.1 Context	123
5.2 Introduction	124
5.2.1 Aims	125
5.3 Methods	125
5.3.1 Protocol registration	125
5.3.2 Inclusion and exclusion criteria	126
5.3.3 Bibliographic search	127
5.3.4 Removal of duplicate publications	127
5.3.5 Full text document retrieval	127
5.3.6 Screening	127
5.3.7 Checking for retractions	128
5.3.8 Data extraction	129
5.3.9 Quantitative data synthesis	130
5.3.10 Data presentation	131
5.3.11 Validation of the systematic map	132
5.4 Results	133
5.4.1 Systematic search results	133
5.4.2 Narrative summary of included publications	134
5.4.3 Quantitative summary of flurothyl-induced seizure data	142
5.4.4 Reporting of measures to reduce risk of bias	145

5.5 Discussion	146
5.5.1 Sources of heterogeneity and experimental reporting	147
5.5.2 Validation of the systematic map	147
5.5.3 Strengths and limitations	148
5.6 Conclusion.....	148
Chapter 6: General discussion and conclusion.....	151
6.1 Introduction	151
6.2 Systematic reviews of fundamental neuroscience research	151
6.3 Curating a corpus of research publications	153
6.3.1 Searching for relevant research publications.....	153
6.3.2 Screening for relevant research publications.....	156
6.3.3 Making use of existing research metadata.....	157
6.3.4 Automated approaches to extracting research metadata	158
6.3.5 Black boxes and transparency.....	160
6.4 Examining research reproducibility and reliability	162
6.5 Building blocks to more sustainable evidence synthesis.....	163
6.6 Conclusion.....	168
Appendix	169
Supplementary materials.....	169
Preregistrations	169
Data and code availability	170
References	171

Table of figures

Figure 1: Adapted PRISMA flow diagram for main review	36
Figure 2: Level of reporting of PRISMA-Pre items within the title section.	39
Figure 3: Level of reporting of PRISMA-Pre items within the introduction section..	40
Figure 4: Level of reporting of PRISMA-Pre items within the methods section.....	42
Figure 5: Level of reporting of PRISMA-Pre items within the results section.	43
Figure 6: Level of reporting of PRISMA-Pre items within the discussion section.	43
Figure 7: Level of reporting of PRISMA-Pre items within other sections.	44
Figure 8: Adapted PRISMA flow diagram for PROSPERO.....	45
Figure 9: Publications included in NDC-SOLES by publication year.	101
Figure 10: Word clouds of terms matched using regex dictionary-based NER.....	102
Figure 11: World map showing institutions producing research.	107
Figure 12: Word clouds of publication discipline.	109
Figure 13: (a) Number of and (b) percent of publications tagged with different publishing models.	110
Figure 14: Publications tagged using ODDPub.	112
Figure 15: Publications tagged using risk of bias tool.	115
Figure 16: Screenshot of NDC-SOLES web application homepage.	116
Figure 17: Screenshots of the (a) basic and (b) advanced search boxes.	116
Figure 18: Screenshot of the available search filters.....	117
Figure 19: Adapted PRISMA flow diagram for systematic review.....	133
Figure 20: Forest plot of quantitative data	144
Figure 21: Traffic light plot of the adapted SYRCLE risk of bias tool.....	146

Table of tables

Table 1: Summary of the 12 systematic reviews included in the main review.....	38
Table 2: Summary of the 19 review registrations identified via PROSPERO.	46
Table 3: Resources for conducting systematic reviews of animal studies.....	50
Table 4: Full labelled dataset split for stratified 5-fold cross validation	64
Table 5: Number of publications assigned to each set	69
Table 6: Classifier performance against the test set.	70
Table 7: Classifier performance against the validation set.	70
Table 8: Initial NER performance at the level of individual annotations.....	74
Table 9: Updated NER performance at the level of individual annotations.	78
Table 10: NER performance where entities only extracted if model gene and species appeared in same sentence.....	79
Table 11: NER performance where entities only extracted if model gene and species appeared in same text format.	80
Table 12: NER performance where entities only extracted if animal sex appeared alongside additional text.	81
Table 13: NER performance where entities only extracted in publications matching model gene and animal species entities on title and abstract text.	81
Table 14: NER performance with multiple constraints.....	82
Table 15: Conditions for recording NER matched tags	97
Table 16: Number of publications matching NER phrases for different details. ...	101
Table 17: Summary of retraction information retrieved from OpenAlex.....	102
Table 18: Reasons for retractions and time to retract.	104
Table 19: Number of publications tagged by language.	105
Table 20: Number of institutions and publications per first author institution continent.	106
Table 21: Number of publications per institution type.	107
Table 22: Top fifteen research funders.	109

Table 23: Number of publications tagged with different publishing models.....	110
Table 24: Number of publications tagged as sharing data and/or code.	112
Table 25: Number of publications tagged as reporting measures to reduce the risk of bias.	113
Table 26: Summary characteristics extracted from included publications.	138
Table 27: Summary of actions, challenges, and barriers.	166

Table of equations

Eq. 1 Recall	59
Eq. 2 Precision.....	59
Eq. 3 Specificity	59
Eq. 4 F1 score	59
Eq. 5 F2 score	59
Eq. 6 Matthew's Correlation Coefficient (MCC)	59

Chapter 1: Introduction

1.1 Context

In this chapter, I introduce neurodevelopmental conditions (NDCs) and how animal models can be used to investigate the biology underlying monogenic NDCs. I discuss areas of research improvement within animal research of NDCs, including model suitability, translation of laboratory findings to the clinic, and experimental reproducibility and replicability. I discuss how evidence synthesis methods can be used to understand reproducibility and inform research improvement, as well as the challenges of traditional systematic review methods. I discuss advances in automation approaches in systematic review and introduce systematic maps as a form of evidence synthesis where automation is increasingly of interest. Finally, I set out the aims of this thesis, to investigate how evidence synthesis methods and automation approaches can be combined to create a systematic map of research using animals to study monogenic NDCs, and to use these methods to inform improvements in experimental design and reporting.

1.2 Neurodevelopmental conditions

Neurodevelopmental conditions (NDCs) are a heterogeneous group of conditions which arise from differences in the development of the nervous system (Thapar et al., 2017). Many conditions — including autism, attention deficit hyperactivity disorder (ADHD), speech and language conditions, specific learning disabilities, epilepsy, schizophrenia, obsessive compulsive disorder (OCD), tics and Tourette syndrome, intellectual disability, genetic syndromes, and congenital brain injuries — can be included under the umbrella of NDCs (Bölte et al., 2023; Thapar et al., 2017). NDCs themselves fall under the larger umbrella of neurodivergence, which

encompasses neurodevelopmental, neurodegenerative, and mental health conditions, and acquired brain injuries.

Autism and intellectual disability, particularly where they co-occur, are the most widely researched NDCs. Autism is characterised by differences in socialising and communicating, repetitive or routinised behaviours, and narrow but typically deep interests (American Psychiatric Association, 2013). Meanwhile, intellectual disability is characterised by difficulties in cognitive and adaptive functioning (American Psychiatric Association, 2013). In the UK, intellectual disability is also commonly referred to as “learning disability”. In this text, I will use the term intellectual disability to avoid confusion with “specific learning disabilities” such as dyslexia and dyspraxia.

NDCs are lifelong, although some, such as schizophrenia and certain forms of epilepsy, may not present until later in life. Estimates of the prevalence of NDCs tend towards the region of 1% for intellectual disability (McKenzie et al., 2016) and 2.8% for autism (Shaw et al., 2023). However, estimates vary and are impacted by inequalities in access to healthcare and diagnosis, socio-economic status, gender-based or racial discrimination, and stigma around certain conditions (Diemer et al., 2022; Scior et al., 2016). Due to these barriers, the true prevalence of NDCs is likely higher than reported estimates.

1.2.1 Co-occurring conditions

It is common for NDCs to co-occur, and neurodivergent people often have multiple diagnoses. For instance, there is a large overlap between autism and ADHD (Antshel & Russo, 2019) and a quarter of people with intellectual disability also have epilepsy (Snoeijen-Schouwenaars et al., 2021). A meta-analysis published in 2023 found that around half of people diagnosed with an NDC also have a congenital heart disease (Gu et al., 2023). Broadly, neurodivergent people also frequently experience gastrointestinal problems (Curtis et al., 2021), mental health problems, and sleep disturbances (Beadle-Brown et al., 2016; Mason et al., 2018).

1.2.2 Altered sensory reactivity

Altered sensory reactivity – including both hypo- and hyper-reactivity, such as tactile defensiveness – is common across many NDCs (Cascio, 2010).

Approximately 90% of autistic people experience altered sensory reactivity (DeBoth & Reynolds, 2017). The biological mechanisms underlying these sensory differences are not yet clear, although there is evidence that they differ across different NDCs. For instance, some studies have suggested that people with ADHD have typical tactile thresholds while autistic people show enhanced sensory perception; however, these studies are limited by small sample sizes (see Cascio (2010)). Altered sensory reactivity may also contribute to sleep problems, behaviour problems, and eating disorders (Cascio, 2010). Touch, taste, and smell hypersensitivities, along with poor interoception (the detection and regulation of internal body functions such as hunger and thirst), can be correlated with an increased risk of developing eating disorders (Nimbley et al., 2022; Trevisan et al., 2021).

1.2.3 Pain and distress

Many of the conditions which co-occur with neurodivergences – such as altered sensory reactivity, poor interoception, gastrointestinal issues, and sleep problems – or specific neurodivergences themselves, such as epilepsy, can cause pain and distress. Many autistic people experience increased or decreased sensitivity to pain, including menstrual pain (Steward et al., 2018; Summers et al., 2017). Pain and distress can also contribute to self-harm and suicide attempts and people who are nonspeaking with limited alternative communication opportunities are more likely to self-harm (Summers et al., 2017). Many sources of distress can be difficult for external observers such as parents, carers, and healthcare workers to perceive unless they can be effectively communicated by the person experiencing them; a recent publication has shown that caregivers may underestimate the sensory sensitivities experienced by autistic adults with intellectual disability (Bradshaw et al., 2024). Additionally, aggressive or injurious behaviours, including self-injurious

repetitive behaviours, are thought to be a response to severe sensory overwhelm or distress (Summers et al., 2017).

1.3 Genetics

Many NDCs have a large genetic component. However, most NDCs do not have a single genetic cause and may arise from multiple genetic alterations that are not pathogenic on their own. This genetic variability of NDCs is thought to account for the heterogeneity in clinical presentations both across and within different NDCs (Parenti et al., 2020). Monogenic NDCs, NDCs with a known single gene cause, are the focus of this thesis.

1.3.1 Rare monogenic NDCs

Monogenic NDCs arise from single gene alterations, often resulting in gain or loss of function of a particular gene. These alterations may be either inherited from parents or occur *de novo* (spontaneously within the sperm or egg). While each known specific genetic alteration is rare, collectively monogenic conditions are more common. Additionally, monogenic NDCs tend to be characterised by a more severe clinical phenotype and almost all people diagnosed with a monogenic NDC have developmental delay or severe intellectual disability. Autistic traits and seizures are also common.

1.3.2 Identifying genes associated with NDCs

Gene panels and chromosomal microarray analysis (CMA) are commonly used to diagnose rare monogenic conditions. Recent advances in next-generation genome sequencing mean that exome or genome sequencing methods are less costly and more feasible to use as diagnostic tests (Chung et al., 2023). A recent meta-analysis concluded that both genomic and exome sequencing can increase the diagnostic rate of rare monogenic conditions by up to three times that of more conventional genetic testing without increasing the rate of variants of uncertain significance (Chung et al., 2023). Published the same year, a large-scale sequencing study of

13,449 families in the UK and Ireland identified 60 new genes which are associated with developmental conditions, leading to new diagnoses for thousands of people (Wright et al., 2023). In 2020, 102 genes with de novo variants associated with autism were recently identified in a 'largest-of-its-kind' study which sequenced the exome (the protein-coding part of the genome) of over 35,500 people, almost 12,000 of whom were autistic (Satterstrom et al., 2020). Below, I will discuss three common monogenic NDCs associated with genes investigated by Satterstrom et al. (2020): Fragile X syndrome, Angelman syndrome, and SCN2A-related disorders.

Fragile X syndrome is an inherited NDC caused by alterations in the *FMR1* (fragile X messenger ribonucleoprotein 1) gene which silences the FMRP protein (fragile X messenger ribonucleoprotein) (Verkerk et al., 1991). The condition is named after the fragile appearance of the long arm of the X chromosome caused by the alteration. Fragile X syndrome is the most common inherited form of intellectual disability, and is more common in boys than girls (Wijetunge et al., 2013). Additionally, boys with the condition are more likely to be diagnosed with intellectual disability than girls (Wijetunge et al., 2013). About a third of people with fragile X syndrome meet autism diagnostic criteria, and other co-occurring conditions include ADHD, anxiety, epilepsy, and sensory hypersensitivities (Wijetunge et al., 2013).

Angelman syndrome is a rare, inherited NDC caused by a loss of function of the *UBE3A* (ubiquitin-protein ligase E3A) gene. People with Angelman syndrome often have, along with several other traits, intellectual disability, ataxia and jerky limb movements, epilepsy, and may be unable to communicate verbally (Buiting et al., 2016). Seizures in Angelman syndrome tend to emerge in infants between 1 and 3 years old, show a distinct EEG (electroencephalogram) pattern, and may be treated with anti-epileptic drugs including valproate, clonazepam, topiramate, lamotrigine and levetiracetam (Buiting et al., 2016).

Finally, SCN2A (Sodium channel protein type 2 subunit alpha)-related disorders arise from de novo alterations to the *SCN2A* gene, resulting in either loss or gain of function of the gene. The *SCN2A* gene is responsible for encoding the neuronal

sodium channel NaV1.2 which in turn plays a critical role in action potential initiation and propagation (Sanders et al., 2018). SCN2A-related disorders are characterised by intellectual disability, autism, and epilepsy (Sanders et al., 2018). Epilepsy onset typically occurs before the first year of life and tends to be drug resistant (Sanders et al., 2018).

1.3.3 Common molecular pathways

Despite this genetic variability, a small number of converging molecular pathways are thought to be involved in most known NDCs; this may account for how often NDCs co-occur (Parenti et al., 2020). Common pathways include those involved with synaptic function and dysfunction, chromatin remodelling, and mTOR (mechanistic target of rapamycin) (Parenti et al., 2020). Additionally, many of the most common single gene causes of autism and intellectual disability include alterations of genes involved in synaptic function (Zoghbi & Bear, 2012). Therefore, despite the heterogeneity of NDCs, a collaborative approach to understanding underlying biology may be key to making discoveries and developing therapeutics.

1.4 Researching genetic NDCs

Once a gene associated with a monogenic NDCs has been identified, the gene's homolog can be altered in a non-human animal species and used to study the NDC. Animal models can be used to understand the molecular, cellular, and circuitry pathways underlying NDCs, in the hopes that this knowledge can help advance our understanding of the causes of phenotypes seen in people with these conditions. Gaining an understanding of the biology underlying NDCs can then highlight therapeutic targets. This can allow novel drug treatments or gene therapies to be tested in animal models and successful drugs may be taken forward to clinical trials.

1.4.1 Developing animal models

Monogenic NDCs offer a unique opportunity to develop animal models with similar genetic alterations to those seen in human patients. Additionally, and with an understanding of the biological pathways involved, such genetic models can be used as an entry point to identify further associated genes, as was done by Vernes et al. (2008) in the case of developmental language disorders.

The first animal models with single gene alterations associated with autism were developed using mice in the early-1990s (The Dutch-Belgian Fragile X Consortium, 1994), almost a decade before the mouse genome was sequenced in 2002. Mice are used extensively to generate monogenic NDC models. More recently, rats are gaining popularity as they are thought to display more complex social behaviours that better recapitulate social differences associated with NDCs (Silverman et al., 2022) and more easily form bonds with the researchers handling them (Sohn, 2017).

Researchers have also developed models in other species, including zebrafish, *Drosophila* (fruit flies), and *Caenorhabditis elegans* (roundworms). Research using these species has been instrumental in investigating gene functions, assessing if certain genetic alterations may be lethal, and providing insights into neural circuits (DeWeerd, 2021). These models can also help understand social and sensory behaviours; fruit flies have been shown to form social networks, fish shown to synchronise their swimming patterns with other fish of the same species, and worms produce chemical signals to help other worms locate food (DeWeerd, 2021).

Overall, the choice of animal model needs to be matched with the biological question being asked in the research. The adage “all models are wrong”, attributed to statistician George Box (Box, 1976), refers to the fact that a model will never be “correct” but can still give valuable insights if applied and interpreted with good judgement and caution.

1.4.2 Research priorities

The heterogeneity of conditions which typically co-occur within NDCs, such as epilepsy and altered sensory reactivity, present a plethora of potential biological targets for fundamental, preclinical and clinical research (Natri et al., 2023). Historically, within autism research particularly, most biomedical research focuses on social and behaviour differences, with little research into how physical health may be impacted by conditions which co-occur with autism, such as gastrointestinal problems, or how problems in physical health may lead to mental health or behaviour problems (Natri et al., 2023). Epilepsy appears to be well researched within the field (Natri et al., 2023) and the topic of sleep problems in NDCs and sensory processing are gaining traction, potentially indicating a change of research focus towards co-occurring conditions (Heraty et al., 2023).

1.4.3 Translation of animal research to the clinic

To date, few findings from monogenic NDC animal research have been successfully translated to the clinic. Several factors may contribute to these clinical trial failures, including issues with clinical trials themselves, or weak evidence from the animal studies which have informed the trial. Such translational failure has been investigated in many human conditions, but is most documented within ischaemic stroke research. A systematic review published in 2006 found ineffective prioritisation of candidate drugs may play a role in clinical trial failures (O'Collins et al., 2006). The authors identified 912 drugs that had been tested in animals and 114 that had been tested in humans (O'Collins et al., 2006). Evidence from animal studies showed the drugs tested in humans were no more effective than those not taken forward for human testing and that 51% of the 114 drugs that had been tested in humans were reported in clinical use before any use in ischaemia models of stroke were published (O'Collins et al., 2006). These findings highlight a lack of effective drug prioritisation as a potential cause of translational failure.

A recent randomised controlled clinical trial in young people with fragile X syndrome (NCT03614663) found that treatment with cannabidiol showed no

statistical significance compared to control groups despite animal evidence suggesting its potential as a therapeutic (Berry-Kravis et al., 2022). To my knowledge, no systematic reviews have been conducted investigating the evidence from these animal studies. However, a systematic review of research testing the efficacy of cannabinoids in animal models of injury-related or pathological persistent pain found low reporting of measures to reduce the risk of bias may weaken the quality of evidence from biomedical animal research (Soliman et al., 2021).

Within autism research, intranasal oxytocin treatment has been frequently investigated as a potential modulator of social behaviours. A systematic review of the effect of intranasal oxytocin administration on social and routinised behaviours in autistic people found some evidence that oxytocin modulated social behaviour but inconclusive evidence on its effect on routinised behaviours (Audunsdottir et al., 2024). A previous narrative review discussing the pharmacokinetics of oxytocin suggested there is little evidence that oxytocin can cross the blood-brain barrier, therefore it is doubtful that oxytocin can effectively modulate if administered intranasally (Leng & Ludwig, 2016). There are several reasons why claims to the contrary may make it into the scientific literature. (Audunsdottir et al., 2024) suggest publication bias (where the findings of published results differ significantly from results left unpublished, for instance where only results showing a positive effect of a drug are published and results showing inefficacy remain unpublished) as a potential cause, and recommend preregistration as a solution. These views are echoed by Leng & Ludwig (2016). However, preregistration does not address the root cause of the inconclusive results. Poor reproducibility in clinical studies and biomedicine is well documented and may be influenced by experimental design and conduct, selective outcome reporting, or researcher biases (Leng & Ludwig, 2016). Differences in the population being studied may also contribute to irreproducibility; the studies included in Audunsdottir's review assessed mainly male participants and no participants had intellectual disability, indicating the current literature does not fully represent the heterogeneity of the autistic population (Audunsdottir et al.,

2024). Additionally, in their earlier review, (Leng & Ludwig, 2016) highlighted the subjectivity in assessing behaviours as outcome measures, and highlighted one particular publication that found that despite intranasal oxytocin showing no impact, “caregivers who believed their children received oxytocin reported greater improvements compared to caregivers who believed their child received placebo” (Guastella et al., 2015). As social and repetitive behaviours are externally assessed with no biological markers, observer biases may be influencing results.

1.5 Reproducibility and replicability in animal research

Over the years, researchers have generated many models of NDCs in a variety of animal species. However, conflicting results often emerge from different models or from different labs working with the same model. Researchers working with mice face challenges including inconsistent behaviours observed in the same model across different labs and quick rate of drug metabolism impacting the reliability of experiments measuring drug effectiveness and toxicity (Sohn, 2017). Additionally, as the popularity of rat models has grown, so has their use in validating the potential robustness of findings previously observed in mice. Speaking to *Spectrum Autism Research News* (now renamed *The Transmitter*), Professor Peter Kind from the University of Edinburgh suggested: “If it doesn’t translate from a mouse to a rat, is it likely to translate through to humans?” (Sohn, 2017).

Irreproducibility of research findings and failure to translate results from animals to humans is a concern echoed throughout biomedical sciences (Cobey et al., 2024). There are many definitions of reproducibility in research, varying in complexity (summarised in Wilson, Ramage, et al. (2023)), with many being contradictory between disciplines and often being used interchangeably (Fidler & Wilcox, 2025). In this thesis, I define an experiment as “replicable” if sufficient detail is given that biological procedures and analyses could be re-done by an

independent researcher and “reproducible” if consistent results are obtained when the experiment is replicated. Therefore, using these definitions, an experiment needs to be in-principle replicable – through open research practices such as transparent reporting of procedures, methods, and analyses and sharing of analysis code – for one to determine potential factors influencing reproducibility.

It can be easy to frame discussions about irreproducibility around “bad scientists doing bad science” but that is far from the truth and such mindsets are unhelpful to individual researchers or efforts to improve reproducibility and public trust in science. Professor Marcus Munafò recently described irreproducibility not as a crisis but as an opportunity to encourage continuous improvement within research (Munafò et al., 2022). On a similar vein, Professor Malcolm Macleod has written about the need to differentiate research integrity (influenced by experimental design, lack of generalisability, and false positive results) from researcher integrity (related to deliberate misconduct) (M. Macleod & the University of Edinburgh Research Strategy Group, 2022). Here I will focus my discussions on Macleod’s definition of research integrity, as these are issues where both individual researchers and institutions can make positive steps towards improving. However, I do not discount the severe impact that deliberate fraud has on research and knowledge generation.

1.6 Factors influencing reproducibility

There is no single cause of irreproducibility within research, instead various factors have been identified which may influence the reproducibility, or lack thereof, of an experiment. These factors include how the experiment has been designed, conducted, analysed, and reported (M. Macleod & the University of Edinburgh Research Strategy Group, 2022). Within animal research, the reproducibility of findings (for instance, the generalisability of findings across different laboratory groups or different animal species or strains) is not the only matter of importance;

results must also translate from animals in the laboratory to humans in clinical settings.

1.6.1 Model suitability

As mentioned earlier with the George Box quote, no model is perfect. However, we base our assumptions that an animal model is a valid, suitable model of a human condition based on at least two factors: construct validity (how well the induction of a model reflects that human condition) and face validity (how well a model looks like the human condition). Most NDCs, including autism, have no identified biomarkers and are clinically diagnosed based on the presence or absence of externally observable features.

Animal models therefore seek to recapitulate observable features. However, not all genetic models recapitulate expected behaviours and underlying biology. For instance, CDKL5 mouse models fail to recapitulate seizures characteristic of CDKL5 deficiency disorder (Fallah & Eubanks, 2020). While this makes mouse models unsuitable for investigating seizures in CDKL5, it instead offers a unique opportunity to explore how gene and protein functions and pathways may have evolved between species.

Other behaviours must be assumed; in rodent models for instance, “excessive” self-grooming is thought to mimic restrictive behaviours and decreased vocalisations in adulthood is thought to mimic social differences. This makes autism research using animals particularly reliant on face validity, that is the degree to which phenotypes observed in animals “look like” those seen in humans with the condition (Silverman et al., 2022). However, as our perception of face validity in autism is built on the externally observed and deficit-focused view of autism, which has been criticised by many autistic autism researchers (Botha & Cage, 2022; Heraty et al., 2023; Natri et al., 2023), many of the advancements in our clinical understanding of autism may not be translated back to animal research.

Due to the limitations in face validity, construct validity is more important and should bear more weight than face validity alone (Silverman et al., 2022). While the heterogeneity and multi-faceted nature of autism and other NDCs can prove challenging, focusing on models of “high confidence” gene alterations could be key to ensuring construct validity (Silverman et al., 2022). Currently, only a portion of self-described autism mouse models are based on high confidence gene alterations, with many induced by low confidence gene alterations or environmental exposure (Silverman et al., 2022).

Another key aspect that can hinder translation of animal research to the clinic, even in high confidence genetic models, are genetic differences between animals and humans. Although humans and rodents share many of the same genes, a fact that allows the generation of genetic models in the first place, several characteristics of brain development, such as expansion of the cerebral cortex, are unique to humans, limiting the aspects of brain development that can be recapitulated in animal models (Zhao & Bhattacharyya, 2018). While there are only subtle differences between human and animal DNA within protein-coding regions, transcribed non-protein coding sequences show greater variation between species, which can result in phenotypic differences between humans and animals even with the same genetic alteration (Zhao & Bhattacharyya, 2018).

Several *in vitro* models have been argued as alternatives, including human pluripotent stem cells, however, the sufficiency of these models has yet to be more systematically evaluated (Zhao & Bhattacharyya, 2018). As a potential solution, pluripotent stem cell-derived neural progenitors can be implanted into the brains of developing or adult rodents, allowing researchers to study the effects *in vivo* (within the body) (Zhao & Bhattacharyya, 2018). However, while *in vitro* models may help understand genetic and cellular biology, they do not offer the opportunity to study behavioural differences caused by genetic alterations.

1.6.2 Sample size and pseudoreplication

The overuse of animals in research is an animal welfare concern, and often studies use small numbers of animals in efforts to reduce the number of animals required. Regrettably, the use of fewer animals can reduce the robustness of animal studies. In the context of publication bias (where studies with large effects and statistical significance are favoured for publication over studies showing smaller, non-statistically significant results), these small studies can lead to false positives and often irreproducible results (Wilson, Ramage, et al., 2023).

The population size of an experiment (denoted as N) can be described as the total number of experimental units in each experimental group, where an experimental unit is defined as the smallest entity within an experiment which can be subjected to an intervention independent of all other experimental units (Lazic et al., 2018). Correct identification of experimental units is therefore essential to ensure that statistical analyses are performed as intended.

It is important to note that the experimental unit within animal studies is not always each individual animal. For instance, if an intervention (e.g. diet) is introduced to several animals sharing the same cage (e.g. the same food source), then the experimental unit is the cage (Wilson, Ramage, et al., 2023). Similarly, if an intervention is applied maternally and effects are investigated in offspring, the experimental unit is the whole litter (Lazic et al., 2018).

Pseudoreplication occurs when the sample size of an experimental group is artificially inflated due to incorrect identification of the experimental unit, and can lead to false positive results (Lazic et al., 2018). Small sample sizes mean that often, when pseudoreplication is apparent, experimental data cannot simply be reanalysed to correct for the pseudoreplication, as the experiment may be statistically underpowered (Lazic et al., 2018).

Researchers are recommended to perform a sample size calculation prior to performing experiments to ensure they use an appropriate number of animals and improve the reproducibility and reliability of their findings (Lazic et al., 2018).

To ensure transparency, researchers should report how their sample size was calculated, and if a sample size calculation was performed *a priori* (before beginning the experiment) (Percie Du Sert et al., 2020).

Additionally, a phenomenon known as the “litter effect”, whereby animals from the same litter are more phenotypically different (and therefore less heterogeneous) to animals from a different litter, has recently been implicated as a potential factor influencing animal experiments of NDCs (Jiménez & Zylka, 2021). Failure to account for the litter effect in research using a rodent model of CHD8 haploinsufficiency (a high confidence autism gene) has been shown to mask animal phenotypes and contribute to irreproducibility (Jiménez & Zylka, 2021).

1.6.3 Heterogeneity and generalisability of results

As described previously, NDCs are extremely heterogeneous. Even monogenic conditions such as fragile X syndrome can vary greatly in clinical presentation (Zhao & Bhattacharyya, 2018). Heterogeneity can occur within experiments (for instance, differences between individual animals used in an experiment) and between experiments (for instance, differences in the effect size reported from two separately conducted experiments which report to use the same population, methodology, and analysis). High between-experiment heterogeneity, especially where within-experiment heterogeneity is lower, can reduce the generalisability (also known as the external validity) of experimental findings.

There are an abundance of biological factors (e.g. animal sex, age, strain, weight, co-morbidities) and technical factors (e.g. housing conditions, animal handling) which are thought to at least partially account for heterogeneity between animal experiments (Wilson, Ramage, et al., 2023). However, these are often under or unreported in publications, meaning the extent of their impact on reproducibility is unknown (Wilson, Ramage, et al., 2023).

1.6.4 Internal validity, risk of bias, and transparent reporting

The internal validity of an experiment can be described as the likelihood that any observed effects between experimental groups are true effects rather than methodological errors. Researcher bias, including behaviours, preconceptions, or “preferred” outcomes, can negatively impact the internal validity of an experiment (Wilson, Ramage, et al., 2023). It is important to note that these biases are typically subconscious and are not examples of outright fraud or data manipulation.

Reporting of any measures taken to reduce these individual biases is important to ensure transparency and improve the reproducibility of individual studies (Wilson, Ramage, et al., 2023). Even if measures were not taken, the fact and the reason why they were not taken should be reported in publications, as otherwise it is impossible to tell whether measure were taken and not reported or not undertaken at all (Percie Du Sert et al., 2020).

Many individual biases are thought to reduce internal validity. The method used to assign of animals to experimental groups can introduce selection bias, whereby there may be (unseen) systematic differences between control and treatment groups (Van Der Worp et al., 2010). These differences between may later be mistaken for actual differences between intervention or model and control groups (Van Der Worp et al., 2010). Randomly allocating animals to experimental groups, and both reporting that animals are randomly allocated and the method of random allocation used, can be employed to reduce this risk of bias (Percie Du Sert et al., 2020). Additionally, researchers should be unaware of which group animals are being randomly allocated to (allocation concealment, a form of blinding) (Percie Du Sert et al., 2020). Where complete randomisation is impossible, such as where the experimental groups are a genetically modified model versus a wildtype control, animal groups can be counterbalanced to ensure differences between groups (such as age, weight, and sex) are accounted for (Percie Du Sert et al., 2020).

Performance bias is introduced by differences in how animals are handled or cared for (aside from any tested interventions) (Van Der Worp et al., 2010). Similar

to selection bias, blinding (which is also known as masking) and randomisation can be undertaken to reduce the risk of bias. All interactions with animals should be taken in a randomised order across groups, and ideally an alpha-numeric identifier should be used in place of “treatment” and “control” group names to ensure researchers and animal house staff are unaware of experimental group (Wilson, Ramage, et al., 2023). Involvement with animal house staff is essential to ensure that group allocation can remain concealed during the experiment and if handling or treatment requires a member of staff to be aware of an animal’s experimental group, that member of staff should be excluded from other phases of the experiment to avoid bias (Wilson, Ramage, et al., 2023).

During outcome assessment, detection bias can be introduced by differences in how outcomes are assessed between different groups (Van Der Worp et al., 2010). For instance, in an experiment comparing a seizure model with a wildtype rodent where number of seizures are a measured outcome, experimenters may be more likely to attribute subtle, jerky motions to normal grooming behaviour in the wildtype and seizures in the model animals if they are aware of experimental group. Once again, masking can be used to reduce the impact bias on the outcomes measured (Percie Du Sert et al., 2020). In this example, it may be necessary to have a member of animal house staff, who is not involved with the experiment, aware of experimental groups to ensure animal welfare.

Finally, unequal drop-outs between experimental groups, for instance where exclusions are applied differently between groups, can introduce attrition bias (Van Der Worp et al., 2010). Exclusion criteria should ideally be defined *a priori* and exclusion decisions made by animal house staff external to the experiment and unaware of group allocation (Wilson, Ramage, et al., 2023). The exclusion criteria should be reported in publications, even if no exclusions were made (Percie Du Sert et al., 2020).

1.7 Research improvement

The goal of efforts to identify and highlight irreproducibility of research is not to discount the research itself but rather bring forth methods to improve and make the best use of research findings (Munafò et al., 2022).

Several tools have been developed to aid researchers in designing, conducting, and reporting their animal experiments. The ARRIVE (Animal Research: Reporting *In Vivo* Experiments) Guidelines are a set of reporting guidelines developed by animal research experts in 2010 and updated in 2020 (Percie Du Sert et al., 2020). The guidelines list several items, including sample size calculation, randomisation, masking, and exclusion criteria, which should be reported in publications to improve the transparency and reproducibility of animal research (Percie Du Sert et al., 2020). Additionally, the PREPARE (Planning Research and Experimental Procedures on Animals: Recommendations for Excellence) checklist was developed to help researchers in the planning stages of their experiments take into consideration reproducibility and animal welfare (Smith et al., 2018).

Open research practices, which aim to improve the transparency of research methods and results, are gaining traction across research disciplines (Munafò et al., 2022; Nosek et al., 2015). Open research practices, such as materials and methods sharing and data sharing, are necessary to ensure enough detail is provided for an experiment to be replicable, and therefore essential to reproducibility however are not sufficient to ensure reproducibility – experiments can be well reported but may use unsuitable models or statistical methods (Devezer et al., 2021; Eleftheriou et al., 2025). Many funder and publisher policies now require open research practices, including data sharing, as a standard. Guidance on how to adopt open research practices in animal research is available (summarised in Wilson, Ramage et al. (2023)). These include posting completed work as preprints on a preprint sharing platform – of which there are many relevant to biomedical sciences (Kirkham et al., 2020) – and sharing datasets

according to FAIR (Findable, Accessible, Interoperable, Reusable) principles (Wilkinson et al., 2016).

Another key open research practice is preregistration, a practice that while more widely adopted in other areas of research such as psychology and clinical trials, is rarely adopted in animal research. Preregistration of experimental designs, a process by which hypotheses and methods for data collection and analysis are pre-defined, and published or embargoed, prior to starting an experiment, is a common strategy thought to improve reproducibility (by thinking more carefully about experimental designs), reduce HARKing (hypothesising after results are known) and prevent P-hacking (misusing statistics to make non-statistically-significant findings appear significant) (Van Der Naald et al., 2022; Wilson, Ramage, et al., 2023). PreclinicalTrials (Van Der Naald et al., 2022) and The Animal Study Registry (Bert et al., 2019) are two platforms specifically designed for the preregistration of animal studies. It's important to recognise, however that the simple act of preregistering a study design does not make that study reproducible or "high quality", as invalid study designs and analysis plans can be, and are, preregistered (Devezer et al., 2021).

Despite much effort being put into the development of tools to aid the improvement of animal research reproducibility and transparency, uptake and implementation of these tools could be better. Potentially, we could see greater uptake through greater involvement of animal researchers in tool development (Wilson, Ramage, et al., 2023). Efforts to improve the reproducibility of animal research do require often substantial changes to the way researchers plan and conduct experiments, including ensuring that there are sufficient staff available to help with tasks such as masking. It is important therefore that researchers and animal house staff feel that any changes will benefit rather than hinder their work, that they feel supported to make these changes, and that they are involved in the research improvement process (Wilson, Ramage, et al., 2023). Additionally, all-or-nothing approaches to reproducibility are also unlikely to help researchers making

the first steps (Wilson, Ramage, et al., 2023), and improvements to reproducibility should be viewed as continual efforts (Munafò et al., 2022).

1.8 Evidence synthesis

The lack of reproducibility and transparency of individual animal experiments makes them unreliable to use on their own to inform future research direction. By systematically identifying all the publications that conduct the same experiment, we can gain stronger evidence and better understand factors that might influence irreproducibility. When looking at research conducted over time, evidence synthesis can be used to benchmark improvements to research reproducibility and transparency.

Evidence synthesis is a research methodology that follows strict processes of selection, critique, and analysis of relevant publications. These processes separate evidence synthesis from more traditional narrative literature reviews which serve the purpose of summarising key findings in a particular research area without systematic identification strategies or formal critique of the evidence.

1.8.1 Systematic review

There are several methods researchers can use to synthesise evidence, depending on the purpose of the research. One of the most commonly used methods of synthesising evidence in healthcare and life sciences are systematic reviews and meta-analyses (Chalmers & Haynes, 1994; Egger et al., 2001; Higgins et al., 2019; Soliman et al., 2020). Systematic reviews have been used since the 1980s to evaluate evidence from clinical trials (Chalmers & Haynes, 1994; Egger et al., 2001; Higgins et al., 2019), with organisations including Cochrane, JBI (formerly the Joanna Briggs Institute), and the Campbell Collaboration playing key roles in shaping the field (Chalmers, 1993; Pilla et al., 2022). More recently, systematic review methods have been adapted to interrogate evidence from animal research and even *in vitro* research (Sander et al., 2022; Soliman et al., 2020), with much work shaped by the CAMARADES (Collaborative Approach to Meta Analysis and

Review of Animal Data from Experimental Studies) and SYRCLE (Systematic Review Center for Experimentation with Laboratory Animals) research groups (Soliman et al., 2020). Systematic review methods and guidance has also been adapted for and is used in other research domains including ecology and evolutionary sciences by organisations such as the Collaboration for Environmental Evidence (CEE) (<https://environmentalevidence.org/>).

Systematic reviews are a structured research method, based on a pre-specified protocol, used to systematically identify and evaluate qualitative evidence, including experiment quality measures such as reporting of methods to reduce the risk of bias, related to a pre-specified research question, in an unbiased manner (Chalmers & Haynes, 1994; Egger et al., 2001; Higgins et al., 2019; Soliman et al., 2020). Meta-analyses are optional to systematic reviews and involve quantitative evaluation of effect sizes (Chalmers & Haynes, 1994; Egger et al., 2001; Higgins et al., 2019; Soliman et al., 2020).

1.8.2 The benefits of systematic reviews of animal research

Animal systematic reviews can be used to inform improvements to experimental design and reporting which can aid the translatability of results and improve overall reproducibility (Sena et al., 2014). Systematic reviews are recommended to identify existing evidence and knowledge gaps within research prior to grant applications, and the findings of reviews are used to inform policy and funding decisions (Soliman et al., 2020). Several systematic reviews have highlighted ethical considerations relating to animal models used and outcomes measures, the findings of which can improve animal welfare and the reduction, refinement, and replacement (3Rs) of animals in research (Ritskes-Hoitinga & Van Luijk, 2019). Systematic reviews can also inform and benchmark research improvement; a 2016 update of a systematic review originally published in 2009 investigating the efficacy of interleukin-1 receptor antagonist in animal models of focal cerebral ischaemia found that reporting of measures to reduce the risk of bias had greatly improved (McCann et al., 2016). Systematic evaluation of research evidence can also help more reliably identify and prioritise candidate drugs most likely to

succeed in clinical trials; this approach has been implemented in the MND-SMART (Motor Neuron Disease-Systematic Multi-arm Adaptive Randomised Trial; NCT04302870) (Wong et al., 2023).

1.8.3 Barriers and limitations to reviewing animal research

Systematic reviews of animal data are generally considered more complex than those of clinical studies, due to the heterogeneity of animal experiment designs, lack of methods reporting, smaller sample sizes per experiment, and larger numbers of individual studies compared to clinical reviews (Bannach-Brown et al., 2021). Guidance on how to best conduct preclinical animal systematic reviews has been rigorously adapted from clinical review guidance (Sena et al., 2014; Soliman et al., 2020), however the complexity of animal research still imposes barriers on the feasibility of conducting reviews.

Systematic reviews typically focus on narrow research questions with specific defined populations (e.g., animal models and/or species), interventions or experimental designs, and outcome measurements. These defined criteria are then translated into search terms to identify publications from bibliographic databases, typically relying on search title, abstract, and keyword text. Unlike clinical trials, which are typically published individually, several animal experiments are typically included in the same publication, limiting the amount of detail that appears in the title and abstract of the publication. There is growing evidence that this approach to identifying studies may not capture all relevant evidence, with reviews in Alzheimer's research estimating only 69-82% of literature was identified through title and abstract searching (Hair, 2022).

Increasing publication rates make it more and more difficult to stay up to date with the literature that we can detect (Bannach-Brown et al., 2021; Hair, Wilson, et al., 2023). This challenge is not limited to reviews of animal research and also impacts clinical systematic reviews (Beller et al., 2013; J. H. Elliott et al., 2017). Larger reviews require additional, specialist manual effort to complete and a lack

of collaborative effort to complete reviews further confounds this barrier (Hair, Wilson, et al., 2023).

Finally, systematic reviews are typically static; once they are completed, new research will continue to be published, some of which may conflict with the findings synthesised in the review. Particularly in fields with high publication rates, such as neuroscience, this means reviews are out of date relatively quickly (Hair, Wilson, et al., 2023). Updating out of date systematic reviews is a difficult task, often akin to starting the review from scratch (Beller et al., 2013; J. H. Elliott et al., 2017; Hair, Wilson, et al., 2023).

1.8.4 Automating the systematic review process

There is a pressing need for research decision-makers to have access to up-to-date evidence (Brody et al., 2023; J. Elliott et al., 2021). Recently, meta-researchers have begun to employ automation approaches, including machine learning and text mining, to reduce the manual effort required to conduct systematic reviews (Bannach-Brown et al., 2021; Thomas et al., 2017). For instance, the ASySD (Automated Systematic Review Deduplicator) tool can aid researchers in quickly and accurately removing duplicate copies of publications from their datasets (Hair, Bahor, et al., 2023), and machine learning classifiers can be trained to aid researchers in screening publications for inclusion in systematic reviews (Bannach-Brown et al., 2019; O'Mara-Eves et al., 2015; Shemilt et al., 2022). Additionally, researchers have developed tools to extract data from publications, including reporting of measures to reduce the risk of bias in animal research (Bahor et al., 2017; Wang et al., 2022b) and in clinical trials (I. J. Marshall et al., 2016), population, intervention, control, and outcome (PICO) elements (Wang et al., 2022a), and metadata related to data and code availability (Riedel et al., 2020).

Researchers at the University of Bond in Australia were able to use automation approaches to conduct a systematic review in two weeks (Clark et al., 2020). The effort has since been replicated by another research team (Tufanaru et al., 2024)

and has also been successfully trialled in the context of ten more complex systematic review questions (Scott et al., 2023).

In addition to reducing the time required to conduct a systematic review, such automation approaches are also able to facilitate so-called 'living' systematic reviews: where new information can be continually identified and evaluated (Bannach-Brown et al., 2021; J. H. Elliott et al., 2017; Thomas et al., 2017).

1.8.5 Systematic maps

Unlike systematic reviews, which have a narrow but deep focus, systematic maps are a method of "big picture" evidence synthesis aiming to gain a broader understanding of a research topic or area, identify areas where evidence exists and where there are gaps, and may be used as the starting point for further evidence synthesis such as systematic reviews (F. Campbell et al., 2023; Hair, Wilson, et al., 2023). Evidence mapping techniques as a terminology has existed for at least two decades (Saran & White, 2018), however the approaches this terminology describes are variable and inconsistent, sometimes referring to narrow scoping reviews and other times referring to broad, interactive, web-based evidence gap maps (F. Campbell et al., 2023). In this thesis, the approach I describe as a systematic map is most similar in concept to an evidence gap map, although I was not aware of the concept or terminology when I set out to conduct this work.

Systematic maps offer an exciting avenue for research discovery and may aid the systematic review process. However, as systematic maps seek to explore vast quantities of published research rather than narrow, selected topic areas, they are even more constrained by time and resource requirements than traditional systematic reviews. Therefore, the implementation of automation approaches is essential for their feasibility. Automation approaches, such as the ones mentioned in the sections above, offer a potential to make systematic maps on broad topic areas with high publication rates feasible. As these approaches have been introduced within the last decade, exploration of the feasibility of this approach is relatively novel. Similar work, using machine learning to enhance systematic

mapping, have been conducted in climate research (Berrang-Ford et al., 2021; Callaghan et al., 2021; Lück et al., 2025), however these appear to be static.

1.9 Aims and objectives

The overall aim of this thesis is to investigate how evidence synthesis methods and automation approaches can be combined to create a systematic map of research using animals to study monogenic NDCs and use these methods to inform improvements to experimental design and reporting.

I narrowed the scope of this effort to animal research exploring “high confidence” genes. These are any of the 102 genes identified via exome sequencing by Satterstrom et al. (2020) or the 92 genes from the SFARI (Simons Foundation Autism Research Initiative) animal model list with a human gene score of 1 (SFARI Gene, 2024a). As there is overlap in these two lists, there are 150 genes of interest in total. These genes, and the lists they appear on, are detailed in Appendix 1A.

The objectives of this thesis are to:

1. Understand the aims and reporting quality of previously conducted systematic reviews in this research area;
2. Combine evidence synthesis methods and automation approaches to build a corpus of research literature that can be systematically interrogated and updated as new research is published;
3. Identify features which may contribute to irreproducibility within the research, and;
4. Make recommendations to aid the improvement of reproducibility and reliability of evidence.

In Chapter 2, I began by systematically identifying previously conducted systematic reviews and evaluating their reporting quality to understand the current state of evidence synthesis.

In Chapter 3, I describe my process of developing and validating methods to systematically identify animal research relevant to monogenic NDCs. This included building systematic search strategies to identify potentially relevant research, training a machine learning classifier to score publication relevance, and using named entity recognition to extract animal model details from publication text documents.

Data identified through these processes then formed the foundations for the development of a systematic map, which I describe in Chapter 4. I describe my process of building a workflow of automation approaches to retrieve and extract additional metadata from relevant publications. I present static data from the systematic map to showcase the type of information offered and discuss use cases and limitations.

In Chapter 5, I validated my systematic map as a comprehensive resource by conducting a systematic review. The review question was developed in collaboration with primary neuroscience researchers to ensure it was of relevance to the scientific community and aimed to investigate factors that may influence heterogeneity in *Syngap1* rodent seizure phenotypes.

I end in Chapter 6 with a discussion of the overall findings of my research, a discussion of the opportunities and challenges of summarising evidence from fundamental neuroscience studies at scale, and highlight recommendations for improvement of primary research conduct and reporting, evidence synthesis guidance, and the use of automation tools in summarising research.

Chapter 2: Review of systematic review reporting quality

2.1 Context

I began my research by investigating systematic reviews that had already been conducted and published in this area. The work described in this chapter has been published in *Brain and Neuroscience Advances* as a Registered Report. I (EW) conceptualised the study, conducted bibliographic searches, removed duplicate copies of publications, and retrieved full text documents, screened and extracted data from publications, conducted all analyses, produced all visualisations, and led project administration. Gillian Currie (GC) provided second screening and data extraction, Malcolm Macleod (MM) reconciled screening decisions, and Emily S Sena (ESS) reconciled data extraction. GC, MM, ESS, and Peter Kind (PK) contributed supervision (systematic review and subject domain knowledge) and review and editing of the manuscript for publication. In section 2.5.1 “Comparison with results from Hunniford et al. (2021)”, I present an additional discussion of my results not present in the published version of the article, requested by my thesis examiners.

2.2 Introduction

Up to 40% of neurodevelopmental conditions (NDCs) presenting with intellectual disability (ID), epilepsy, and autism are caused by single genetic alterations, often occurring de novo in that individual (Brunet et al., 2021). These single gene alterations provide an opportunity to develop animal models harbouring those alterations, which may give insights into NDC pathophysiology and treatment

more generally, including NDCs with polygenic or environmental causes. Such models can be used to investigate associated cellular, circuit, and behavioural traits. Researchers have developed numerous genetically modified animal models to study NDCs, and research using these models plays a vital role in developing our understanding of the biology underlying NDCs (basic research) and in the testing of potential drug treatments (preclinical research). However, despite extensive research, successful translation of laboratory findings to the clinic is rare (Pankevich et al., 2013; Silverman et al., 2022).

Systematic review provides a well-developed method for identifying areas of poor methodological quality or high risk of bias within research literature. Systematic review is a research method used to summarise and appraise all available evidence related to a pre-specified topic (Egger et al., 2001) and can identify areas for improvements, which might increase internal validity and research rigour and reproducibility.

Clinicians have used systematic reviews to inform evidence-based healthcare since the 1980s. Findings from such reviews have led to considerable improvements in the way clinical trials are conducted and reported (Plint et al., 2006). Over the last two decades, researchers have adopted systematic review methods to summarise and appraise evidence from laboratory animal studies (De Vries et al., 2014) and have achieved similar success in research improvement (McCann et al., 2016; Ramirez et al., 2017). However, the fundamental differences between clinical and animal studies mean that systematic review methodologies must be appropriately adapted (Hunniford et al., 2021). The development of several tools has been instrumental in supporting researchers conducting systematic reviews and meta-analyses of animal studies, including checklists for assessing the risk of bias (Hooijmans et al., 2014) and reporting quality (M. R. Macleod et al., 2004) of laboratory animal studies, and a protocol template for systematic reviews of animal intervention studies (De Vries et al., 2015). Despite this, reporting quality of animal systematic reviews is low (Hunniford et al., 2021; Mueller et al., 2014). Currently in development is an extension to the Preferred Reporting Items for

Systematic reviews and Meta Analyses (PRISMA) guidelines specifically designed for preclinical systematic reviews (PRISMA-Pre). PRISMA-Pre are a set of reporting guidelines that aim to improve the reporting quality of preclinical systematic reviews (Hunniford et al., 2021).

2.2.1 Aims

Here, we aim to evaluate the quantity, characteristics, and reporting quality of systematic reviews which synthesise research using genetically modified animals to model NDCs. We assessed the reporting quality of included systematic reviews using the PRISMA-Pre checklist. We were interested in (1) models with alterations in any of the 102 high-confidence genes identified via large-scale exome sequencing by Satterstrom et al. (2020) and (2) models of Fragile X syndrome (FXS), Rett syndrome (RTT), and Angelman syndrome which have alterations in the Fmr1, Mecp2, and Ube3a genes, respectively. These additional three genes are not included in the list of 102 genes; however, they are extensively researched monogenic NDCs.

This review does not synthesise evidence from existing reviews. The rationale behind this review is that by identifying the quantity of and assessing the reporting quality of existing systematic reviews in this area, we can inform guidance on how future systematic reviews within this research field should be conducted. Findings from this review have been used to inform the development of a living evidence summary of research using genetically modified animals to model NDCs, a preliminary protocol for which is available on the Open Science Framework (OSF; DOI: <https://doi.org/10.17605/OSF.IO/GFTZP>).

2.3 Methods

2.3.1 Protocol registration

Stage 1 of this Registered Report received peer review via *Peer Community In Registered Reports* (PCI-RR) and is preregistered on OSF (DOI:

<https://doi.org/10.17605/OSF.IO/952QK>). At the time of preregistration, EW had completed preliminary searches and optimised our search strategy to inform the development of our study protocol. Deviations from the protocol are listed in Appendix 2A.

2.3.2 Bibliographic search

On 23 January 2023, EW conducted a systematic literature search on three electronic data sources: PubMed including Medline (accessed via NCBI), Embase (accessed via Ovid), and Web of Science Core Collection. Our search strategy included three components: (1) broad terms related to NDCs, intellectual disability, epilepsy, and autism, and associated genes; (2) terms related to animal models, and (3) terms related to systematic reviews or meta-analyses. Terms used to identify animal models were taken from van der Mierden et al. (2022) and terms used to identify systematic reviews and meta-analyses were taken from Langendam et al. (2021). Full search terms are given in Appendix 2B and the Stage 1 Registered Report. Where citations appeared in multiple databases, EW removed duplicate versions of the citation using the ASySD tool (Hair, Bahor, et al., 2023).

2.3.3 Screening

We uploaded our search results, with duplicate citations removed, to the Systematic Review Facility (SyRF) platform (RRID:SCR_018907; (Bahor et al., 2021)) for screening, data extraction and management of records. SyRF displays records to reviewers in a random order, and reviewers were unaware of the decisions or annotations made by other reviewers, or in the case of reconciliation, unaware of which reviewer made which decision or annotation.

Two independent reviewers (EW and GC) screened each publication for inclusion, and any disagreements were reconciled by a third independent reviewer (MM). We planned one round of screening, where we screened the full texts of all studies retrieved from our searches against our inclusion and exclusion criteria, to avoid potentially excluding systematic reviews where the decision for inclusion rests on

information contained in the full text but not in the abstract (Wilson, Cruz, et al., 2023).

EW retrieved full text PDF files using the find full text feature in EndNote 20 (with institutional subscription) or via hand-searching. Where we could not access the full text using our institutional subscriptions, we requested the full text via interlibrary loan. In total, we used 24 interlibrary loans. Where the full texts of relevant articles were not written in English, we planned to use Google Translate. However, our search did not return any relevant non-English language publications.

2.3.4 Inclusion and exclusion criteria

Studies were screened according to the criteria outlined below:

- Study design – We only included systematic reviews or meta-analyses that include animal studies, either as a review limited to animal studies or those which included them alongside other study types (e.g., clinical studies). We excluded all other study designs.
- Animal models – We only included systematic reviews synthesising research using genetically modified animals to model NDCs where the modified gene appeared on the list of 102 genes identified via large-scale exome sequencing by Satterstrom et al. (2020); or genetically modified animal models of FXS, RTT, or Angelman syndrome; or other genetic models of NDCs characterised by intellectual disability, epilepsy, and autism. Animal models may be of any species. A diverse range of other models of NDCs are available but are not the focus of this review and will not be included.
- Publication type – We included systematic reviews published in (1) peer-reviewed journals, (2) as conference abstracts, or (3) as preprints (where they are identified in searches). We did not search dedicated preprint servers. We placed no restriction on publication date or language.

2.3.5 Data extraction

Two independent reviewers (EW and MM) conducted data extraction.

Discrepancies between reviewer decisions were reconciled by a third independent reviewer (ESS). We carried out data extraction using the SyRF platform and collected the following information:

2.3.5.1 Bibliographic data

We extracted the names of first authors, year of publication, title, and digital object identifier (DOI) of each included review.

2.3.5.2 Characteristics

To understand the purpose and scope of included systematic reviews, we extracted the following characteristics from each:

- The aim of each systematic review and the primary research question each review seeks to ask.
- Whether the review only included animal studies, or also included clinical or *in vitro* studies.
- Which animal models the review included.
- The total number of studies included in the systematic review.
- The total number of studies investigating relevant animal models.

2.3.5.3 Reporting quality

There are two broad approaches to evaluating the quality of systematic reviews: addressing the completeness of reporting or assessing the risks of bias arising from the approaches which were taken. Precise evaluation of the second requires completeness of the first, so these are overlapping but distinct.

Here, we assessed the completeness of reporting (reporting quality) of each included systematic review using the 46-point checklist developed by Hunniford et al. (2021). The checklist is adapted from the PRISMA guidelines for systematic reviews and is more specific to systematic reviews of animal studies. However, the

checklist is not currently an official extension to PRISMA. Although it is not recommended to use the general PRISMA guidelines to assess *in vivo* systematic review reporting quality, the adapted checklist for preclinical systematic reviews has been designed for this purpose. There are no validated tools for assessing risks of bias in systematic reviews of *in vivo* research, and we will not conduct a formal assessment of risk of bias. The checklist items are written in full in Appendix 2C and in the Stage 1 Registered Report.

2.3.6 PROSPERO search

Additionally, we searched the PROSPERO (International prospective register of systematic reviews) database to identify the status of ongoing but unpublished preregistered systematic reviews. On 25 September 2023, we searched PROSPERO using the following terms: neurodevelopment OR neurodevelopmental OR autism OR autistic OR ASD OR intellectual disability OR epilepsy. We limited this search to reviews of animal studies for human health protocols.

Two reviewers (EW and GC) screened each of the search results based on our inclusion criteria, and disagreements were reconciled by a third reviewer (ESS). For each relevant review, EW noted the data of registration, expected start date, expected end date, and current stage of the review. We had planned to contact authors of PROSPERO registrations to establish if their review has been published; however, we did not complete this due to time constraints. We did, however, search for published versions of preregistrations labelled “ongoing” in the PROSPERO system.

2.3.7 Data synthesis

We did not conduct a meta-analysis or perform statistical analyses on our data. We have presented a descriptive summary of the bibliographic, characteristics, and reporting quality data extracted from each included systematic review. We scored each included systematic review using the PRISMA-Pre checklist and have provided summary graphs detailing which items of the checklist each review met. Additionally, we assessed which tools are currently being used to conduct

systematic reviews in this field, including the tools used to screen studies (Checklist Item 6a), extract numerical data (Checklist Item 17a), and measure study quality or risk of bias (Checklist Item 19).

2.4 Results

2.4.1 Systematic search results

Our systematic searches of PubMed, Embase, and Web of Science Core Collection returned a total of 1,753 records (441 from PubMed, 640 from Embase, and 672 from Web of Science Core Collection; Figure 1). 428 records were removed using the ASySD tool, leaving 1,325 unique records remaining. Of these, only twelve records met our inclusion criteria via full text screening. Ten of the included records were peer-reviewed journal articles and two were conference abstracts. We did not identify any preprints through our searches. The same systematic review was identified as a conference abstract (X.-Y. Zhang & Spruyt, 2022) and a peer-reviewed journal article (X. Zhang et al., 2021). As our evaluation is primarily concerned with reporting quality of these publications, we assessed the conference abstract and journal article as separate publications.

2.4.2 Characteristics of included reviews

All the reviews included in this evaluation were published between 2016 and 2022 (Table 1). Most of the reviews, including both conference abstracts, were published in 2022. Only two reviews (a journal article and a conference abstract) conducted a meta-analysis alongside their systematic review. The remaining ten included only qualitative synthesis of evidence. Three of the reviews (all journal articles) included studies with animal or human participants, while the remaining reviews included only animal studies. None included *in vitro* or *in silico* work alongside animal studies. Each of the reviews covered a distinct research question. The review aims and research questions were coded into the following topics: sleep disturbances; microbiome differences; stress (specifically the hypothalamic-pituitary-adrenal

axis); neurobiology; sensory differences (specifically auditory and olfaction); model suitability; and biomarker identification (Table 1). A variety of genetic alterations were included in the reviews evaluated, but this only represented 11 of the total 105 genetic modifications of our interest. Many reviews looked at multiple genes. The genes assessed were *Fmr1* (6 reviews); *Mecp2* (5 reviews); *Ube3a*, *Pten*, and *Shank3* (4 reviews each); *Adnp* (2 reviews); and *Cacna2d3*, *Chd8*, *Kcnma1*, *Nrxn1*, and *Tbr1* (1 review each). *Fmr1*, *Mecp2*, and *Ube3a* alterations are strongly associated with FXS, RTT, and Angelman syndrome, respectively. Additionally, *Pten*, *Shank3*, *Adnp*, and *Nrxn1* are associated with Cowden syndrome, Phelan-McDermid syndrome, Helsmoortel-Van der Aa syndrome, and Pitt-Hopkins-like syndrome 2, respectively (SFARI Gene, 2024b). *Cacna2d3*, *Chd8*, and *Tbr1* are associated with autism and intellectual disability generally but not with any named syndromes. The number of studies included in each review varied greatly, ranging from thirteen studies to 531 (mean = 98, median = 26).

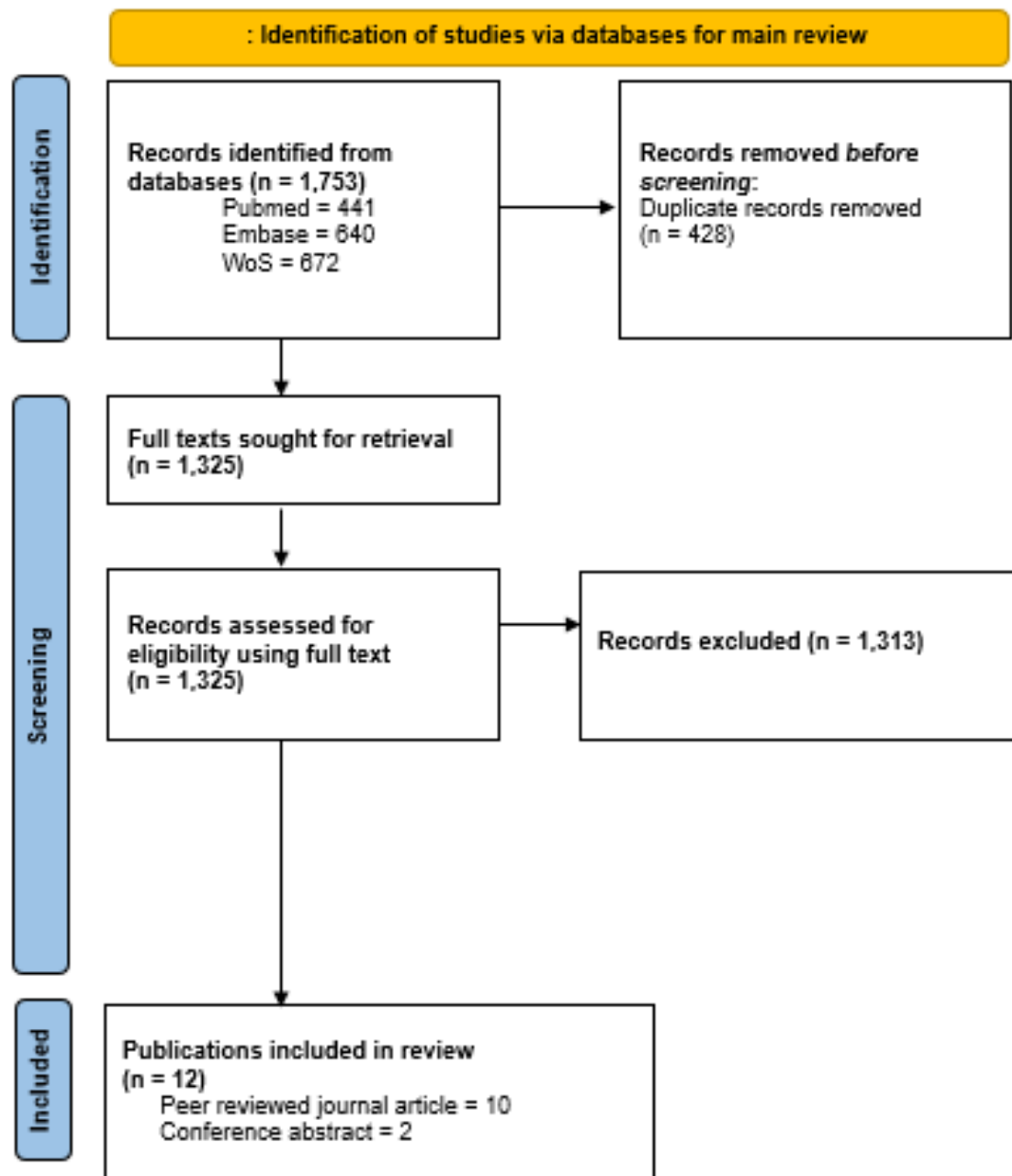


Figure 1: Adapted PRISMA flow diagram for main review.

Diagram templates are shared under a Creative Commons By Attribution (CC-BY) license (<https://creativecommons.org/licenses/by/4.0/>) and available at <https://www.prisma-statement.org/prisma-2020-flow-diagram>.

Publication	Journal	Publication Type	Review Type	Review Focus	Study Types	Gene Models	N Included Studies
(Hardiman & Bratt, 2016)	Physiol Behav	Journal article	SR	Stress	Animal and human studies	Fmr1	17 (9)
(Nakai et al., 2018)	Front Neurosci	Journal article	SR	Neurobiology	Animal studies only	Fmr1; Mecp2; Ube3a	22 (22)
(Kundap et al., 2020)	Pharmaceuticals	Journal article	SR	Model suitability	Animal studies only	Ube3a	23 (23)
(Sysoeva et al., 2020)	Clin Neurophysiol	Journal article	SR	Biomarkers	Animal and human studies	Mecp2	28 (6)
(Lyons-Warren et al., 2021)	Neurosci Biobehav Rev	Journal article	SR	Sensory differences (olfactory)	Animal studies only	Adnp; Nr1h3; Pten; Tbr1	119 (119)
(X. Zhang et al., 2021)*	J Neurosci Res	Journal article	SR	Sleep disturbances	Animal studies only	Mecp2	13 (13)
(Alamoudi et al., 2022)	Front Cell Infect Microbiol	Journal article	SR	Microbiome	Animal and human studies	Shank3	22 (9)

(Kat et al., 2022)	Neurosci Biobehav Rev	Journal article	SR + MA	Model suitability	Animal studies only	Fmr1	266 (266)
(Panzenhagen et al., 2022)	Laboratory Animals	Conference abstract	SR + MA	Model suitability	Animal studies only	Ube3a; Pten; Shank3; Mecp2; Fmr1	531 (531)
(Thawley et al., 2022)	Front Immunol	Journal article	SR	Biomarkers	Animal studies only	Fmr1; Shank3; Pten	28 (28)
(Wilde et al., 2022)	J Neurodev Disord	Journal article	SR	Sensory differences (auditory)	Animal studies only	Adnp; Cacna2d3; Chd8; Fmr1; Mecp2; Pten; Shank3; Kcnma1; Ube3a	88 (88)
(X.-Y. Zhang & Spruyt, 2022)*	Sleep Medicine	Conference abstract	SR	Sleep disturbances	Animal studies only	Mecp2	13 (13)

*Table 1: Summary of the 12 systematic reviews included in the main review. SR = Systematic review. MA = Meta-analysis. N included studies shows the total number of studies included in each review and the total number of animal studies (in brackets). * indicates journal article and conference abstract version of the same review.*

2.4.3 Evaluation of reporting quality

The PRISMA-Pre checklist items are in six categories divided by the section of a manuscript which they refer to: title, introduction, methods, results, discussion, and other. A PRISMA-Pre checklist for abstracts does not yet exist, so we were mindful when evaluating the two conference abstracts and have marked reporting as not applicable for items where reporting would be unfeasible, for instance, the inclusion of a PRISMA flow diagram or summary table of included studies.

2.4.3.1 Items related to reporting in the title

Eleven out of the twelve reviews identified the report as a systematic review, and eleven identified that the report contained animal data (Figure 2).

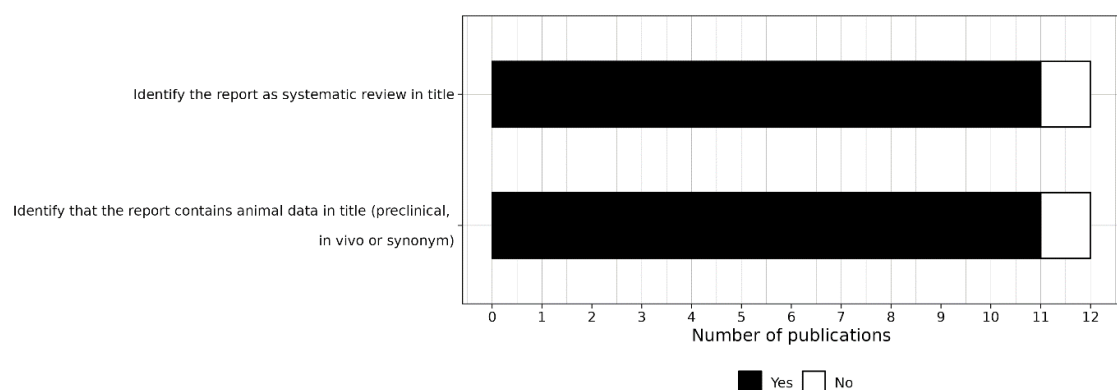


Figure 2: Level of reporting of PRISMA-Pre items within the title section.

2.4.3.2 Items related to reporting in the introduction

In the introduction section, all the reviews described the human condition being modelled, but only six reviews provided an explicit statement of the questions being addressed (Figure 3). None of the included reviews focused on the effects of interventions, so the reporting item related to describing the biological rationale for testing the intervention was not applicable for any of the reviews. Only two reviews indicated whether a protocol was registered (both were registered *a priori* on the PROSPERO platform). However, only one indicated whether any deviations were made to the protocol (Figure 4).

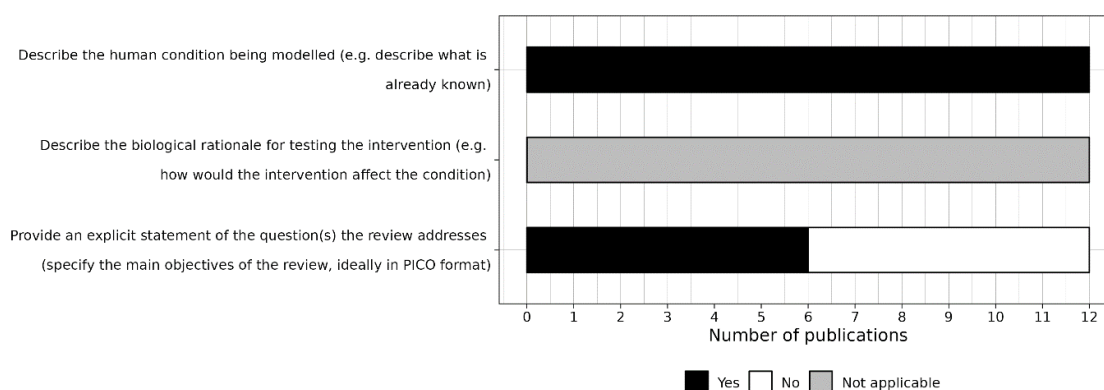


Figure 3: Level of reporting of PRISMA-Pre items within the introduction section. PICO = Population, Intervention, Comparison, Outcome.

2.4.3.3 Items related to reporting in the methods

The eligibility criteria related to animal species, models, and outcomes of interest were reported in seven, nine, and eight of the reviews, respectively (Figure 4). Items relating to reporting interventions of interest and the timing of intervention delivery were not applicable to any of the reviews. Nine of the twelve reviews gave the full search strategy of all databases searched, and only four described inclusion limits added to the search (Figure 4). Three reviews described the study selection process, two reported the platform used for screening (Rayyan and Microsoft Excel), four reported the number of independent screeners, and two reported the number of reviewers extracting data. Three reviews reported methods for assessing the risk of bias of the studies they included (one used both CAMARADES and SYRCLE tools, one used only the SYRCLE tool, and one used an unknown tool), none reported methods for construct validity assessment, and only one reported the methods for assessing publication bias (Figure 4). Several of the reporting criteria for the methods section are related to meta-analyses, so were only applicable for the two reviews that reported a meta-analysis. One reported eligibility criterion related to any controls or comparators, described the methods of data extraction, reported the platforms or tools used to extract numerical data (Rayyan), and described methods for synthesising the quantitative effect measures of included studies, any required data transformations, heterogeneity assessment, and sub-group or sensitivity analysis (Figure 4). The other with limited reporting

was a conference abstract and may have been affected by abstract word limits. Neither meta-analysis reported methods for handling shared control groups or effect sizes over multiple time points.

2.4.3.4 Items related to reporting in the results

All twelve reviews reported the number of individual reports included in the review (Figure 5). Nine provided a summary table of individual studies with data and references, and nine also included a PRISMA flow diagram. These items were not applicable for the two conference abstracts. For reporting of study characteristics, eight reviews reported the animal species, ten reported model details, five reported a measure of the sample size, and four reported individual study designs or intentions. Reporting of intervention details was not applicable for any review. Of the three reviews which reported methods to conduct a risk of bias assessment, and one review which reported methods to analyse publication bias, all reported the results of these assessments (Figure 5). The meta-analysis published in a peer-reviewed journal article reported the number of studies included in quantitative analysis, the outcome effects of included studies and associated confidence intervals, a measure of heterogeneity between included studies, and results from sub-group or sensitivity analysis (Figure 5). These items were not reported in the conference abstract.

2.4.3.5 Items related to reporting in the discussion

Two of the three reviews that conducted a risk of bias assessment discussed the impact of this on the included studies (Figure 6). Six reviews discussed the limitations of the individual studies they included, but only three discussed the limitations of the review itself.

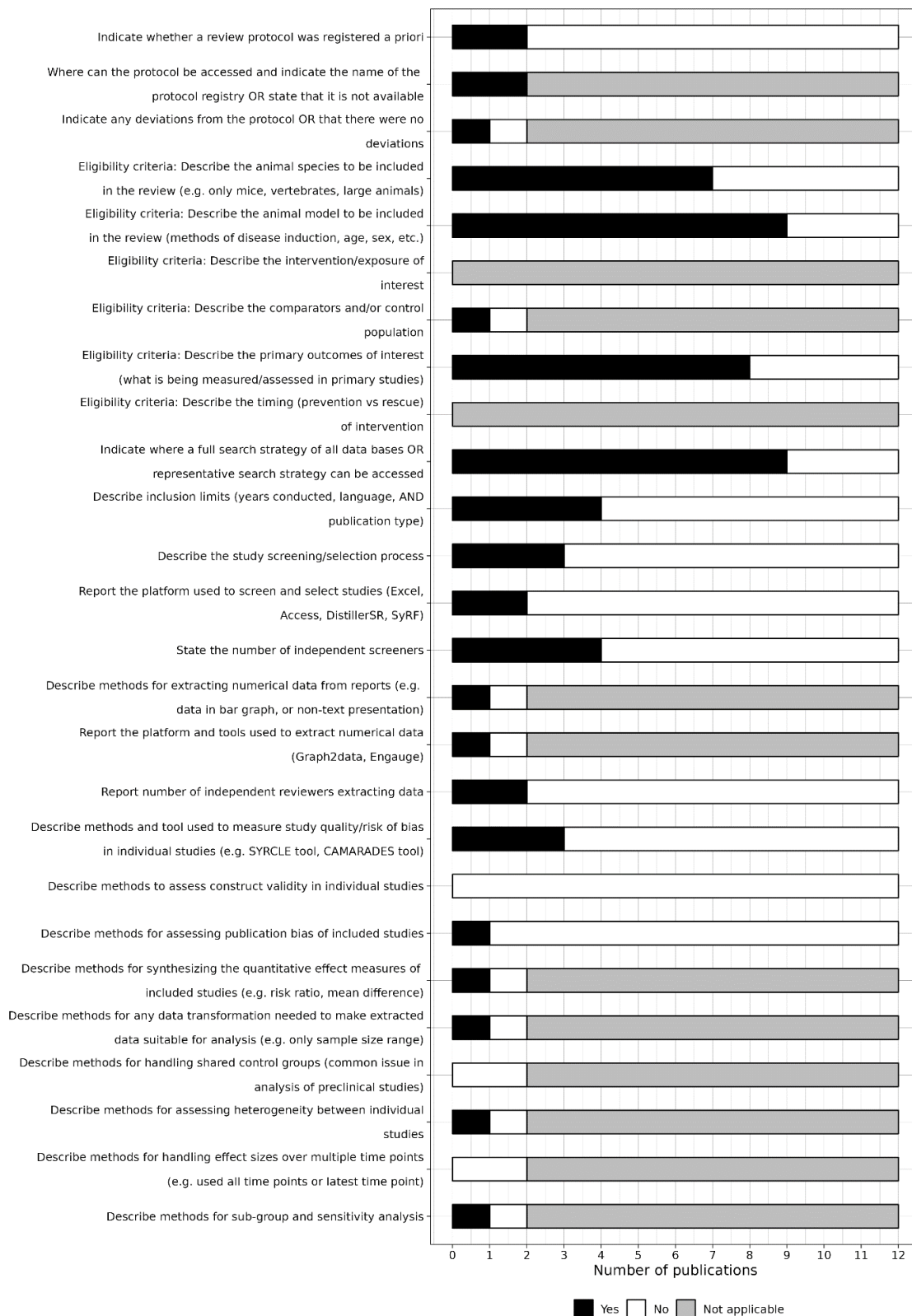


Figure 4: Level of reporting of PRISMA-Pre items within the methods section. CAMARADES = Collaborative Approach to Meta-Analysis and Review of Animal Data from Experimental Studies. DistillerSR = Distiller Systematic Review Software. SYRCLE = SYstematic Review Center for L Laboratory animal Experimentation. SyRF = Systematic Review Facility.

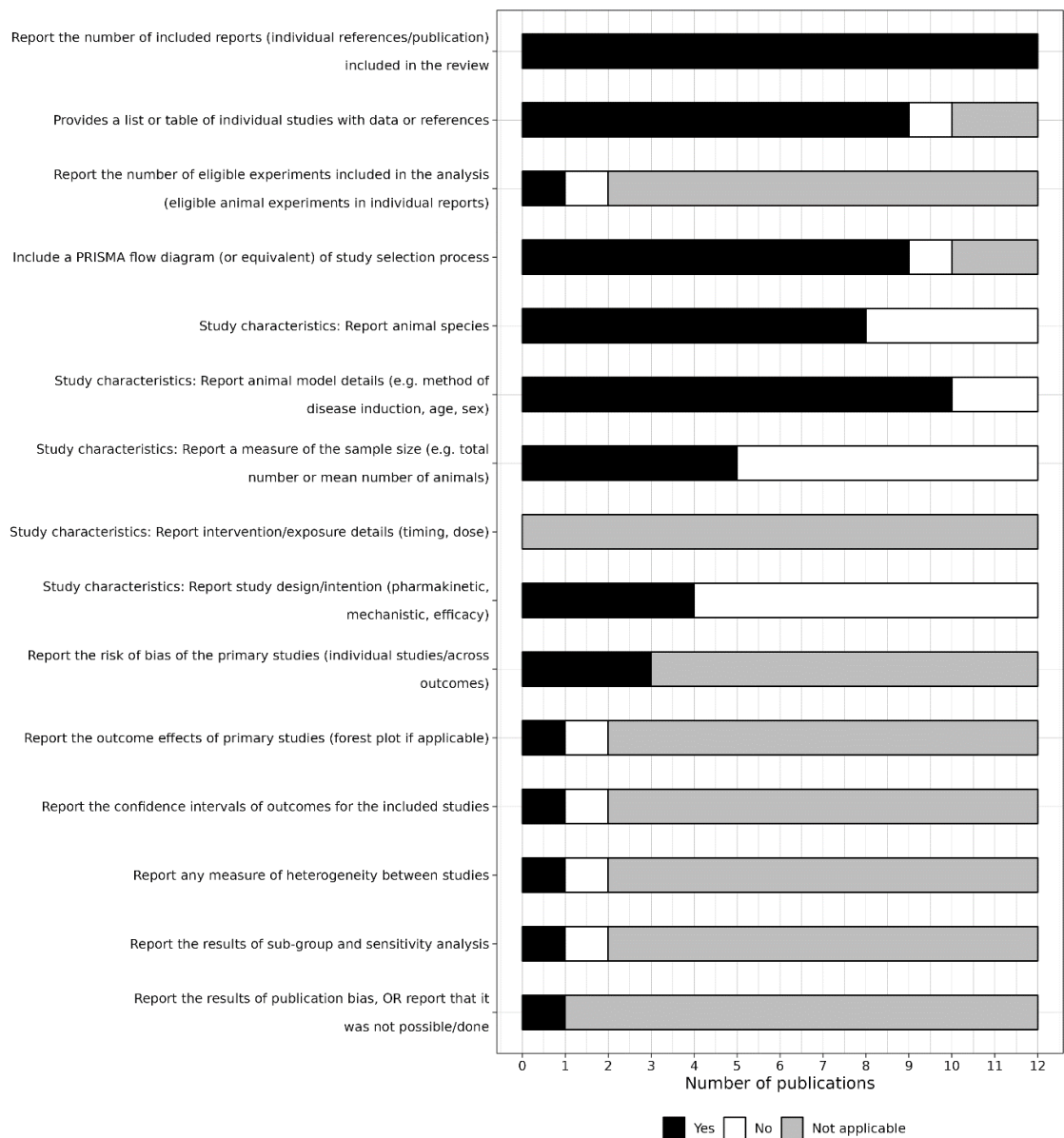


Figure 5: Level of reporting of PRISMA-Pre items within the results section. PRISMA = Preferred Reporting Items for Systematic reviews and Meta-Analyses.

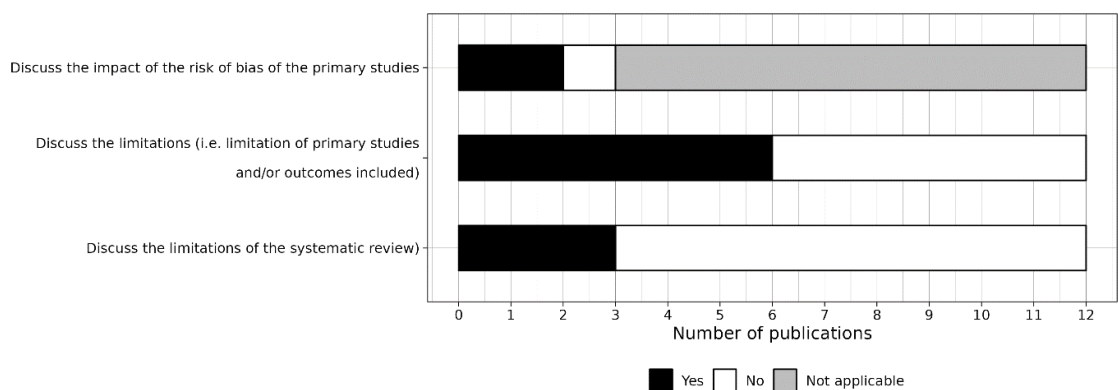


Figure 6: Level of reporting of PRISMA-Pre items within the discussion section.

2.4.3.6 Items related to reporting in other sections

Finally, ten reviews (all the journal articles) reported the source of funding for the review, and four reviews included a data availability statement (Figure 7).

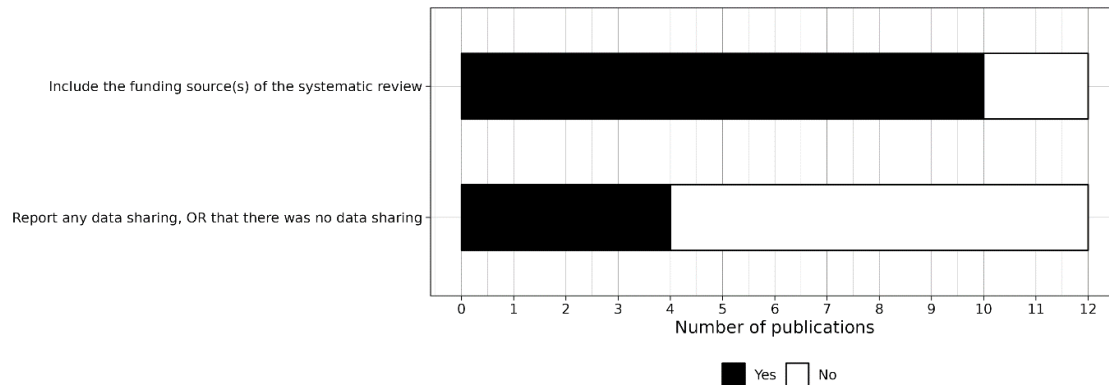


Figure 7: Level of reporting of PRISMA-Pre items within other sections.

2.4.4 Identification of ongoing reviews via PROSPERO

Our search of PROSPERO identified 82 ongoing or completed reviews. After screening, we identified 19 review registrations fitting our inclusion criteria (Table 2). It was often difficult to tell if reviews would include genetic models, so we opted to be overinclusive, only excluding from our analysis if it was clear that genetic models would be excluded. Many of the protocols did not specify specific models of interest, instead often stating that they will include “all” animal models of NDCs or “all” animal models of autism. Only two of these reviews were published at the time of our analysis; these were the two reviews from our main search which reported a protocol. The remaining 17 reviews were marked as ongoing. We did, however, identify that the unpublished conference abstract we found in our main review was registered in PROSPERO, although this registration was not reported in the conference abstract (Figure 8).

All the registrations were made between 2018 and 2023. Six registrations, all of which are ongoing, were made on the same date by the same first author. The mean number of days authors anticipated to complete reviews was 291 (minimum 52 days and maximum 715 days). Some of the ongoing reviews anticipated a very quick completion time (less than 100 days).

The two published reviews were registered in 2020 and 2021, and authors anticipated 508 and 395 days, respectively, to complete their reviews. In practice, they took 528 days and 347 days, respectively, from start date to journal submission (Table 2).

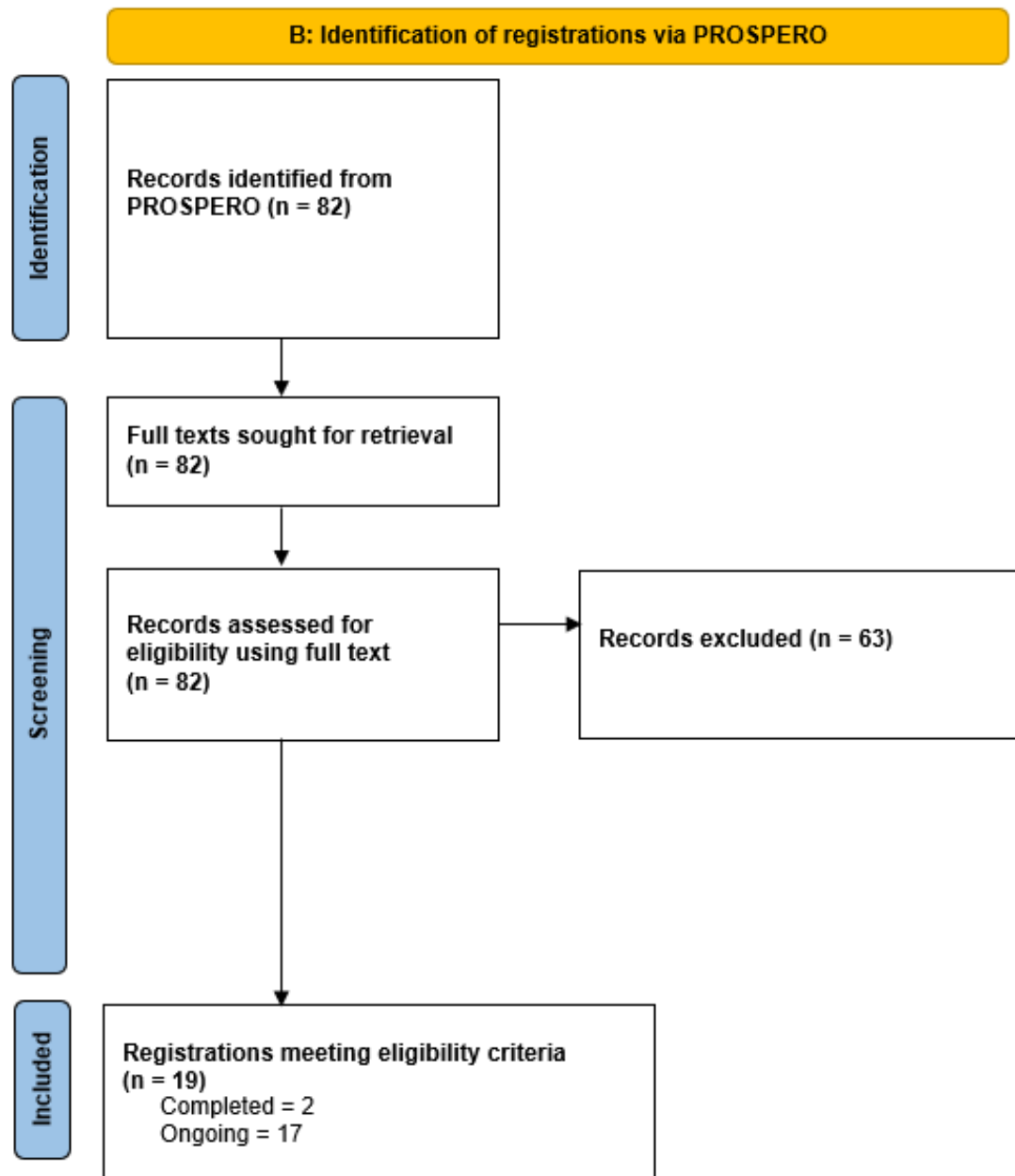


Figure 8: Adapted PRISMA flow diagram for PROSPERO. Diagram templates are shared under a Creative Commons By Attribution (CC-BY) license (<https://creativecommons.org/licenses/by/4.0/>) and available at <https://www.prisma-statement.org/prisma-2020-flow-diagram>.

PROSPERO Identifier	Anticipated Start Date	Anticipated End Date	Journal Submission Date (If Published)	Days Elapsed	Days Anticipated
CRD42018103628	2018-07-16	2018-10-16		1907	92
CRD42019122991	2018-11-01	2019-08-31		1799	303
CRD42019122981	2018-11-01	2019-08-31		1799	303
CRD42019122990	2018-11-01	2019-08-31		1799	303
CRD42019118855	2018-11-01	2019-08-31		1799	303
CRD42019122983	2018-11-01	2019-08-31		1799	303
CRD42019122988	2018-11-01	2019-08-31		1799	303
CRD42020184971	2020-05-05	2020-09-30		1248	148
CRD42020181581	2020-04-08	2020-07-01		1275	84
CRD42020191070*	2020-06-08	2021-10-29	2021-11-18	528	508
CRD42020196337	2020-05-15	2021-04-30		1238	350
CRD42021226299**	2020-11-15	2022-10-31		1054	715
CRD42022306558*	2021-03-01	2022-03-31	2022-02-11	347	395
CRD42022335883	2021-02-01	2022-12-01		976	668
CRD42023392578***	2023-01-02	2023-12-01		276	333
CRD42023382270	2022-12-21	2023-04-21		288	121
CRD42023434189	2023-06-10	2023-08-01		117	52
CRD42023425261***	2023-05-08	2023-11-12		150	188
CRD42023448404	2023-07-10	2023-09-10		87	62

Table 2: Summary of the 19 review registrations identified via PROSPERO. * indicates the registered review appears in the main review as a journal article. ** indicates the registered review appears in the main review as a conference abstract. *** indicates the anticipated end date has not past at the time of analysis. Time of analysis (which is also used to calculate days elapsed) was 5 October 2023.

2.5 Discussion

Our searches identified ten published systematic reviews, two conference abstracts, and seventeen ongoing reviews related to genetic animal models of NDCs. Most of the reviews are recent, with the earliest published in 2016. The first PROSPERO registration was in 2018; before 2018, PROSPERO did not allow registration of animal systematic reviews (Pieper & Rombey, 2022). We made the decision to include both (X. Zhang et al., 2021) and (X.-Y. Zhang & Spruyt, 2022) in our analysis, as despite the publications referring to the same systematic review research, they are separate publications and their reporting quality can be assessed separately.

From the published literature, we can see that reporting against the PRISMA-Pre checklist was mixed. Reporting in the title to identify publications as (1) systematic reviews and (2) related to animal research was overall good. One of the reasons we chose to screen based on the full text rather than one round of title and abstract screening was that we were unsure if this would cause us to erroneously exclude publications with unclear titles or abstracts.

Within the introduction section, authors consistently reported descriptions of the conditions being modelled, but only half explicitly reported the review aims. The lack of clear reporting in this area made it difficult to code the review aims in our summary table and may impact the use of individual reviews if their purpose is not clear.

Methods and results reporting were mixed. Few reviews reported having a preregistered protocol. Those that did preregister did so through PROSPERO, a dedicated repository for protocols of systematic reviews related to human health. Protocol registration is often considered a fundamental step in the systematic review process, to ensure that the review methodology is transparent and reduces the risk of bias being introduced in the systematic review (Soliman et al., 2020).

Eligibility criteria and search strategies were reported well overall, with room for improvement, while reporting of methods to screen studies was limited. The reporting of the number of studies included, and inclusion of summary tables and PRISMA flow diagrams were better. The small number of reviews that included a meta-analysis limits our ability to interpret reporting of items related to quantitative analysis with any confidence.

Few of the reviews conducted a risk of bias assessment, meaning the potential impact of risk of bias within primary studies was often not discussed. Many tools have been developed to assess the quality of evidence included in systematic reviews of animal data and associated risks of bias, including the CAMARADES checklist (M. R. Macleod et al., 2004) and the SYRCLE risk of bias tool (Hooijmans et al., 2014). Only one of the reviews that did not conduct a risk of bias assessment stated their reasoning, that the reporting of the studies they included was too poor to conduct a risk of bias assessment. However, the purpose of a risk of bias assessment is to determine precisely the reporting quality and present this evidence (Soliman et al., 2020). Systematic reviews of animal studies often observe very limited reporting quality and present that many primary studies they include are at high risk of bias (Soliman et al., 2020).

Only half of the reviews discussed the limitations of the studies they had included, and a quarter discussed the limitations of the review itself. Systematic reviews, just like any other form of research, are susceptible to research bias. Systematic reviews often play a key role in research and healthcare decision making, so it is important that the findings of reviews are as free from bias as possible. As discussed above, protocol registration is one method for mitigating at least some of this bias and improving the transparency of the work; however, this does not negate the need to discuss limitations.

Finally, funding sources of the reviews were reported in all peer-reviewed journal articles, likely because of publishing policies. Data availability statements, which should report whether data are available or not available, were included in less than half of the reviews, despite this also being a requirement for most journals.

Systematic review is only one method used to bring understanding to translational challenges within biomedical science. In NDC research, there are many potential explanations for this translational barrier, including limitations in how research is designed, conducted, and reported (Wilson, Ramage, et al., 2023), and research that is rooted in the medical model and fails to take into consideration advances in our understanding of NDCs achieved through neuroaffirmative research approaches (Heraty et al., 2023).

Systematic review of animal studies of NDCs is still an emerging activity. There are fundamental differences between clinical and animal research, such as the experimental methods used, number of participants/animals per publication, and number of relevant publications retrieved in systematic literature searches, meaning systematic review methods have to be adapted from the clinical space to fit animal research. Systematic reviews of animal research related to other biomedical conditions are ahead in this space, and many resources have been developed to help researchers conduct their reviews with rigour (De Vries et al., 2014, 2015; Hooijmans et al., 2014; M. R. Macleod et al., 2004; Sena et al., 2014). Unfortunately, our findings demonstrate that study authors of the reviews we have included have not engaged with much of this guidance and the available resources.

Systematic reviews are an empirical form of research, and if we are to rely on their findings, they must be conducted rigorously (Sena et al., 2014). When assessing the quality of research, including research like systematic reviews, we cannot directly assess the methodological quality; therefore, we assess the reporting quality as a proxy. Where the reporting of systematic review is limited, our understanding of the review's methodological quality is limited, and in turn we have less certainty around the reliability of the review's findings. We identified 17 systematic reviews in progress, demonstrating that this is an area of research growth. Moving forward, we encourage authors to make use of existing resources to maximise the value and impact of their work. Table 3 summarises some key resources for conducting systematic reviews of animal studies.

Reference	Resource	Type
(Sena et al., 2014)	Systematic reviews and meta-analysis of preclinical studies: why perform them and how to appraise them critically	General guidance article
(De Vries et al., 2015)	A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies	Systematic review protocol template
(Pieper & Rombey, 2022)	PROSPERO	Systematic review preregistration platform
(Hunniford et al., 2021)	Epidemiology and reporting characteristics of preclinical systematic reviews	Systematic review reporting guidelines (PRISMA-Pre checklist)
(Hooijmans et al., 2014)	SYRCLE's risk of bias tool for animal studies	Assessing risk of bias in animal research (SYRCLE checklist)
(M. R. Macleod et al., 2004)	Pooling of animal experimental data reveals influence of study design and publication bias	Assessing reporting quality in animal studies (CAMARADES checklist)
(Bahor et al., 2021)	Systematic Review Facility (SyRF)	Systematic review screening and annotation

Table 3: Resources for conducting systematic reviews of animal studies

Most of the systematic reviews we identified were published in 2022, and likely conducted during 2020-2021. We cannot know with any certainty if laboratory closures during the COVID-19 pandemic influenced the uptick in reviews conducted and whether this activity will continue to rise. However, it is important that the reporting quality, as well as the design and conduct, of systematic reviews in this area follow the advances in systematic review methodology we see in other areas, including focal cerebral ischaemia, so that they can have similar effects on primary research direction, conduct, and policy (McCann et al., 2016).

2.5.1 Comparison with results from Hunniford et al. (2021)

To better understand how my findings compare with animal systematic reviews more broadly, I compared my findings (NDC SRs) with those of Hunniford et al. (2021) (animal SRs). This was an unplanned analysis, limited by my small sample size (N=12). As only two publications included a meta-analysis, I do not compare findings on meta-analysis-specific criteria.

Identification of publications as systematic reviews was very well reported in both animal and NDC SRs (>90%), however a higher percentage of NDC SRs identified the review as containing animal data in the title than animal SRs more broadly (92% versus 54%). In the introduction section, description of the human condition being modelled was similar (100% in NDC SRs versus 82%) while an explicit PICO statement was reported more often in animal SRs than NDC SRs (73% versus 50%).

In the methods section, indication that a protocol was registered a priori was poorly reported in both NDC and animal SRs. There was similar mixed reporting of eligibility criteria in NDC and animal SRs, while reporting of full search strategies was similarly high in both NDC and animal SRs (75 and 75% respectively).

Processes used for screening, the number of screeners, and the number of reviewers for data extraction were all well reported in animal SRs but poorly reported in NDC SRs, while the platform used for screening was poorly reported in both NDC and animal SRs (10% and 17% respectively). 49% of animal SRs reported methods for a risk of bias assessment compared to only 25% of NDC SRs, while

methods for assessing construct validity and publication bias were poor in both samples.

In the results section, the number of studies included, a list of studies included, and a PRISMA flow diagram were all well reported in both NDC and animal SRs, while reporting of study characteristics were mixed. In the discussion section, reporting of the impact of risk of bias (where this was conducted as part of the methods) was higher in NDC SRs than animal SRs (67% versus 31%), although only 3 NDC SRs conducted a risk of bias assessment. Discussion of limitations of both the primary studies included in reviews and limitations of the systematic reviews themselves were reported more often in animal SRs than NDC SRs (65% and 56% versus 50% and 25% respectively). Finally, funding was reported similarly in both NDC and animal SRs (83% versus 78%) while data sharing was more commonly reported in NDC SRs than animal SRs (33% versus 19%) although both could be much improved.

Overall, I found room for improvement across the board. While NDC SRs showed better reporting in some aspects (e.g., data sharing and identification that reviews investigate animal data), they showed poorer reporting in others (e.g., use of PICO statements and reporting of methods used to screen and annotate, including the number of screeners and reviewers). The median publication year of NDC SRs was 2022 while three animal SRs were published from 2015 to 2018 inclusive, meaning improvements in practices like data sharing may be as a result of a more general uptake of open research practices within biomedical sciences (Deeb et al., 2024). Similarly, poorer reporting of screening and reviewing methodologies seems to highlight a lack of awareness of specific systematic review best practices, as mentioned earlier in the discussion. Importantly, however, most other criteria had similar reporting between NDC and animal SRs, and there is indication that reporting quality improvements are required across all animal systematic reviews.

2.5.2 Strengths and limitations

A major strength of our study is that, by preregistering our study using the Registered Report format, our rationale and proposed methods have been peer-reviewed, meaning we have been able to make improvements to our proposed study design prior to beginning our research.

To maximise the sensitivity (recall) of our search, we used broad search terms related to NDCs, intellectual disability, epilepsy, and autism, in parallel with terms related to individual associated genes, and screened studies based on full text to avoid erroneously excluding systematic reviews which do not report their inclusion criteria in their abstract.

A limitation, however, is that by only searching primarily English-language databases we may have missed non-English reviews (possibly evidenced by the lack of such reviews in our results).

Although we identified protocols for ongoing reviews via PROSPERO, we did not analyse the reporting of these protocols. Additionally, from our results, we know that the majority of published reviews were not preregistered, suggesting that we may be unaware of many more ongoing and unpublished reviews. We used a draft extension of the PRISMA guidelines specifically designed for systematic reviews of animal studies (PRISMA-Pre (Hunniford et al., 2021) also included in Appendix 2C) to assess the reporting quality of included systematic reviews. This extension remains in draft, is not yet an official extension to PRISMA and may be subject to change. The checklist was designed to improve the reporting quality of journal articles. In hindsight, it may not have been appropriate for us to apply the checklist to conference abstracts. Conference abstracts have heavily limited word counts and their purpose is to act as short communication pieces where further details are given in presentations (oral or poster).

Additionally, many aspects of the PRISMA-Pre checklist were not applicable for this dataset. All the systematic reviews identified were interested in mechanistic studies, so intervention-related questions within the checklist were not relevant.

Additionally, only two of the reviews included a meta-analysis. An improvement to the PRISMA-Pre checklist may involve adding additional subheadings to separate the meta-analysis-related questions from “core” systematic review reporting items.

2.6 Conclusion

To conclude, recent years have seen the publication of a diverse range of systematic reviews investigating genetic animal models of NDCs. Systematic evaluation of published research can help summarise research findings and strengthen evidence most effectively when effort is taken to minimise bias within the review itself. Within our sample, the reporting of published systematic reviews is mixed. We rely on clear and transparent reporting as a proxy to determine sound methodological quality. Therefore, researchers conducting systematic reviews in this space should engage with systematic review guidance, where available, to strengthen and ensure the robustness and reliability of their reviews. Increased awareness of reporting guidelines may help authors plan and report their systematic reviews with more details to increase their transparency and reproducibility.

Chapter 3: Systematic approach to identifying research publications

3.1 Context

In this chapter, I set out to test and validate an approach to systematically identify relevant research publications. This approach would then form the foundation for a systematic map. Fiona Brown (academic librarian) and Peter Kind (subject matter expert) provided guidance that helped me build my systematic search strategy. Fiona Brown also provided guidance on how to validate the search strategy. The machine learning classifier I use in this chapter was developed by James Thomas (EPPI-Centre, University College London). Malcolm Macleod, Sean Smith, Francesca Tinsdeall, and Ian Simpson provided guidance on how best to validate the machine learning classifier. Kaitlyn Hair helped me conceptualise how to validate NER approaches.

3.2 Introduction

Neurodevelopmental conditions (NDCs) is an umbrella term representing a highly heterogeneous group of lifelong disabilities arising from differences in how the brain and nervous system develops (Thapar et al., 2017). Research into the neurobiology of NDCs is similarly diverse, and a wide breadth of both animal and organoid models are used (Damianidou et al., 2022). Despite convergence of the genetic pathways implicated in many monogenic NDCs, much of the research into those conditions exists in silos (Peall et al., 2024). By summarising evidence from

animal research modelling monogenic NDCs, we may be able to identify and assess similarities and differences between these conditions and improve our overall knowledge on the underlying neurobiology.

AutDB (Autism Database) was a genetic reference tool developed in 2009 for research related to autism (Basu et al., 2009). In 2011, the resource added to its collection a curated database of “autism-related” animal models (Kumar et al., 2011). The AutDB animal model content was manually annotated by autism researchers from one research group. Content related to mouse models (but not rat, fruit fly, or zebrafish models) from the AutDB resource is mirrored on the SFARI Gene Animal Model database (SFARI Gene, 2024a). However, neither AutDB and SFARI is systematically comprehensive; only PubMed was searched, not all relevant publications related to each animal model appear to be listed, and the methods used to search for publications and extract data from publications are not transparently and reproducibly reported. Additionally, AutDB appears to be no longer maintained, and my efforts to contact the authors have been unsuccessful.

One major barrier to broad, systematic curation of research literature is the rate of publication. Within the biomedical sciences, publication rate has rapidly risen over recent years; a search of MEDLINE for the Medical Subject Heading (MeSH) term “neurodevelopmental disorder” yielded over 211,000 results on 23rd November 2023, with 14% of results published between 2020 and 2023. This publication rate is impossible to contend with manually. Nonetheless, it is important that researchers and research-users can stay on top of the published literature to make the best use of research findings and to make evidence-based decisions.

Recently, evidence synthesis methodologists have trialled the implementation of automated approaches to improve the feasibility of conducting systematic reviews in research-intensive areas (Bannach-Brown et al., 2021). Several steps in the systematic review process can be assisted using machine learning algorithms, natural language processing (NLP) and other textual analysis approaches, and application programming interfaces (APIs) (Bannach-Brown et al., 2021). In this

chapter, I discuss, test, and validate how such approaches might help identify relevant publications for a systematic map.

Traditional systematic review search strategies are typically built by incorporating P, I, and O elements from PICO (population, intervention, control, and outcome)-structured systematic review questions. Systematic maps, however, aim to capture publications relevant to a broader research area. For instance, a systematic map of animal research modelling Alzheimer's disease uses only the term "Alzheimer's" to identify relevant research; animal studies are then identified using a trained machine learning classifier (Hair et al., 2024). Due to the great variability in terminology and models used in monogenic NDC animal research, it is likely that search strategies will have to include several broad terms to systematically capture relevant research.

Systematic searches will retrieve irrelevant publications that must be removed through a process called screening. In systematic reviews, typically two independent reviewers will screen each search result for relevance and any disagreements will be reconciled either by discussion or by a third reviewer. Support vector machines (SVMs) are supervised machine learning algorithms commonly used in text classification problems, when texts can fall into one of two categories (binary classification). These binary classifiers can be trained on pre-defined inclusion and exclusion criteria and used as part of the screening process (Bannach-Brown et al., 2019, 2021).

Named entity-recognition (NER) is a type of natural language processing (NLP) used to extract structured information from unstructured text (Leser & Hakenberg, 2005). General uses of NER include extraction of telephone numbers, email addresses, and names from text. In medical and biomedical fields, NER can be used to extract information such as gene, protein or drug names and, where synonyms exist, standardise these to a preferred name (Leser & Hakenberg, 2005). In a systematic map context, NER can be used to identify entities within text, such as the population being studied and the outcomes tested. There are many approaches to NER, including dictionary-based approaches and more flexible

machine learning based approaches (Leser & Hakenberg, 2005). Dictionary-based approaches, where texts are searched for a list of possible words or phrases related to entities, are particularly useful for NER of medical and biomedical texts, however they are limited in their scalability as dictionaries may need to be manually updated over time. Regular expressions (commonly shortened to regex) are a form of pattern matching that can be used in dictionary-based NER to find and extract relevant words or phrases and account for case-sensitivity, spelling, and synonyms.

To assess how well approaches such as systematic searches, binary machine learning classifiers, and dictionary-based NER approaches work we require labelled data for testing and/or training. Systematic searches require a list of known relevant publications that can be checked against search results. Binary classifiers require a sample of text publications (typically the title and abstract) each labelled with either a binary “include” or “exclude” decision. Dictionary-based NER approaches require text data from a sample of publications annotated with all correct named entities.

Labelled data are used to test and validate the performance of systematic searches and NER approaches. Binary classifiers, however, are additionally trained on a proportion of labelled publications (training set), while a further portion is used to calibrate the classifier to the optimal performance (test set), and a third proportion is used to validate this performance and ensure the classifier will likely be stable when shown new data (validation set).

The number of true positives (TP) is calculated as the number of publications from the test or validation set that the classifier correctly included or where NER correctly extracted a named entity, and the number of true negatives (TN) is the number of publications correctly excluded. Conversely, the number of false positives (FP) is the number of publications incorrectly classified as included and the number of false negatives (FN) is the number of publications incorrectly classified as excluded. Ideally, the number of TP and TN should be as high as possible, and the number of FP and FN should be as low as possible. Performance

of both binary classifier and NER approaches is calculated through various metrics utilising TP, TN, FP, and / or FN. Commonly used metrics include recall (sensitivity), precision, specificity, and F1 score (the harmonic mean of recall and precision):

$$recall = \frac{TP}{TP + FN}$$

Eq. 1 Recall

$$precision = \frac{TP}{TP + FP}$$

Eq. 2 Precision

$$specificity = \frac{TN}{TN + FP}$$

Eq. 3 Specificity

$$F1\ score = \frac{2 \times precision \times recall}{precision + recall}$$

Eq. 4 F1 score

Additionally, where higher recall is sought after and lower precision can be handled, such as is the case in evidence synthesis, the F2 score may be more appropriate as it puts more emphasis on recall over precision:

$$F2\ score = \frac{1 + 2^2}{\frac{2^2}{recall} + \frac{1}{precision}}$$

Eq. 5 F2 score

The Matthew's Correlation Coefficient (MCC), also known as the phi coefficient, takes all four of TP, TN, FP, and FN into consideration and can be particularly useful if there is class imbalance in datasets, for instance more expected negatives or "excludes" than positives or "includes":

$$MCC = \frac{TN \times TP - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Eq. 6 Matthew's Correlation Coefficient (MCC)

3.2.1 Aims

Here, my overall aim is to develop a method to systematically identify research publications using animals to model monogenic NDCs. Firstly, I aim to develop and

validate a systematic search strategy that is feasible to run and broad enough to capture the diversity of relevant evidence. I envision that this broad searching approach will also retrieve many irrelevant publications. Therefore, secondly, I aim to train and validate a machine learning algorithm, developed for the purposes of aiding screening of systematic search results, to correctly identify research publications. Finally, I aim to develop, test and validate dictionary-based NER approaches to identify features relevant to genetic animal models: gene name, animal species, animal sex, and the experimental tests conducted using the animal model.

3.3 Methods

3.3.1 Protocol registration

A protocol for binary classifier development is available online on the Open Science Framework as part of a wider protocol for the systematic map (DOI: <https://doi.org/10.17605/OSF.IO/GFTZP>). Deviations between the protocol and methods described in this chapter are:

- The inclusion criteria were expanded to include Level 1 genes on the SFARI Gene list.
- The method of validating performance (k-fold cross validation) was added in this chapter.

A protocol for developing and validating NER approaches is available online on the Open Science Framework (<https://doi.org/10.17605/OSF.IO/5H4FB>). Deviations between the protocol and methods described in this chapter are:

- Regular expressions for experimental tests were also tested and validated.

3.3.2 Inclusion and exclusion criteria

I will include research which meets the following criteria:

- Study design – I included original research publications (mechanistic or interventional) using animals to model NDCs. I excluded studies which only presented results from clinical, *in vitro*, or *in silico* research.
- Animal models – I included publications using genetically-modified animals of any species, where the modified gene(s) are any of the 150 genes total identified via large-scale exome sequencing in Satterstrom et al. (2020) or a Level 1 (high confidence) gene on the SFARI Gene list (SFARI Gene, 2024a). All 150 genes are listed in Appendix 1A. I excluded all research which using other types of models, e.g., chemically induced models such as sodium valproate.
- Article type – I included peer-reviewed journal articles, preprints, and conference abstracts describing original research. As screening was based on title and abstract alone. I excluded review articles, book chapters, comments, letters to the editor, news articles, and opinion pieces.

I placed no restriction on language, although I acknowledge I searched primarily English-language data sources. I restricted my search to articles published in or after 1994.

3.3.3 AutDB animal model resource

I compiled a list of 484 known relevant publications, matching my inclusion criteria, from the AutDB animal model resource (Kumar et al., 2011). I then used this list to validate the search strategy and aid in training a machine learning classifier. All the publications listed in the AutDB animal model resource are indexed in PubMed.

3.3.4 Systematic searching

I developed search strategies to use on three online bibliographic sources easily accessible via free APIs: PubMed, Scopus, and Web of Science Core Collection. I first developed a search strategy using PubMed syntax, which included terms relevant to NDCs (including intellectual disability, epilepsy, and autism), the 150 relevant gene names, and relevant animal model species names. I developed

search terms with the help of a subject matter expert (Professor Peter Kind) and an academic librarian (Fiona Brown). I then calculated the recall of the PubMed search against the list of 484 known relevant publications from AutDB (Kumar et al., 2011) to assess how well it performed. Finally, I translated the search terms into the correct syntax for Scopus and Web of Science Core Collection searches. Full search terms are given in Appendix 3A.

3.3.5 Screening publications for relevance

3.3.5.1 Generating a labelled dataset for binary classification

I downloaded all search results from PubMed, Scopus, and Web of Science Core Collection and removed duplicates using the automated and manual duplicate removal features in ASySD (Hair, Bahor, et al., 2023). From this combined and duplicated dataset, I randomly sampled 2,000 publications using the *sample()* function in base R (RRID:SCR_001905).

I uploaded the title and abstract text from the 2,000 publications to the SyRF (RRID:SCR_018907) (Bahor et al., 2021) for manual screening. Each publication's title and abstract text was screened by two independent reviewers who were unaware of the other reviewer's decision. Screening decisions were made using the above inclusion criteria. Reviewers also noted publications that were otherwise relevant but investigated monogenic NDC genes not included on the gene list. Disagreements between two reviewers were reconciled by a third independent reviewer in the SyRF platform. The reconciling reviewer was unaware of which initial reviewer made which screening decision and reviewers were unaware as to whether their decision was an initial screening decision or reconciliation of a disagreement. SyRF presented publications to reviewers and reconcilers in a random order for screening.

To account for the low yield of included publications in the random sample, I combined this data with the 484 known relevant AutDB publications.

In total, the labelled dataset consists of the title and abstract text data from 2,484 publications and an assigned screening score: either “included” (denoted as 1) or “excluded” (denoted as 0) in relation to the inclusion criteria.

3.3.5.2 Machine learning development and error correction

In an initial test, using the R *sample()* function, I randomly split all the labelled data into two sets: 80% of the data was assigned to the training set and 20% assigned to the test set. I then trained the classifier using labelled data from the training set and tested the performance by withholding the test set labels from the classifier and comparing machine scores with the test set labelled data. The classifier gives a decimal score from 0 to 1, with a score closer to 0 meaning the classifier has more certainty that the publication belongs in the excluded set and a score closer to 1 meaning the classifier has more certainty that the publication belongs in the included set.

I conducted a round of error correction, by taking the 100 publications from the test set where the labelled score (binary) and the machine score (decimal) differed greatest and uploaded these data to the SyRF platform for rescreening by a new independent reviewer. The process of error correction acts on the basis that the discrepancy between the labelled data and the machine score may indicate that the publication was labelled incorrectly during the initial screening phase (Bannach-Brown et al., 2019). In hindsight, the process of error correction would have been better placed after stratified 5-fold cross validation (described below) so that potential screening errors across all the labelled data could have been assessed.

3.3.5.3 Stratified 5-fold cross validation and classifier performance

When training and assessing the performance of a machine learning classifier, there is a danger of overfitting and overestimating the performance. To combat this as much as possible, I used a process called stratified k-fold cross validation, which is better able to estimate the performance and stability of a machine learning classifier when faced with new data. There are various methods and

terminologies used in k-fold cross validation. The process I used is explained in Table 4 and in the paragraph below.

	Full Labelled Dataset (100%)					
	Fold / Set 1 (16%)	Fold / Set 2 (16%)	Fold / Set 3 (16%)	Fold / Set 4 (16%)	Fold / Set 5 (16%)	Validation Set (20%)
Iteration 1	Train	Train	Train	Train	Test	Validate
Iteration 2	Train	Train	Train	Test	Train	Validate
Iteration 3	Train	Train	Test	Train	Train	Validate
Iteration 4	Train	Test	Train	Train	Train	Validate
Iteration 5	Test	Train	Train	Train	Train	Validate

Table 4: Full labelled dataset split for stratified 5-fold cross validation. Each fold and validation set contains the same proportion of included publications.

The 'k' in k-fold refers to the number of folds (or sets) used to evaluate the algorithm performance; in this case I used 5. To achieve 5-fold cross validation, I randomly split the error corrected labelled dataset of 2,484 publications into 6 sets. 80% of the data was split into 5 sets of equal size (16% of the total data each), which were used to generate the 5 folds, and the remaining 20% of the data was kept aside as the validation set. I used stratified random sampling to ensure that each of the 6 overall sets had the same proportion of included and excluded publications. I split the 5 folds into training (4 folds) and test (1 fold) sets, across 5 iterations. In each iteration, labelled data each of the training sets were used to train the algorithm and the test and validation sets were passed through the algorithm without labels to receive machine scores.

Machine scores given to the publications in the test sets were then compared to the publications' manual labels to calculate the number of TP, FP, TN, and FN, which were then used in performance metric calculations (given in the chapter introduction) for each possible machine score threshold from 0.01 to 1.00. The optimal machine score threshold was determined as the threshold that achieved the highest specificity at a recall of at least 0.95.

The machine scores given to publications in the validation set were finally compared to the publications' manual labels, at the chosen threshold for each

iteration, and used to calculate the performance at the chosen thresholds on the unseen dataset. The optimal iteration was determined to be the one with the highest F2 score.

3.3.5.4 Applying the trained classifier to all search results

The training data from the optimally performing iteration of the binary classifier was taken forward and used to screen the deduplicated systematic search results (see Methods 3.2.4 Generating a labelled dataset for machine learning assessment for further details) to identify the most likely relevant publications.

3.3.6 Extracting animal model features from texts

3.3.6.1 Generating a labelled dataset for NER approaches

From the subset of data scored as “included” by the machine learning classifier, I retrieved full text documents in XML and PDF formats. To obtain XML files, I first used the `rcrossref` R package (Chamberlain et al., 2014) to retrieve PubMed Central identifiers (PMIDs) for publications using the DOI, then used the `europemc` R Package (Jahn, 2016) to retrieve available XML files using PMIDs.

I used the `xml2` R Package (Wickham et al., 2015) to read the retrieved XML files into R and the `tidypmc` R Package (Stubben, 2019) to retrieve only the main body text (title, abstract, introduction, methods, results, discussion, and conclusion). Using these methods, I was able to exclude reference list text from XML files but not PDF files. I used custom-made code developed by colleagues within the SOLES team to retrieve PDF files. The code first attempts to retrieve full text documents from CrossRef via the `rcrossref` R package (Chamberlain et al., 2014) or UnPaywall via the `roadoi` R package (Jahn, 2017). Where full texts cannot be retrieved by either of these methods, the code will attempt to retrieve full texts using the Elsevier API (Elsevier, 2025) and Wiley Text and Data Mining API (Wiley, 2025). I then used two methods to convert PDF files into text (TXT) files for machine readability: (1) the `readtext` R Package (Benoit & Obeng, 2017), which uses the `pdftools` R Package (Ooms, 2016), and (2) the `pdftotext` software from XpdfReader

(Glyph & Cog, 2025). I aimed to test if one text conversion software performed better than the other.

Using the *sample()* base R function (RRID:SCR_001905), I randomly sampled 400 of the publications where I was able to retrieve both machine readable PDF (e.g. converted to TXT) and XML files and uploaded them to SyRF (RRID:SCR_018907) (Bahor et al., 2021) for manual data extraction. Each publication was first screened to ensure it was relevant to the above inclusion criteria and included publications were annotated with relevant entities related to four topics (1) gene altered in the animal model used; (2) the animal species used; (3) the sex of animals used; and (3) any behavioural or electrophysiology experiments conducted on relevant animal models. Specific genetic models were not extracted, and neither were biochemical experiments. SyRF presented publications in a random order for data extraction and information from each publication was extracted once by myself.

I split the 400 labelled publications into two equal-sized, using the *sample()* R function and ensuring each set had equal proportion of data screened as included and excluded.

3.3.6.2 NER regex development and analysis

I developed regexes prior to extracting data from publications. 150 gene model regexes were created using the gene symbol only (e.g., “FMR1” for animal models of Fragile X syndrome) for each of the 150 relevant genes. Five species regexes were created using synonyms for the five common animal model species included in the search strategy: mice, rats, fruit flies, zebrafish, and roundworm. Three sex regexes were created looking for the words, male/males, female/females, or both sexes. And finally, 45 experimental test regexes were created through subject matter knowledge of commonly used behavioural and electrophysiology experimental tests. All regexes, including both regex string and named entity they are designed to capture are given in full in Appendix 3B.

I used the *tidypmc* R Package (Stubben, 2019) to tokenise the XML text into sentences and run NER on the sentences, and the *quanteda* R Package (Benoit et

al., 2015) to tokenise the PDF text into sentences and run NER on sentences. I ran NER on set 1 from the labelled data on five different text formats (1) title and abstract text; (2) PDF converted the TXT file using XpdfReader pdftotxt software; (3) PDF converted to TXT using pdftools; (4) XML text (all main body text); and (5) XML text (methods section text only). I then compared the entities captured through manual data extraction with those data captured through regex NER on the various text formats. I assessed the number of TP, TN, FP, and FN at the label of individual annotations and at the level of each publication. Using These values, I calculated performance metrics including recall, specificity, precision, and F2 score. The definition of optimal performance was decided through consultation with laboratory researchers.

Where manually extracted entities were not captured by NER (FN, false negatives), I assessed the publication text and determined if the regexes could be improved. Improved regexes are listed in full in Appendix 3B. After making improvements to the regexes, I rechecked the performance on Set 1, including at various thresholds from 1-5 regex mentions in the text required to count as a match. I then checked the performance of the updated regexes on Set 2 to check consistency with Set 1 results and avoid overfitting the NER approach to the data in Set 1.

Finally, using the performance values, I assessed if any of the NER methods could be used to filter some of the false positive publications erroneously included by the classifier, to further improve the overall relevance of the resulting dataset. Specifically, I assessed (1) only extracting model gene and animal species where the respective named entities appear in the same sentence; (2) only extracting model gene and animal species where the respective named entities both appear in the same text format (not necessarily in the same sentence); (3) only extracting animal sex where terms such as “were used” or “we included” are used in the same sentence; and (4) only extracting animal sex and experimental tests from publications that matched with both model gene and animal species entities in the title and abstract text.

3.4 Results

3.4.1 Performance of systematic search

I ran systematic searches on 2 December 2022 and retrieved a total of 203,058 publications. 66,458 publications were retrieved from PubMed, 69,943 from Scopus, and 66,657 from Web of Science Core Collection.

In testing the PubMed search, I found that 451 of the 484 known relevant publications from AutDB (93%) were retrieved and 33 were missed (recall = 0.9318). Investigating the 33 publications not retrieved by the search, I found that these publications were all from at least five years ago and typically did not mention neurodevelopmental conditions or specific genetic models in their title and abstract.

3.4.2 Performance of binary classifier for screening

3.4.2.1 Machine classifier initial training and error correction

After removal of duplicate copies of publications, the combined search yielded 99,086 publications. Of the 2,000 randomly sampled publications from this dataset, 132 publications met the inclusion criteria and 1868 did not. When supplemented with the 484 known relevant publications, this gave us a total set of 2,484 publications, 616 labelled as included and 1868 labelled as excluded. 1,987 publications (80% of the total set) were randomly assigned to the training set (495 labelled included, 1,492 labelled excluded), and 497 (20% of the total set) were randomly assigned to the test set (121 labelled included, 376 labelled excluded).

From the test set machine scores, I determined the 100 publications with the biggest difference between manual and machine label scores and re-screened these. From this, 18 publications were re-labelled from excluded to included, and 7 were re-labelled from included to excluded. The dataset now contained 627 publications labelled as included and 1857 labelled as excluded.

3.4.2.2 Machine classifier k-fold cross validation performance

The labelled data were split into 6 sets (Table 5). Sets 1-5 formed the training and test sets and Set 6 was held as a validation set.

Set	% Total Dataset	Total Publications (n)	Publications Labelled 1 / Included (n)	Publications Labelled 0 / Excluded (n)
1	16	421	110	311
2	16	369	106	263
3	16	390	98	292
4	16	391	95	296
5	16	403	88	315
6	20	510	130	380

Table 5: Number of publications assigned to each set

Sets 1-5 were assigned to train and test sets across 5 iterations to determine algorithm performance at a recall of at least 0.95. Results from the 5 iterations showed the algorithm performance was fairly stable across iterations 1-4 with slightly poorer performance in iteration 5 (Table 6). The chosen machine score threshold ranged from 0.25 to 0.44 (mean 0.35, median 0.36). Specificity ranged from 0.86 to 0.94 (mean 0.91, median 0.93), precision from 0.71 to 0.85 (mean 0.79, median 0.83), F2 score from 0.90 to 0.93 (mean 0.90, median 0.90), and MCC from 0.76 to 0.87 (mean 0.82, median 0.86). The algorithm performance on the validation set (set 6), using the chosen thresholds determined in each iteration appeared consistent with the test set performance, although recall appeared slightly higher and specificity slightly lower (Table 7). The recall ranged from 0.95 to 0.98 (mean 0.97, median 0.97). Specificity ranged from 0.86 to 0.95 (mean 0.91, median 0.92), precision from 0.71 to 0.86 (mean 0.80, median 0.81), F2 score from 0.91 to 0.94 (mean 0.93, median 0.93), and MCC from 0.77 to 0.87 (mean 0.83, median 0.84). Iteration 4 showed the highest F2 score (0.94) and MCC (0.87).

Iteration	Test Set Number	Optimal Threshold	Recall	Specificity	TP	TN	FP	FN	Precision	F1	F2	MCC
1	5	0.36	0.95	0.90	84	282	33	4	0.72	0.82	0.90	0.77
2	4	0.33	0.96	0.94	91	278	18	4	0.83	0.90	0.93	0.86
3	3	0.44	0.96	0.94	94	275	17	4	0.85	0.90	0.93	0.87
4	2	0.39	0.95	0.93	101	245	18	5	0.85	0.90	0.93	0.86
5	1	0.25	0.95	0.86	105	269	42	5	0.71	0.82	0.90	0.76

Table 6: Classifier performance against the test set. Data is shown in each iteration at a pre-specified recall of at least 0.95. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives. MCC = Matthew's correlation coefficient.

Iteration	Chosen Threshold	Recall	Specificity	TP	TN	FP	FN	Precision	F1	F2	MCC
1	0.36	0.97	0.92	126	350	30	4	0.81	0.88	0.93	0.84
2	0.33	0.98	0.90	127	343	37	3	0.77	0.86	0.93	0.82
3	0.44	0.95	0.95	123	360	20	7	0.86	0.90	0.93	0.87
4	0.39	0.97	0.94	126	357	23	4	0.85	0.90	0.94	0.87
5	0.25	0.99	0.86	128	328	52	2	0.71	0.83	0.91	0.77

Table 7: Classifier performance against the validation set. Data is shown in each iteration at the threshold found in test set assessments (Table 6). TP = true positives; TN = true negatives; FP = false positives; FN = false negatives. MCC = Matthew's correlation coefficient.

3.4.3 Performance of NER for extracting animal model features

3.4.3.1 Named entity recognition development

Searches were updated on 4 October 2024 to retrieve newly published publications and search results were deduplicated (see full details and numbers in Chapter 4 Results). Following deduplication, 151,319 unique publications remained from the search and 14,867 (9.8%) of these were included by the classifier. Of those, I was able to retrieve a PMCID for 9,603 publications and a full text XML file for 5,330 publications (35.9% of included publications), and a machine-readable PDF file for 11,140 publications (74.9% of included publications). Combined, 4,653 publications had both an XML files and machine-readable PDF file (31.3% of included publications); the random sample of 400 publications for NER development, testing, and validation was chosen from within this set.

Of the 400 publications randomly sampled, only 105 (26.3%) met the inclusion criteria. 109 were relevant on most of the criteria expect but used a genetic model that was not on the inclusion list; 55 were animal studies that did not use a genetic model (used e.g., valproate models); 117 were not animal studies; and 14 were not primary research articles. I did not annotate publications that did not meet the inclusion criteria. Set 1, used to initially test and develop the regexes, contained 200 publications (52 relevant and 148 not relevant). Set 2, used to validate the NER approaches once optimised on Set 1, contained 200 publications (53 relevant and 147 not relevant). Optimal performance was determined, through consultation with laboratory researchers, as achieving high recall primarily with high precision as a secondary goal. Therefore, high F2 score was chosen as the performance metric to determine best performance.

For model gene, highest recall (0.98) was achieved in both TXT formats, although these also had the lowest precision. A slightly lower recall but higher precision was achieved from title and abstract text (Table 8). For animal species, all methods had a recall of 0.98 or higher and highest precision was in title and abstract text. For

animal sex, recall was highest in full text (0.97 and higher) and methods only (0.88) formats, but precision was lower than title and abstract text. Finally, for experimental tests, TXT formats gave the highest recall (0.70-0.71) although title and abstract and methods only text gave higher precision. Overall, precision was low (0.41 or less) for all NER types and text formats. Best F2 scores were achieved through title and abstract text for model gene and species, and full text formats for sex and experimental tests.

3.4.3.2 Named entity recognition improvement

I adapted several of the regexes that had resulted in false negatives in Set 1. This included (1) adding full gene names and strongly associated syndrome names to all gene model regexes; (2) adding eighteen new animal species regexes so that now regexes covered all species listed on the Understanding Animal Research species webpage (Understanding Animal Research, 2025); (3) adding additional synonyms to 15 of the 45 experimental test regexes; and (4) adding two additional experimental text regexes (47 in total). I was unable to write regexes to capture all of the tests mentioned in the Set 1 texts as some were not reported in enough detail to determine exactly what experimental procedure had been used, for instance a publication mentioning “social interaction test” could actually be referring to the Three Chamber Social Interaction Test, Tube Test for Social Dominance, or perhaps a novel test, but without methodological details this could not be eluded. Animal sex regexes remained unchanged.

The updated regexes were re-run on the Set 1 texts, and performance was assessed as previously, and additionally at various thresholds (i.e., at threshold 1 the regex must match the named entity at least once within the text to count as a recorded match, at threshold 2 the regex must match the named entity at least twice, and so on). I hoped that requiring named entities to appear more than once in text might help improve the precision. I found that thresholds over 1 increased the number of false positives, therefore, to optimise recall one mention of each NER is optimal (Table 9). NER run on title and abstract text gave the highest F2

scores for model gene and animal species, and on full text formats for animal sex and experimental tests.

I found that only extracting model gene and animal species where the respective named entities both appear in the same sentence improved the performance of the NER for model gene and species at a standard threshold of 1 match required (Table 10). Title and abstract was still the text format that gave the highest F2 score; F2 scores improved by 0.05 and 0.16 for model gene and animal species respectively. However, this improvement was caused by an increase in specificity and precision but a decrease in recall.

Similarly, I found that using a less restrictive filter, only extracting model gene and animal species where the respective named entities both appear in the same text format, helped improve F2 score of the NER for model gene and animal species (Table 11). Title and abstract was still the text format that gave the highest F2 score; F2 scores improved by 0.05 and 0.14 for model gene and animal species respectively. Recall decreased for both but not as much as in Table 10.

I found only extracting animal sex where terms such as “were used” or “we included” are mentioned in the same sentence decreased the F2 score when run on title and abstract text and increased the F2 score when run on other text formats, however these increases in F2 were due to improved specificity and precision but poorer recall (Table 12).

Finally, I found NER of animal sex and experimental tests improved when I only extracted these from publications where both model gene and animal species were extracted from title and abstract. The F2 score of animal sex regexes improved by 0.18 and F2 score of experimental test regexes improved by 0.09, and full text formats gave the highest F2 scores (Table 13).

Type	Format	TP	TN	FP	FN	Recall	Specificity	Precision	F1	F2
Model Gene	TiAb*	48	148	77	5	0.91	0.66	0.38	0.54	0.71
	TXT (pdftools)	52	148	485	1	0.98	0.23	0.10	0.18	0.35
	TXT (pdftotext)	52	148	487	1	0.98	0.23	0.10	0.18	0.35
	XML	51	148	325	2	0.96	0.31	0.14	0.24	0.43
	XML (methods)	46	148	117	7	0.87	0.56	0.28	0.43	0.61
Animal Species	TiAb*	51	148	125	1	0.98	0.54	0.29	0.45	0.66
	TXT (pdftools)	52	148	348	0	1.00	0.30	0.13	0.23	0.43
	TXT (pdftotext)	52	148	349	0	1.00	0.30	0.13	0.23	0.43
	XML	52	148	271	0	1.00	0.35	0.16	0.28	0.49
	XML (methods)	51	148	175	1	0.98	0.46	0.23	0.37	0.59
Animal Sex	TiAb	21	155	36	47	0.31	0.81	0.37	0.34	0.32
	TXT (pdftools)	68	155	242	0	1.00	0.39	0.22	0.36	0.58
	TXT (pdftotext)	68	155	242	0	1.00	0.39	0.22	0.36	0.58
	XML*	66	155	225	2	0.97	0.41	0.23	0.37	0.59
	XML (methods)	60	155	180	8	0.88	0.46	0.25	0.39	0.59
Exp. Test	TiAb	19	161	27	163	0.10	0.86	0.41	0.17	0.12
	TXT (pdftools)	128	161	322	54	0.70	0.33	0.28	0.41	0.54
	TXT (pdftotext)*	129	161	342	53	0.71	0.32	0.27	0.40	0.54
	XML	122	161	298	60	0.67	0.35	0.29	0.41	0.53
	XML (methods)	91	161	150	91	0.50	0.52	0.38	0.43	0.47

Table 8: Initial NER performance at the level of individual annotations. Formats with optimal F2 scores are highlighted with an asterisk. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

Type	Format	Threshold	TP	TN	FP	FN	Recall	Specificity	Precision	F2
Model Gene	TiAb*	1*	50	148	84	3	0.94	0.64	0.37	0.72
		2	48	148	84	5	0.91	0.64	0.36	0.70
		3	44	148	84	9	0.83	0.64	0.34	0.65
		4	38	148	84	15	0.72	0.64	0.31	0.57
		5	33	148	84	20	0.62	0.64	0.28	0.50
	TXT (pdftools)	1*	52	148	533	1	0.98	0.22	0.09	0.33
		2	52	148	533	1	0.98	0.22	0.09	0.33
		3	52	148	533	1	0.98	0.22	0.09	0.33
		4	51	148	533	2	0.96	0.22	0.09	0.32
		5	51	148	533	2	0.96	0.22	0.09	0.32
	TXT (pdftotext)	1*	52	148	541	1	0.98	0.21	0.09	0.32
		2	52	148	541	1	0.98	0.21	0.09	0.32
		3	52	148	541	1	0.98	0.21	0.09	0.32
		4	51	148	541	2	0.96	0.21	0.09	0.32
		5	51	148	541	2	0.96	0.21	0.09	0.32
	XML	1*	52	148	354	1	0.98	0.29	0.13	0.42
		2	51	148	354	2	0.96	0.29	0.13	0.41
		3	50	148	354	3	0.94	0.29	0.12	0.41
		4	50	148	354	3	0.94	0.29	0.12	0.41
		5	49	148	354	4	0.92	0.29	0.12	0.40
	XML (methods)	1*	46	148	122	7	0.87	0.55	0.27	0.61
		2	43	148	122	10	0.81	0.55	0.26	0.57
		3	38	148	122	15	0.72	0.55	0.24	0.51
		4	30	148	122	23	0.57	0.55	0.20	0.41
		5	24	148	122	29	0.45	0.55	0.16	0.34

Animal Species	TiAb*	1*	51	148	126	1	0.98	0.54	0.29	0.66
		2	44	148	126	8	0.85	0.54	0.26	0.58
		3	41	148	126	11	0.79	0.54	0.25	0.55
		4	30	148	126	22	0.58	0.54	0.19	0.41
		5	25	148	126	27	0.48	0.54	0.17	0.35
	TXT (pdftools)	1*	52	148	592	0	1.00	0.20	0.08	0.31
		2	52	148	592	0	1.00	0.20	0.08	0.31
		3	52	148	592	0	1.00	0.20	0.08	0.31
		4	52	148	592	0	1.00	0.20	0.08	0.31
		5	52	148	592	0	1.00	0.20	0.08	0.31
	TXT (pdftotext)	1*	52	148	587	0	1.00	0.20	0.08	0.31
		2	52	148	587	0	1.00	0.20	0.08	0.31
		3	52	148	587	0	1.00	0.20	0.08	0.31
		4	52	148	587	0	1.00	0.20	0.08	0.31
		5	52	148	587	0	1.00	0.20	0.08	0.31
	XML	1*	52	148	457	0	1.00	0.24	0.10	0.36
		2	52	148	457	0	1.00	0.24	0.10	0.36
		3	52	148	457	0	1.00	0.24	0.10	0.36
		4	52	148	457	0	1.00	0.24	0.10	0.36
		5	52	148	457	0	1.00	0.24	0.10	0.36
	XML (methods)	1*	51	148	336	1	0.98	0.31	0.13	0.43
		2	50	148	336	2	0.96	0.31	0.13	0.42
		3	48	148	336	4	0.92	0.31	0.13	0.41
		4	48	148	336	4	0.92	0.31	0.13	0.41
		5	47	148	336	5	0.90	0.31	0.12	0.40
Animal Sex	TiAb	1*	21	155	36	47	0.31	0.81	0.37	0.32

		2	13	155	36	55	0.19	0.81	0.27	0.20
		3	12	155	36	56	0.18	0.81	0.25	0.19
		4	7	155	36	61	0.10	0.81	0.16	0.11
		5	5	155	36	63	0.07	0.81	0.12	0.08
	TXT (pdftools)	1*	68	155	242	0	1.00	0.39	0.22	0.58
		2	63	155	242	5	0.93	0.39	0.21	0.55
		3	58	155	242	10	0.85	0.39	0.19	0.51
		4	50	155	242	18	0.74	0.39	0.17	0.44
		5	45	155	242	23	0.66	0.39	0.16	0.40
	TXT (pdftotext)	1*	68	155	242	0	1.00	0.39	0.22	0.58
		2	63	155	242	5	0.93	0.39	0.21	0.55
		3	57	155	242	11	0.84	0.39	0.19	0.50
		4	51	155	242	17	0.75	0.39	0.17	0.45
		5	44	155	242	24	0.65	0.39	0.15	0.39
	XML*	1*	66	155	225	2	0.97	0.41	0.23	0.59
		2	60	155	225	8	0.88	0.41	0.21	0.54
		3	49	155	225	19	0.72	0.41	0.18	0.45
		4	41	155	225	27	0.60	0.41	0.15	0.38
		5	35	155	225	33	0.51	0.41	0.13	0.33
	XML (methods)	1*	60	155	180	8	0.88	0.46	0.25	0.59
		2	52	155	180	16	0.76	0.46	0.22	0.52
		3	31	155	180	37	0.46	0.46	0.15	0.32
		4	20	155	180	48	0.29	0.46	0.10	0.21
		5	13	155	180	55	0.19	0.46	0.07	0.14
Exp. Tests	TiAb	1*	20	161	29	162	0.11	0.85	0.41	0.13
		2	0	161	29	182	0.00	0.85	0.00	0.00

		3	0	161	29	182	0.00	0.85	0.00	0.00
		4	0	161	29	182	0.00	0.85	0.00	0.00
		5	0	161	29	182	0.00	0.85	0.00	0.00
	TXT (pdftools)*	1*	138	161	347	44	0.76	0.32	0.28	0.57
		2	130	161	347	52	0.71	0.32	0.27	0.54
		3	121	161	347	61	0.66	0.32	0.26	0.51
		4	103	161	347	79	0.57	0.32	0.23	0.44
		5	91	161	347	91	0.50	0.32	0.21	0.39
	TXT (pdftotext)	1*	139	161	367	43	0.76	0.30	0.27	0.56
		2	130	161	367	52	0.71	0.30	0.26	0.53
		3	123	161	367	59	0.68	0.30	0.25	0.50
		4	110	161	367	72	0.60	0.30	0.23	0.46
		5	94	161	367	88	0.52	0.30	0.20	0.40
	XML	1*	131	161	323	51	0.72	0.33	0.29	0.55
		2	98	161	323	84	0.54	0.33	0.23	0.43
		3	82	161	323	100	0.45	0.33	0.20	0.36
		4	65	161	323	117	0.36	0.33	0.17	0.29
		5	55	161	323	127	0.30	0.33	0.15	0.25
	XML (methods)	1*	99	161	163	83	0.54	0.50	0.38	0.50
		2	63	161	163	119	0.35	0.50	0.28	0.33
		3	31	161	163	151	0.17	0.50	0.16	0.17
		4	19	161	163	163	0.10	0.50	0.10	0.10
		5	13	161	163	169	0.07	0.50	0.07	0.07

Table 9: Updated NER performance at the level of individual annotations. Asterisk highlights best performance. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

Type	Format	TP	TN	FP	FN	Recall	Specificity	Precision	F2
Model Gene	TiAb*	48	148	42	5	0.91	0.78	0.53	0.79
	TXT (pdftools)	52	148	272	1	0.98	0.35	0.16	0.49
	TXT (pdftotext)	51	148	268	2	0.96	0.36	0.16	0.48
	XML	50	148	173	3	0.94	0.46	0.22	0.57
	XML (methods)	45	148	57	8	0.85	0.72	0.44	0.72
Animal Species	TiAb*	48	148	36	4	0.92	0.80	0.57	0.82
	TXT (pdftools)	52	148	233	0	1.00	0.39	0.18	0.53
	TXT (pdftotext)	52	148	220	0	1.00	0.40	0.19	0.54
	XML	51	148	174	1	0.98	0.46	0.23	0.59
	XML (methods)	44	148	88	8	0.85	0.63	0.33	0.65

Table 10: NER performance where entities only extracted if model gene and species appeared in same sentence. Run on Set 1 texts at the level of individual annotations. Asterisk shows text format with highest F2 score. Data shown at a threshold of one match required. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

Type	Format	TP	TN	FP	FN	Recall	Specificity	Precision	F1
Model Gene	TiAb*	49	148	59	4	0.92	0.71	0.45	0.77
	TXT (pdftools)	52	148	533	1	0.98	0.22	0.09	0.33
	TXT (pdftotext)	52	148	541	1	0.98	0.21	0.09	0.32
	XML	52	148	343	1	0.98	0.30	0.13	0.43
	XML (methods)	46	148	114	7	0.87	0.56	0.29	0.62
Animal Species	TiAb*	49	148	50	3	0.94	0.75	0.49	0.80
	TXT (pdftools)	52	148	546	0	1.00	0.21	0.09	0.32
	TXT (pdftotext)	52	148	541	0	1.00	0.21	0.09	0.32
	XML	52	148	368	0	1.00	0.29	0.12	0.41
	XML (methods)	46	148	194	6	0.88	0.43	0.19	0.51

Table 11: NER performance where entities only extracted if model gene and species appeared in same text format. Run on Set 1 texts at the level of individual annotations. Asterisk shows text format with highest F2 score. Data shown at a threshold of one match required. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

Type	Format	TP	TN	FP	FN	Recall	Specificity	Precision	F1
Animal Sex	TiAb	0	155	3	68	0.00	0.98	0.00	0.00
	TXT (pdfutils)	52	155	110	16	0.76	0.58	0.32	0.60
	TXT (pdftotext)*	54	155	113	14	0.79	0.58	0.32	0.62
	XML	49	155	109	19	0.72	0.59	0.31	0.57
	XML (methods)	40	155	92	28	0.59	0.63	0.30	0.50

Table 12: NER performance where entities only extracted if animal sex appeared alongside additional text. Additional text included (was / were / we) (included / analysed / used / conducted / assessed / generated). Run on Set 1 texts at the level of individual annotations. Asterisk shows text format with highest F2 score. Data shown at a threshold of one match required. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives

Type	Format	TP	TN	FP	FN	Recall	Specificity	Precision	F1
Animal Sex	TiAb	21	155	12	47	0.31	0.93	0.64	0.34
	TXT (pdfutils)*	66	155	93	2	0.97	0.63	0.42	0.77
	TXT (pdftotext)*	66	155	93	2	0.97	0.63	0.42	0.77
	XML	64	155	84	4	0.94	0.65	0.43	0.76
	XML (methods)	58	155	67	10	0.85	0.70	0.46	0.73
Exp. Tests	TiAb	19	161	13	163	0.10	0.93	0.59	0.13
	TXT (pdfutils)*	131	161	133	51	0.72	0.55	0.50	0.66
	TXT (pdftotext)*	132	161	140	50	0.73	0.53	0.49	0.66
	XML	126	161	121	56	0.69	0.57	0.51	0.65
	XML (methods)	94	161	65	88	0.52	0.71	0.59	0.53

Table 13: NER performance where entities only extracted in publications matching model gene and animal species entities on title and abstract text. Run on Set 1 texts at the level of individual annotations. Data shown at a threshold of one match required. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives.

Type	Format	TP	TN	FP	FN	Recall	Specificity	Precision	F2
Model Gene	TiAb*	53	147	73	1	0.98	0.67	0.42	0.77
	TXT (pdftools)	53	147	450	1	0.98	0.25	0.11	0.37
	TXT (pdftotext)	53	147	453	1	0.98	0.25	0.10	0.37
	XML	53	147	316	1	0.98	0.32	0.14	0.45
	XML (methods)	46	147	88	8	0.85	0.63	0.34	0.66
Animal Species	TiAb*	53	147	53	0	1.00	0.74	0.50	0.83
	TXT (pdftools)	53	147	522	0	1.00	0.22	0.09	0.34
	TXT (pdftotext)	53	147	516	0	1.00	0.22	0.09	0.34
	XML	53	147	371	0	1.00	0.28	0.13	0.42
	XML (methods)	47	147	161	6	0.89	0.48	0.23	0.56
Animal Sex	TiAb	12	157	9	46	0.21	0.95	0.57	0.24
	TXT (pdftools)	58	157	101	0	1.00	0.61	0.36	0.74
	TXT (pdftotext)	58	157	101	0	1.00	0.61	0.36	0.74
	XML*	58	157	91	0	1.00	0.63	0.39	0.76
	XML (methods)	54	157	73	4	0.93	0.68	0.43	0.75
Exp. Test	TiAb	25	162	17	88	0.22	0.91	0.60	0.25
	TXT (pdftools)	82	162	139	31	0.73	0.54	0.37	0.61
	TXT (pdftotext)*	84	162	143	29	0.74	0.53	0.37	0.62
	XML	79	162	123	34	0.70	0.57	0.39	0.60
	XML (methods)	63	162	70	50	0.56	0.70	0.47	0.54

Table 14: NER performance with multiple constraints. Model gene and animal species entities only extracted where model gene and species both appeared in the title and abstract, and animal sex and experimental test entities were only extracted in publications matching model gene and animal species. Run on Set 2 texts at the level of individual annotations. Asterisk shows text format with highest F2 score. Data shown at a threshold of one match required. TP = true positives; TN = true negatives; FP = false positives; FN = false negatives. z

3.4.3.3 Named entity recognition validation

Through validating the optimal NER approach (extracting model gene and animal species only where they both appear in the title and abstract and extracting animal sex and experimental tests only in publications where model gene and animal species have been matched), I found model gene and animal species F2 scores run on title and abstract text to be consistent compared with Table 11 (no difference and decrease of 0.03 respectively) (Table 14). F2 scores for animal sex were consistent with Table 13 on full text formats text formats (decrease of 0.00 to 0.03) and experimental tests were consistent with Table 13 on full text formats (decrease of 0.04 to 0.05). The text formats with highest F2 scores in Set 2 were broadly consistent with Set 1 and showed very little different between XML and TXT formats. Despite many F2 scores being somewhat lower in Set 2 than Set 1, recall values were higher.

3.5 Discussion

Here, I was able to develop, test, and validate an approach to systematically identify monogenic NDC animal research publications, using complex searches of scientific databases, a trained binary machine learning classifier, and dictionary-based NER. My search strategy performs with a recall of 0.9318 when tested on a sample of 484 known publications from the AutDB animal resource, meaning it can retrieve the vast majority of relevant publications but still misses some relevant publications. The search appears more likely to miss older publications than more recently published research. I succeeded in training a machine learning classifier to identify relevant publications from my search strategy. In stratified 5-fold cross validation, the algorithm performance was fairly stable and performed with an average F2 score of 0.93. And finally, I was able to use NER to extract model gene, animal species, animal sex, and experimental tests from texts included by the classifier. Through testing and validation of regexes, I was able to adapt and improve NER approaches. The best performing approach involved only

extracting animal sex and experimental tests from publications that matched with both model gene and animal species entities in the title and abstract text.

3.5.1 Comparison with other approaches

While use of APIs to query and retrieve records from databases is a common automation method (Bannach-Brown et al., 2021; Thomas et al., 2017), a plethora of approaches have been tested to automate or semi-automate more cognitive tasks including screening for inclusion and identifying features in texts. For screening, these include use of different technology for screening, including SVMs (Bannach-Brown et al., 2019; Berrang-Ford et al., 2021; Hair et al., 2024; Noel-Storr et al., 2021; Shemilt et al., 2022) and BERT models, thought to more accurately handle complex language (Callaghan et al., 2021; Lück et al., 2025). For identification of features, some evidence syntheses have opted to use regex, which can be used to identify occurrences of particular patterns of text (Hair et al., 2024); topic modelling, an unsupervised machine learning that can identify “topics” emerging from clusters of words frequently appearing together (Berrang-Ford et al., 2021); or pre-trained BERT models that can be used for multi-classification of texts against a pre-specified set of “topics” (Callaghan et al., 2021).

Other approaches include priority screening, using machine learning to rank publications as they are screened with more likely relevant publications being presented to screeners earlier once the algorithm is trained (O'Mara-Eves et al., 2015). Although this does not reduce the number of publications needed to screen, it can help ensure a larger proportion of included publications are screened quicker, especially where resources for screening are limited. The use of stopping criteria alongside priority screening approaches has been proposed to determine when screening can safely stop without missing relevant literature; however recent work has found these methods are rarely robust (Repke et al., 2025).

Other approaches include using machine learning alongside manual screening. This may be helpful to ensure included records are actually relevant; as with a low

inclusion rate, even a high precision value can still result in many false positives (discussed in Section 3.5.2 Strengths and limitations). The COVID-19 Study Register aimed to include primary, human studies broadly relevant to COVID-19 and was trained on 59,513 screened records, of which 20,878 were included studies (Shemilt et al., 2022). The generation of such an extremely large labelled dataset was only possible thanks to crowdsourcing via the Cochrane Crowd platform. Cochrane Crowd is Cochrane's "citizen science" platform, allowing people with internet access from around the world to contribute to "microtasks" (such as labelling records as relevant or irrelevant to an eligibility criteria) (Noel-Storr et al., 2021). In another project, using 1,021,277 labelled data from 188,897 Cochrane Crowd contributors, Cochrane have been able to identify 178,437 randomised controlled trials (RCTs) for their Central Register of Randomised Controlled Trials with high recall. Adding machine-learning to the workflow allowed "undoubted gains" in the processing of records, however, the authors highlight that using machine learning for the "easy to reject" records meant that the records available for manual screening were more difficult and the microtask became less suitable for beginners, who made up the majority of the Cochrane Crowd (Noel-Storr et al., 2021).

Overall, the approach used to identify likely relevant research literature are highly dependent on both the particularities of the research domain, such as clarity and complexity of reporting, and the needs of particular evidence syntheses. There is no single right or wrong approach, but there is a need for clearer guidance on how best to choose appropriate methods and report transparently the methods used.

3.5.2 Strengths and limitations

A key strength of my approach is that I worked with a subject matter expert and academic librarian to develop the search strategy. Working with an academic librarian ensured I was able to test the search adequately. Meanwhile, working with a subject matter expert ensured I was able to capture the breadth of terms that could be used to describe relevant research while also limiting the terms as much as possible to avoid retrieving too many irrelevant records. This included the

decision to search for common animal model species names rather than using the entire animal search filter I used previously in Chapter 2. I discuss the animal search filter further in Chapter 6.

A key challenge was capturing relevant publications and information in searches, machine learning classification, and NER extraction. Broadly, inconsistent reporting within biomedical research publications, including both animal and *in vitro* research, mean that publications will be missed if they do not report key terms in their title and abstract text (Hair, 2022; Wilson, Cruz, et al., 2023).

Optimising identification strategies can only go so far, and I believe that the most progress will be made by adhering to minimum reporting standards. The ARRIVE 2.0 Guidelines (criteria 11) and original ARRIVE guidelines (items 1 and 2) recommend reporting animal species, strain, and sex details, as well as accurate methods, results, and conclusions in the publication abstract to aid transparency and systematic reviews (Kilkenny et al., 2010; Percie Du Sert et al., 2020). Uptake of these reporting guidelines will likely improve the recall of systematic review searches, screening algorithms, and NER systems. However, this does rely on both authors being aware of the guidelines and journal editors only publishing work which adheres to the guidelines.

Bias within datasets used in to train machine learning algorithms and classifiers can impact how these approaches perform and can cause real-world harms. Over-representation of certain features within training sets can lead algorithms to assume those features as the “default”. For instance, computer scientist Dr Abeba Birhane has previously shown that after inputting photos of her own face into an algorithm designed to depixelise blurred pictures of faces, the algorithm output white faces and in one example a masculine face (Birhane & Guest, 2020). This output occurred because the developers used a dataset of majority white faces and is an example of a pattern of racism that underlies many human data training sets used in machine learning and artificial intelligence (AI) (Birhane & Guest, 2020).

There are several biases in my own dataset. For instance, some NDCs (e.g., FMR1) are more commonly researched than others, some animal species (e.g., mice) are more common than others, and some countries may produce more research than others. Even the fact I have primarily searched English-language databases serves as a bias. It is unlikely that datasets will ever be completely free of bias, but it is important that I seek to address these issues as much as possible to ensure that I limit any negative impacts of these biases.

I took several approaches to mitigate the bias in my dataset as much as possible. The labelled data I used to train, test, and validate the algorithm came from a random sample of systematic search results supplemented by known relevant publications covering several genetic models from the AutDB database. While this approach ensured that a variety of publications were represented in the dataset, it still likely means certain types of publication are more represented than others. Taking a stratified random sample, where different features such as those listed above are more equally represented in the dataset, could have better mitigated this bias. However, this would likely introduce new challenges as some NDCs may be researched in very few (e.g., less than ten) publications and machine learning classifiers require thousands of labelled data.

I was able to limit some of the bias in my assessment of my classifier's performance by using a stratified 5-fold cross validation approach. In hindsight, I would have conducted the error correction stage during the cross validation instead of before, therefore allowing errors across the whole labelled dataset to be corrected. However, I only learned about the stratified 5-fold cross validation approach after this stage was complete. Prior to beginning my PhD, I trained a machine learning classifier to identify publications reporting *in vitro* research (Wilson, Cruz, et al., 2023). While the classifier achieved a high performance on sensitivity and specificity metrics, I did not account for bias in my dataset, and I did not validate the performance on a separate dataset so cannot be sure of the stability of the classifier performance when shown different publication texts. Additionally, I did not know at the time which performance metrics were most

appropriate to use nor did I know that k-fold validation was recommended to check classifier stability.

When analysing NER approaches, it became apparent that despite the high performance values seen in the machine learning classifier, many false positive publications still remain after classification. However, considering the mathematics behind this equation and considering the massive class imbalance of the systematic search results (irrelevant publications massively outweighing relevant publications), this finding is not surprising. The classifier excluded 136,452 publications from the deduplicated search results (151,319 publications reduced to 14,867) meaning only 9.8% of the search results were included. As the specificity of the classifier is less than 100%, we expect to find a small number of false positives in the included dataset. However, because of the class imbalance, that small number of false positives (compared with true negatives) actually represents quite a large proportion of the included dataset (false positives compared with true positives). As mentioned above, the ARRIVE guidelines (both the initial guidelines published in 2010 and the updated 2.0 guidelines published in 2020) already recommend authors describe animal models and species in title and abstract text to aid discoverability (Kilkenny et al., 2010; Percie Du Sert et al., 2020). Adherence to these guidelines would likely reduce class imbalance in search results (meaning less irrelevant publications in searches) and may help the performance of classifiers.

In NER development, using separate testing and validation sets helped ensure that performance was stable when shown new publication texts. However, a limitation was that not all entities were represented in samples and so not all regular expressions were adequately tested. This was a challenge because mitigating this limitation would involve having known examples of each entity.

A further limitation is the lack of availability of full text files. Not all publications had a full text in any file format, a minority of publications did not have a DOI to retrieve a full text, and less than half the publications in my included dataset (35.9%) had a full text XML file. While I expected XML methods section to perform

better than full text, a reason for the poorer performance may be that some XML files included only a brief methods section or no methods section, instead linking to online supplements accessed via the journal website. Text extraction methods like NER rely on the necessary text being contained within the main article full text file. The decision of how methods are made available and presented in full text files is at the discretion of individual journals.

As part of my validation of NER-approaches, I compared two software programmes to convert PDF to TXT. The Xpdfreader pdftotext software has been used within the CAMARADES lab group for many years as part of the AutoAnnotation function (Liao, 2021) but the code is no longer actively maintained meaning if a dependency the code relies on changed, the code would break. The pdftools R package, however, is actively maintained on CRAN meaning it's more likely to be sustainable in the long-term. I found that there was no difference in the performance of the software, meaning one software did not result in additional false negatives. However, my regexes were very short in length, only intended to match short strings of text. One key issue with PDF to TXT conversion is that two column documents (a format used by many scientific journals) tend to be read across the whole page rather than up and down one column before the next when converting to TXT, meaning sentences can often become broken and words appear in the wrong order. Therefore, the performance of NER on TXT converted documents should be reevaluated where longer regexes are used.

A key strength of my overall approach is that by prioritising recall, I have generated a dataset more likely to include false positives than exclude false negatives, meaning that the precision of the dataset could be improved through manual screening of publications included by the classifier and data extracted using NER approaches. This approach is therefore more feasible than a fully manual approach and likely to be more sustainable long-term.

3.6 Conclusion

To conclude, I was able to successfully develop a workflow to identify published research of genetic animal models used to study neurodevelopmental conditions. This includes a broad search strategy implemented across three data sources, and a trained machine learning classifier. Dictionary-based NER allows extraction of model gene, animal species, animal sex, and experimental tests with high recall, however, is limited by a high number of false positives remaining in the dataset following machine learning classification. However, the dataset remains functional as a research discovery tool due to high recall and, if desired, could be further improved through manual screening of data.

Chapter 4: Automated workflow to summarise research publications

4.1 Context

In this chapter, I use the methods described in the previous chapter as a foundation for a systematic map. Code underlying the generation of this systematic map has been written by myself, Kaitlyn Hair, Sean Smith, and Maria Economou. As a team, we have tried to write code that can be used in systematic map projects across different research areas. In the methods section of this chapter, I have stated where I have written my own code or adapted others code for my own purposes.

4.2 Introduction

Systematic maps aim to provide high-level overviews of a broad research domain (F. Campbell et al., 2023; Hair, Wilson, et al., 2023). Systemic online living evidence summaries (SOLES) are an approach to systematic mapping described by myself and colleagues within the CAMARADES research group, focused on use cases within biomedical research (Hair, Wilson, et al., 2023). By using a broad **systematic** search to identify potentially relevant literature, and a trained machine learning classifier to identify the publications from searches most likely to be relevant (both discussed in the previous chapter), I have generated a dataset that can be used as the starting point of a systematic map, or SOLES, of animal research investigating genetic neurodevelopmental conditions. Through further annotation of this

dataset with existing article metadata and metadata derived through text mining approaches, I can provide an **evidence summary**, in the form of a systematic map, of this dataset. I can make this process '**living**' (i.e., integrating new data as it becomes available) by making use of code software packages to retrieve, process, and annotate new publications. And finally, the resulting systematic map can be shared **online** through a web application allowing users to interrogate and investigate the data.

Searching multiple databases is an essential part of systematic review methods to ensure that as much of the relevant published literature as possible can be captured. This is because not all scientific publications and journals are indexed in all scientific databases. Some journals may be indexed in Web of Science and not in PubMed, and vice versa. This however also means that some publications will be indexed in more than one database. To ensure that systematic search results only contain one copy of each publication, deduplication methods are used to remove duplicate copies of publications. Many scientific databases, such as those that index scientific publications and/or their related metadata, have APIs which can be used to search for and retrieve publications or metadata relating to scientific publications. API calls can be scheduled to retrieve new publications on a regular basis. RISmed is an R package that can be used to retrieve information from National Center for Biotechnology Information (NCBI) databases, including PubMed (Kovalchik, 2011). Similarly, scopusAPI (Belter, 2020) can be used to retrieve information from Scopus (Elsevier), and rwoslite (Casajus & Mouquet, 2022) can be used to retrieve information from the Web of Science databases (Clarivate).

Searching multiple databases means that multiple copies of the same publication will appear in the combined search results. (Bramer et al., 2016) describes a comprehensive manual strategy for removal of duplicates in the reference management software EndNote. More recently, the ASySD software was developed to facilitate both automated and manual duplicate removal (Hair, Babor, et al., 2023). ASySD is available as a web application and an R package, was

able to remove more duplicates than EndNote approaches, and performed with a sensitivity of 0.95 to 0.99 on different test datasets (Hair, Bahor, et al., 2023).

Publication metadata can be retrieved from databases during search result retrieval, including publication title, abstract text, authors, publication year, journal name, PMID (PubMed identifier) and PMCID, and DOI. Additional metadata such as open access status (paywalled, gold open access, green open access, etc.), author institution, grant funder, full text links, and retractions can be identified through additional databases such as CrossRef and OpenAlex. Both of these data sources have associated R packages, *rcrossref* (Chamberlain et al., 2014) and *openalexR* (Aria et al., 2022) respectively. Additionally, OpenAlex links author institution with data from the Research Organization Registry (ROR; <https://ror.org/>), which contains location data such as country and Global Positioning System (GPS) coordinates.

Publication DOIs can be used to download PDF files, which can then facilitate the extraction of further metadata via text mining. (Wang et al., 2022b) recently developed a pre-trained natural language processing software that can be used to estimate the reporting of five measures to reduce the risk of bias in animal experiments. These are: (1) random allocation of animals to experimental groups, (2) blinded assessment of outcome (3) presence of a conflict of interests' statement (regardless of conflicts of interest), (4) statement of compliance with animal welfare regulations, and (5) reporting animals excluded from analysis. When evaluated, the Bidirectional Encoder Representations from Transformers (BERT) transformer model performed with an F1 score above 0.8 for all items except reporting of exclusions which performed with an F1 of 0.466 (Wang et al., 2022b). The software runs in Python but can be run in the R programming language using a wrapper function. Additionally, ODDPub (Open Data Detection in Publications) is a text mining tool that detects data and code availability statements in scientific publications (Riedel et al., 2020). In a validation set, the software performed with a sensitivity of 0.73 and specificity of 0.97, meaning it

likely misses a proportion of shared data but rarely falsely detects shared data (Riedel et al., 2020).

Due to their broad inclusion criteria, the quantity of metadata received from SOLES projects is far greater than typical systematic reviews. As previously alluded to, reference management software such as EndNote are commonly used in systematic review projects to store search results and remove duplicates.

However, from previous experience, reference management software struggles greatly with large datasets in the thousands. Designing a workflow that operates using the R programming language allows the integration of many of existing software tools described above that can be scheduled to run regularly. Relational SQL (Structured Query Language) database tables can be used to store large amount of tabular data that can be quickly read and written in R. Finally, R Shiny Apps is a software that allows the creation of dynamic web application than can be used to allow the sharing and interrogation of data (Chang et al., 2012).

4.2.1 Aims

In this chapter, I aim to implement an automated workflow to build a continually updating systematic map of monogenic NDC animal research publications.

4.3 Methods

4.3.1 Protocol registration

A protocol for the systematic map development is available online on the Open Science Framework (DOI: <https://doi.org/10.17605/OSF.IO/GFTZP>). Deviations between the protocol and methods described in this chapter are:

- Full text retrieval and NER regex tagging were updated following findings from Chapter 3.
- Retracted articles are tagged using data from Retraction Watch
- Open access data, and other data, are retrieved from OpenAlex rather than CrossRef.

4.3.2 Use of previous data

I began with the data collected in the previous chapter: full search results from PubMed, Scopus and Web of Science Core Collection searches, and unique search results following removal of duplicates using the automated deduplication on ASySD. Search result numbers are described in full in the results section of this chapter.

4.3.3 Data storage

I store all data in SQL tables hosted, set up by my colleague Dr Kaitlyn Hair, on Amazon AWS servers. Data in individual tables are linked by unique identifier (UID, described below) or DOI. Data are accessed (reading and writing) in the R programming languages through the DBI R package (R Special Interest Group on Databases (R-SIG-DB) et al., 2001). I run the workflow described below on a weekly basis to identify new publications, which are appended to existing SQL tables. The structure of the SQL tables was determined through discussions with the SOLES team to ensure consistence across projects.

4.3.4 Bibliographic search

The initial systematic search process, retrieving publications published from 1994 onwards, is described in the previous chapter. In the weekly workflow, I conduct systematic searches of PubMed, Scopus, and Web of Science Core Collection using the search terms described in Chapter 3 (also in Appendix 3A). Weekly searches are limited to retrieve only the publications indexed within the last week. Searches are scheduled to run using APIs accessed via the RISmed (Kovalchik, 2011), scopusAPI (Belter, 2020) and rwoslite (Casajus & Mouquet, 2022) R packages to search PubMed, Scopus, and Web of Science Core Collection respectively. I use custom R code, developed by the SOLES Team, to combine these three datasets. At this stage, a unique identifier (UID) is created for each publication by combining the lowercase name of the data source (pubmed, scopus, wos) with the publication's accession number from that data source.

4.3.5 Removal of duplicate publications

I removed duplicates from the larger dataset using the automated and manual deduplication features of the ASySD (Hair, Bahor, et al., 2023) software. I also conducted a further manual check of remaining publications where DOI or PubMed identifier (if available) are duplicated. In the weekly workflow, I remove duplicates using only the automated ASySD functions. My colleague Dr Kaitlyn Hair, who developed the ASySD software, amended the ASySD code to work within this workflow. This includes checking for duplicates not only within newly retrieved publications but also within publications retrieved in the previous two months. This process accounts for differences in the time taken to index publications across different data sources.

4.3.6 Screening using a trained machine learning classifier

The training and validation of the machine learning classifier, and the inclusion criteria for the systematic map, are described previously in Chapter 3. I run the title and abstract text of all publications, with duplicates removed, through the classifier which then labels each publication as “included” or “excluded” based on the optimal training data as determined in Chapter 3. Excluded publications remain in the dataset of unique publications but are not taken forward for further data extraction and are not displayed in outputs. Each week, I pass newly retrieved and deduplicated publications through the classifier for screening.

4.3.7 Full text document retrieval

I retrieve full text documents of included publications as PDF files using custom-made code developed by the SOLES team. The code first attempts to retrieve full text documents from CrossRef via the rcrossref R package (Chamberlain et al., 2014) or UnPaywall via the roadoi R package (Jahn, 2017). Where full texts cannot be retrieved by either of these methods, the code attempt to retrieve full texts using the Elsevier (Elsevier, 2025) and Wiley Text and Data Mining (Wiley, 2025) APIs. Retrieved PDF files are converted to text files via the ‘pdftotext’ software from

XpdfReader (Glyph & Cog, 2025) or the pdftools R package (Ooms, 2016). I retrieve full text in XML format by first retrieving PMCID where available from DOIs via CrossRef using the rcrossref R package (Chamberlain et al., 2014), and then using the europepmc R package (Jahn, 2016) to download an XML file associated with that PMCID if it exists.

4.3.8 Tagging model, animal, and experimental test details

I extract information about model, species, sex, and experimental tests using the optimal NER approach determined in the previous chapter. Details are tagged using a dictionary-based approach which includes commonly used phrases and synonyms. Tags are recorded and applied to publications when certain criteria are met (Table 15); criteria were chosen when testing NER performance in the previous chapter. NER is performed using a dictionary-based regular expressions approach. Where XML full texts are available, these are used in preference over PDF files converted to text files. NER regular expressions are run on title and abstract text and text files using the Quanteda R package (Benoit et al., 2015), and on XML files using the tidyxml and XML2 R packages (Stubben, 2019; Wickham et al., 2015).

Detail	Conditions for recording
Experimental model	At least one NER match in the title and abstract text AND at animal species also matched at least once in the title and abstract text of same publication
Animal species	At least one NER match in the title and abstract text AND experimental model also matched at least once in the title and abstract text of same publication
Animal sex	At least one NER match in the full text AND both experimental model and animal species matched at least once in the title and abstract text of same publication
Experimental tests	At least one NER match in the full text AND both experimental model and animal species matched at least once in the title and abstract text of the same publication

Table 15: Conditions for recording NER matched tags

4.3.9 Using metadata from OpenAlex

As part of the workflow, I annotate included publications with additional metadata from OpenAlex, which contains information from other data sources including CrossRef and Retraction Watch, using the openalexR package (Aria et al., 2022). I wrote the code to implement the openalexR package into the workflow. By searching OpenAlex using publication DOIs, I can retrieve information on retractions, retractions, article abstract language, author institution, grant funder, and discipline, and open access status. Discipline tagging is a feature exclusive to OpenAlex. Additionally, by downloading the Research Organization Registry (ROR) data for author institution, and writing code to call the ROR API, I am able download longitude and latitude coordinates for research institutions, allowing the number of publications produced by first authors are various research institutions to be displayed on a world map produced using the leaflet R package (Cheng et al., 2015). Additionally, as all publications require updated information on retractions and citation count, I wrote an update function to retrieve new information for publications where OpenAlex information was last retrieved over six months ago.

As part of a manual evaluation, I checked the data retrieved from OpenAlex for any inconsistencies or unusual data. Where publications did not have a DOI, I used PubMed identifier where available or publication title to check for retractions manually in the Retraction Watch Database.

4.3.10 Assessing data and code sharing

I use the ODDPub (Riedel et al., 2020) tool to extract information about the availability of open data and code from included publications. The tool applies named entity recognition to full texts (PDFs converted to text) and outputs a binary true or false decision label for each publication as to whether the article contains an open data and / or open code statement. The code for implementing the tool within the workflow was written by the SOLES Team.

4.3.11 Assessing risk of bias reporting

I assess reporting of measures to reduce the risk of bias in included publications where I have been able to retrieve a full text document, using a natural language processing (NLP) tool developed by a former colleague (Wang et al., 2022a). The tool is written in the Python programming language, and the SOLES Team have developed a wrapper function allowing it to be used as part of an R workflow, reading data from PDF files converted to text. The tool is trained to detect reporting of five common measures to reduce the risk of bias in biomedical animal research: (1) random allocation of animals to experimental groups, (2) blinded assessment of outcome, (3) reporting whether or not animals were excluded from analysis, (4) presence of a conflict of interests' statement (regardless of conflicts of interest), and (5) statement of compliance with animal welfare regulations. The tool outputs a decimal score; a score closer to 1 means the item is more likely to be reported and a score closer to 0 means the item is more likely to be not reported. The tool uses threshold of 0.50 to translate decimal score into a reported or not reported decision label.

4.3.12 Web application

Data from included publications are displayed in a web application hosted using R Shiny Apps (Chang et al., 2012). The web application allows users to interrogate, filter, search, and download data. The majority of the app code was written by the SOLES Team, allowing all apps to have a consistent appearance, however I have adapted most of the visualisation code.

4.4 Results

The systematic map dataset is updated weekly. The data below represent a static overview of the dataset as it appeared on 4 October 2024.

4.4.1 The systematic map dataset

As of 4 October 2024, I have retrieved a total of 318,183 publications from database searches (104,557 from PubMed, 107,880 from Scopus and 105,746 from Web of Science Core Collection. After automated removal of duplicates with ASySD, the dataset was reduced to 155,427 unique publications. On 8 October 2024, I conducted an additional round of deduplication on all search results, using both of ASySD's automated and manual features. The additional round of automated deduplication identified and removed 3,470 additional duplicate copies, and through manual deduplication (which is not a typical feature of the automated workflow), I identified and removed an additional 631 duplicate copies. In total, 4,101 additional publications were removed using ASySD and the new total number of unique publications was 151,326. Through further inspection of publications in the new dataset, I identified 4 duplicates (based on DOI) and 3 duplicates (based on PMID) that were not picked up by ASySD. Removing these duplicates resulted in a total set of 151,319 unique publications. 14,867 (9.8%) of the 151,319 unique publications were labelled as meeting the inclusion criteria (Figure 9).

From the 14,867 publications included in my systematic map, 144 (1.0% of included publications) were missing an abstract, and 282 (1.9%) were missing a DOI; 98 of these publications were missing both abstract and DOI. Data were missing from a diverse range of journals published from 1995-2024; from inspection of the data, DOIs were more often missing from journals published outside of Europe or North America. Using DOI where available, I was able to retrieve full text documents (PDF format) for 13,743 publications (92% of included publications). Using PMID and PMCID, I was able to retrieve full texts in XML format for 5,330 publications (36% of included publications).

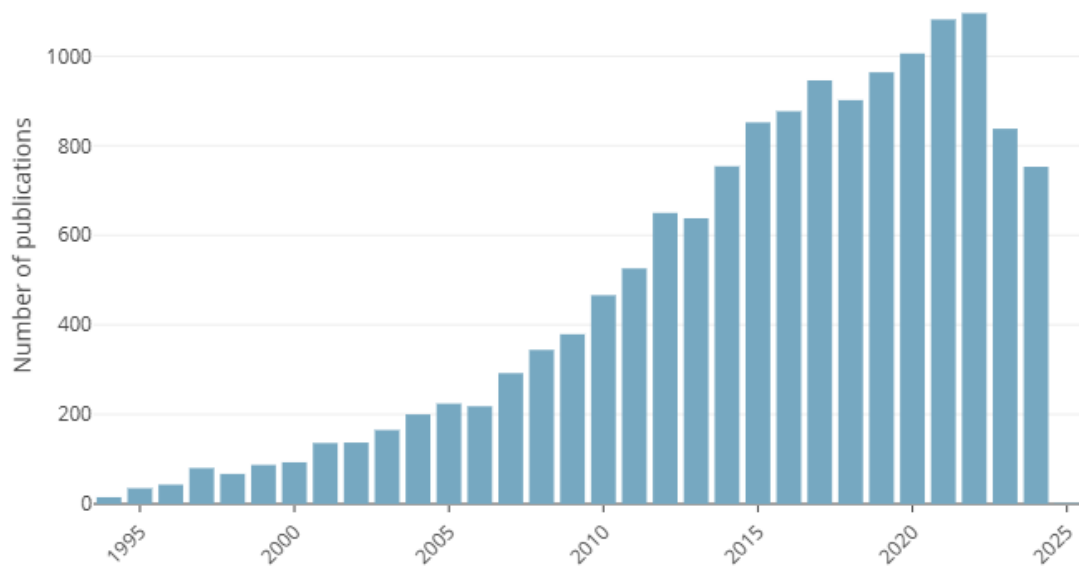


Figure 9: Publications included in NDC-SOLES by publication year. An up-to-date interactive plot is available at: <https://camarades.shinyapps.io/NDC-SOLES/>.

4.4.2 Model, animal, and experimental test detail tags

I ran NER for experimental model, animal species, animal sex, and experimental text extraction on the 14,723 publications with an abstract and 11,599 publications with a machine-readable full text. Not all publications matched NER phrases (Table 16). The most common tags were Fmr1 for experimental model (Figure 10A), mouse for animal species (Figure 10B), male for animal sex, and electrophysiology and open field test for experimental test (Figure 10C).

Detail	Number of publications tagged
Experimental model	6,624
Animal species	6,624
Animal sex	4,597
Experimental test	3,539

Table 16: Number of publications matching NER phrases for different details.

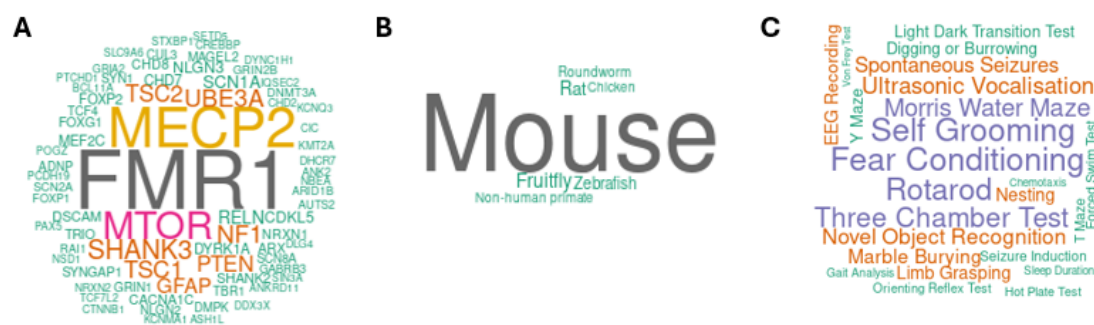


Figure 10: Word clouds of terms matched using regex dictionary-based NER related to (a) genetic model, (b) animal species, and (c) experimental tests. Word frequency > 9.

4.4.3 OpenAlex metadata

4.4.3.1 Retractions

I was able to retrieve retraction information relating to the 14,585 included publications in with a DOI. 12 publications (0.08% of total included) were labelled as retracted (Table 17). This data was correct as of 18 March 2024. On 9th October 2024, I manually checked the 282 publications that do not have a DOI to see whether they appeared in the Retraction Watch database. This did not uncover any additional retractions.

Status	Number of publications
Not retracted	14,573
Retracted	12
Unknown (no DOI)	282

Table 17: Summary of retraction information retrieved from OpenAlex

I was able to investigate the reason for retractions and the time taken to retract publications manually from searches of the Retraction Watch Database. Most retracted publications had multiple reasons for retraction, including image manipulation or duplication; concerns or errors with data, analysis, or the reliability of results; and copyright claims and concerns about authorship.

First author, Year (DOI)	Reasons for retraction	Publication Date	Retraction Date	Time to Retract
Bangash, 2011 (10.1016/j.cell.2011.03.052)	Duplication of Image; Manipulation of Image	2011/05/27	2013/01/17	1-2 years
Deacon, 2017 (10.1111/gbb.12373)	Complaints about Author; Concerns/Issues About Authorship/Affiliation; Objections by Author; Objections by Third Party	2017/02/17	2017/09/01	< 1 year
Feng, 2019 (10.1371/journal.pone.0211066)	Error in Analyses; Unreliable Results	2019/01/25	2021/10/14	> 2 years
He, 2016 (10.1038/srep30368)	Duplication of Image; Investigation by Journal/Publisher; Original Data not Provided; Unreliable Image	2016/08/04	2018/10/30	> 2 years
Kauffman, 2015 (10.1371/journal.pone.0131981)	Copyright Claims; Objections by Author; Objections by Company/Institution	2015/07/06	2017/10/30	> 2 years
Kumar, 2014 (10.1371/journal.pone.0110482)	Concerns/Issues About Data; Concerns/Issues About Image; Duplication of Image	2014/10/17	2019/07/19	> 4 years
Lee, 2013 (10.3389/fncel.2013.00043)	Concerns/Issues About Data; Duplication of Image; Investigation by Journal/Publisher; Unreliable Results	2013/04/18	2023/08/29	> 10 years
Li, 2004 (10.1074/jbc.m401878200)	Notice - Limited or No Information; Removed	2004/05/28	2015/06/05	> 10 years
Lose, 2021 (10.17912/micropub.biology.000407)	Error by Journal/Publisher; Removed	2021/06/24	2021/09/14	< 1 year

Wang, 2021 (10.1155/2021/9983438)	Concerns/Issues About Data; Concerns/Issues About Results; Concerns/Issues about Referencing/Attributions; Concerns/Issues with Peer Review; Investigation by Journal/Publisher; Investigation by Third Party; Paper Mill; Randomly Generated Content; Unreliable Results	2021/04/15	2024/01/24	> 2 years
Yin, 2014 (10.1016/j.ijdevneu.2014.01.006)	Breach of Policy by Author; Copyright Claims; Objections by Third Party	2014/12/01	2015/08/01	< 1 year
Zou, 2011 (10.1111/j.1601- 183x.2011.00702.x)	Duplication of Image; Error in Data; Error in Image; Falsification/Fabrication of Image; Investigation by Company/Institution	2011/06/14	2013/06/22	> 2 years

Table 18: Reasons for retractions and time to retract. Reasons for retraction taken from Retraction Watch database. Dates given YYYY/MM/DD.

One retracted publication was from a suspected paper mill, one was due to a journal or publisher error, and one had a limited retraction notice (Table 18). Only two retractions occurred less than one year following article publication. Most occurred shortly after two years of publication, and two articles were retracted ten years after publication.

4.4.3.2 Publication abstract text language

I was able to retrieve abstract text language for 14,554 publications. Most articles have an English language abstract; this makes sense as I searched predominantly English-language databases (Table 19). Full text language is not indexed in databases such as CrossRef and OpenAlex so I was not able to extract this information, which would be more useful, as articles in languages other than English still often have an English language abstract.

Language	Number of publications
English	14,524
Languages other than English	30
<i>French</i>	9
<i>Spanish</i>	6
<i>Japanese</i>	5
<i>Russian</i>	5
<i>German</i>	1
<i>Greek</i>	1
<i>Italian</i>	1
<i>Polish</i>	1
<i>Romanian</i>	1
Unknown	313
<i>Publication has DOI but no OpenAlex data</i>	31
<i>Publication has no DOI</i>	282

Table 19: Number of publications tagged by language. Data from OpenAlex.

4.4.3.3 First author institution

I was able to retrieve information on the institution of the first author for 13,392 publications. Research in my dataset are published by first authors in 2,295 unique institutions across 80 countries (Table 20). Despite the apparent breadth of research sources, most work is published by first authors at institutions in the Global North; this is not surprising as I searched predominantly English-language databases. OpenAlex identified most research as coming from first authors affiliated with education institutions, such as universities (Table 21). Through identification of longitude and latitude co-ordinates from ROR, I was able to produce a world map showing global affiliations of first authors (Figure 11).

First Author Institution Continent / Country	Number of institutions	Number of publications
Europe	772	3,798
<i>United Kingdom</i>	<i>113</i>	<i>539</i>
North America	748	6,507
South America	57	139
Asia	635	2,685
Africa	9	12
Oceania	61	251
Unknown	---	1,475
<i>Publication has DOI⁵ but no OpenAlex data</i>	---	<i>1,193</i>
<i>Publication has no DOI⁵</i>	---	<i>282</i>

Table 20: Number of institutions and publications per ⁵rst author institution continent. Data from OpenAlex. Data from the United Kingdom is also displayed.

Institution type	Number of publications
Education	8,488
Healthcare	2,007
Facility	1,623
Non-profit	645
Government	348
Company	157
Archive	2
Other	124
Unknown	1,475
<i>Publication has DOI' no OpenAlex data</i>	<i>1,191</i>
<i>Publication has no DOI</i>	<i>282</i>

Table 21: Number of publications per institution type. Data from OpenAlex.

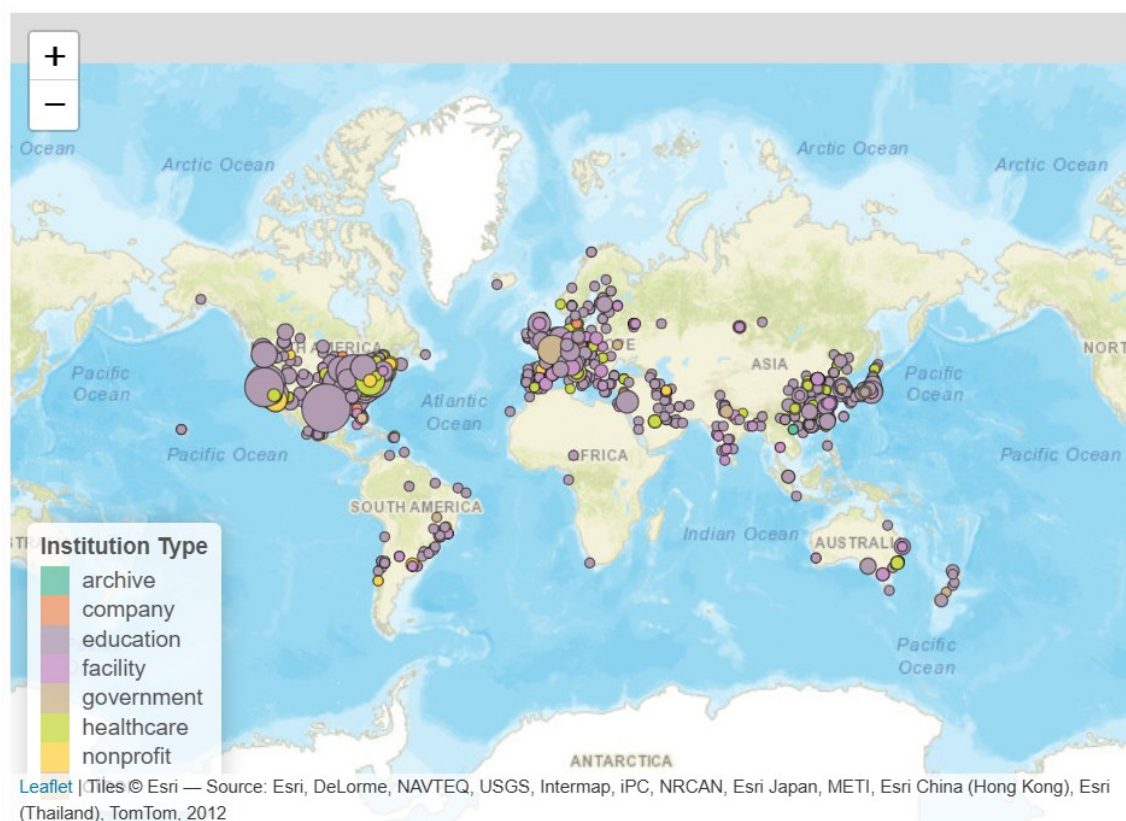


Figure 11: World map showing institutions producing research. The map is produced using the leaflet R Package using data from Open Alex and Research Organization Registry. Location data relates to the first affiliated institution listed for the first author in OpenAlex. An up-to-date interactive plot is available at: <https://camarades.shinyapps.io/NDC-SOLES/>.

4.4.3.4 Grant funders

I was able to retrieve information about research funders and grants for 5,536 publications, which identified 1,898 unique funders. Where funders were not identified, it is unclear if this means research was unfunded or that funder metadata is unavailable. 18 funders were listed on over 100 unique publications (Table 22), however as most funder data is missing it is unclear if the data displayed below is representative of the biggest funders across the full dataset.

Funder	Number of publications
<i>National Institutes of Health</i>	1,146
<i>National Institute of Neurological Disorders and Stroke</i>	545
<i>National Natural Science Foundation of China</i>	476
<i>National Institute of Mental Health</i>	445
<i>Deutsche Forschungsgemeinschaft</i>	280
<i>Japan Society for the Promotion of Science</i>	224
<i>Canadian Institutes of Health Research</i>	215
<i>Simons Foundation</i>	183
<i>National Research Foundation of Korea</i>	166
<i>National Institute of General Medical Sciences</i>	154
<i>Ministry of Education, Culture, Sports, Science and Technology</i>	143
<i>Agence Nationale de la Recherche</i>	138
<i>Eunice Kennedy Shriver National Institute of Child Health and Human Development</i>	127
<i>Japan Agency for Medical Research and Development</i>	121
<i>FRAXA Research Council</i>	106

<i>Medical Research Council</i>	105
<i>Brain and Behaviour Research Foundation</i>	103
<i>Natural Sciences and Engineering Research Council of Canada</i>	101
Unknown	9,331
<i>Publication has DOI' no OpenAlex data</i>	9,049
<i>Publication has no DOI</i>	282

Table 22: Top ⁵ fifteen research funders. Data from OpenAlex.

4.4.3.5 OpenAlex disciplines

Level 0, 1, and/or 2 OpenAlex discipline data with a score of 0.4 or above was available for 14,516 publications, and many publications were tagged with more than one discipline. Level 0 was at the level of subject (e.g., biology, medicine), Level 1 the discipline (e.g., neuroscience, cell biology), and Level 2 the topic (e.g., gene, hippocampus, synapse) (Figure 12A-C).



Figure 12: Word clouds of publication discipline. Data from OpenAlex at (a) level 0 (word frequency > 9), (b) level 1 (word frequency > 9), and (c) level 2 (word frequency > 99).

4.4.3.6 Open access publishing

I was able to retrieve information about open access publication and publishing model for 14,532 publications. OpenAlex labelled 10,888 publications (73% of the total dataset) as being available open access. However, manual inspection of the publishing model (gold, green, closed, etc.) revealed an additional 148 publications were available open access under hybrid, green, or bronze models (11,036 publications in total; 74% of the total dataset). The most common publishing

model overall is gold open access (Table 23, Figure 13A), and the proportion of articles published gold open access has been steadily increasing since around 2010 (Figure 13B).

Publishing model	Number of publications
Open access	10,888
<i>Gold</i>	4,551
<i>Bronze</i>	2,574
<i>Green</i>	2,370
<i>Hybrid</i>	1,524
<i>Diamond</i>	17
Closed access	3,496
Unknown	335
<i>Publication has DOI' no OpenAlex data</i>	53
<i>Publication has no DOI</i>	282

Table 23: Number of publications tagged with different publishing models. Data from OpenAlex. Corrected values shown.

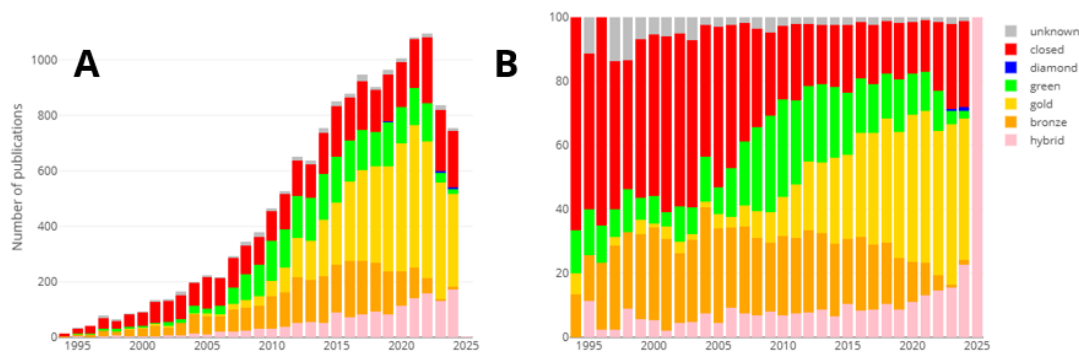


Figure 13: (a) Number of and (b) percent of publications tagged with different publishing models. Corrected data from OpenAlex. An up-to-date interactive plot is available at: <https://camarades.shinyapps.io/NDC-SOLES/>.

Gold open access models rely on authors or their institutions paying a fee (article processing charge; APC) to publish a work that is then made freely available read (Eve, 2014). While closed access represents a small number of articles over time (Figure 13A), the proportion of closed access articles was much greater prior to 2010. Additionally, although we can have seen a decrease in the proportion of

closed access articles since then, their proportion per year has increased over the last three years (Figure 13B).

Bronze, the second most common open access model, is attributed to articles that appear free to read on from the publisher but are not shared with an identifiable open license detailing how a work may be used, reused, or shared (Piwowar et al., 2018). However, as we see a decrease in the proportion on bronze open access articles over time (Figure 13B), this may indicate an improvement in the clarity of licensing.

Green open access refers to a version of an article being made free to read on a repository, typically hosted by a university, although a paid-access version may still be available on a publisher website, while hybrid refers to articles, published in a subscription journal, that are free to read on the publisher website (Eve, 2014). Although green open access offers a way for authors to take control of their work, the freely available version of the publication is often the author accepted manuscript, not the final published version, so there may be minor differences in formatting. Finally, diamond open access journals are both free to read and free to publish in, typically subsidised by university libraries or other funding (Eve, 2014). Diamond open access is still uncommon, especially in STEM research published in the Global North, but the number and proportion of diamond open access publications appears to have grown over 2023-2024 (Figure 13A-B).

4.4.4 Data and code sharing practices

Using ODDPub, I was able to extract information on data and code sharing from 13,471 publications (Figure 14A-D). 2,157 publications (15% of the total included) and 308 (2% of the total included) were tagged by ODDPub as sharing either data or code, respectively. 228 publications were tagged as sharing both data and code (Table 24). Although still low, the proportion of publications sharing data and code has increased over time (Figure 14B, 14D), however it's not clear if untagged papers used code but did not share it or did not use code.

Data and code sharing	Number of publications
Shared either or both of data and code	2,237
Data and code shared	228
Data shared only	1,929
Code shared only	80
Did not share data or code	11,234
Unknown	1,396
<i>Publication has full text ' not machine readable</i>	842
<i>Publication has DOI' no full text retrieved</i>	272
<i>Publication has no DOI</i>	282

Table 24: Number of publications tagged as sharing data and/or code. Data from ODDPub.

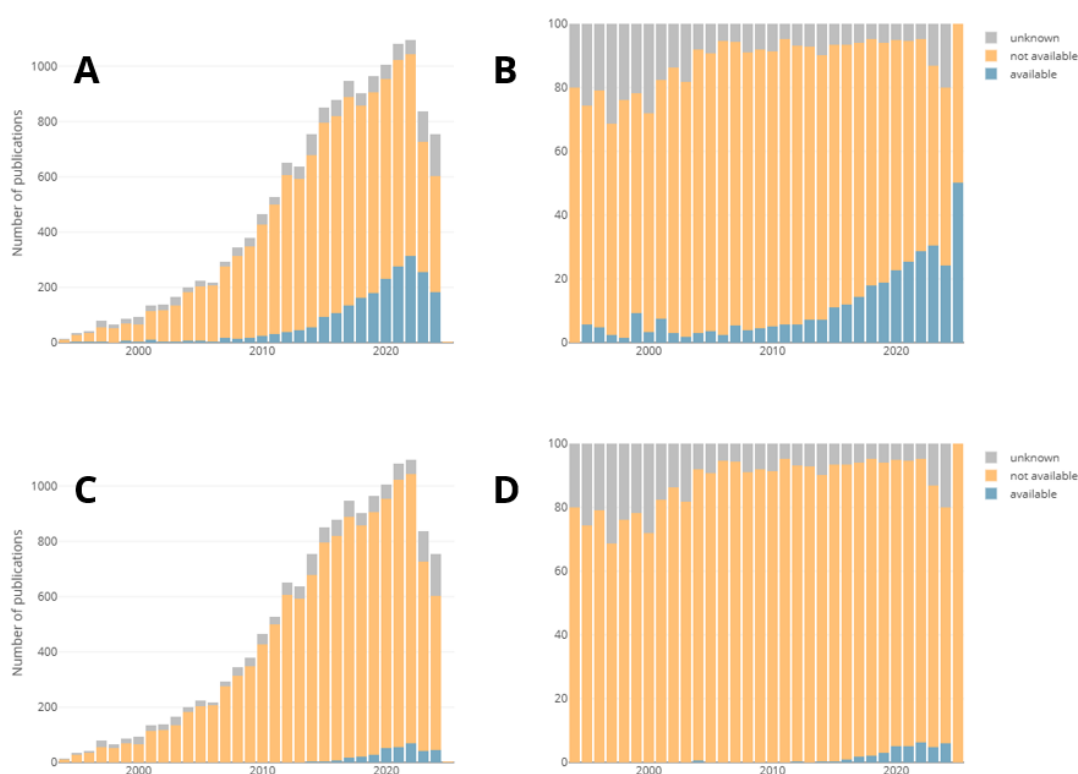


Figure 14: Publications tagged using ODDPub. Number and percent of publications tagged as (a,b) sharing data and (c,d) sharing code, respectively, according to ODDPub. An up-to-date interactive plot is available at: <https://camarades.shinyapps.io/NDC-SOLES/>.

4.4.5 Risk of bias reporting

Using the NLP software, I was able to extract reporting of measures to reduce risk of bias in animal experiments from 13,365 publications (Table 25). 12,109 publications (81% of total included) were detected to have reported at least one measure to reduce risks of bias. Reporting over time for random allocation blinding, exclusions, conflict of interests, and welfare are shown in Figure 15A-J. Randomisation is less likely to be relevant in this dataset, as most publications will likely compare wild type and genetically altered groups, so this data are not shown as a figure; counterbalancing is more likely to be relevant but is not picked up by the NLP tool.

Reporting of measure to reduce risk of bias	Number of publications
Reported random allocation of animals to experimental groups	1,553
Reported blinded outcome assessment	4,007
Reported criteria for animal exclusions	2,126
Included a conflicts of interest statement	10,404
Included an animal welfare statement	8,465
Did not report any measures to reduce the risk of bias	1,256
Unknown	1,502
<i>Publication has machine readable full text ' 5le size too big for software</i>	106
<i>Publication has full text ' full texts not suitable for software</i>	842
<i>Publication has DOI ' no full text retrieved</i>	272
<i>Publication has no DOI</i>	282

Table 25: Number of publications tagged as reporting measures to reduce the risk of bias. Data from risk of bias tool.

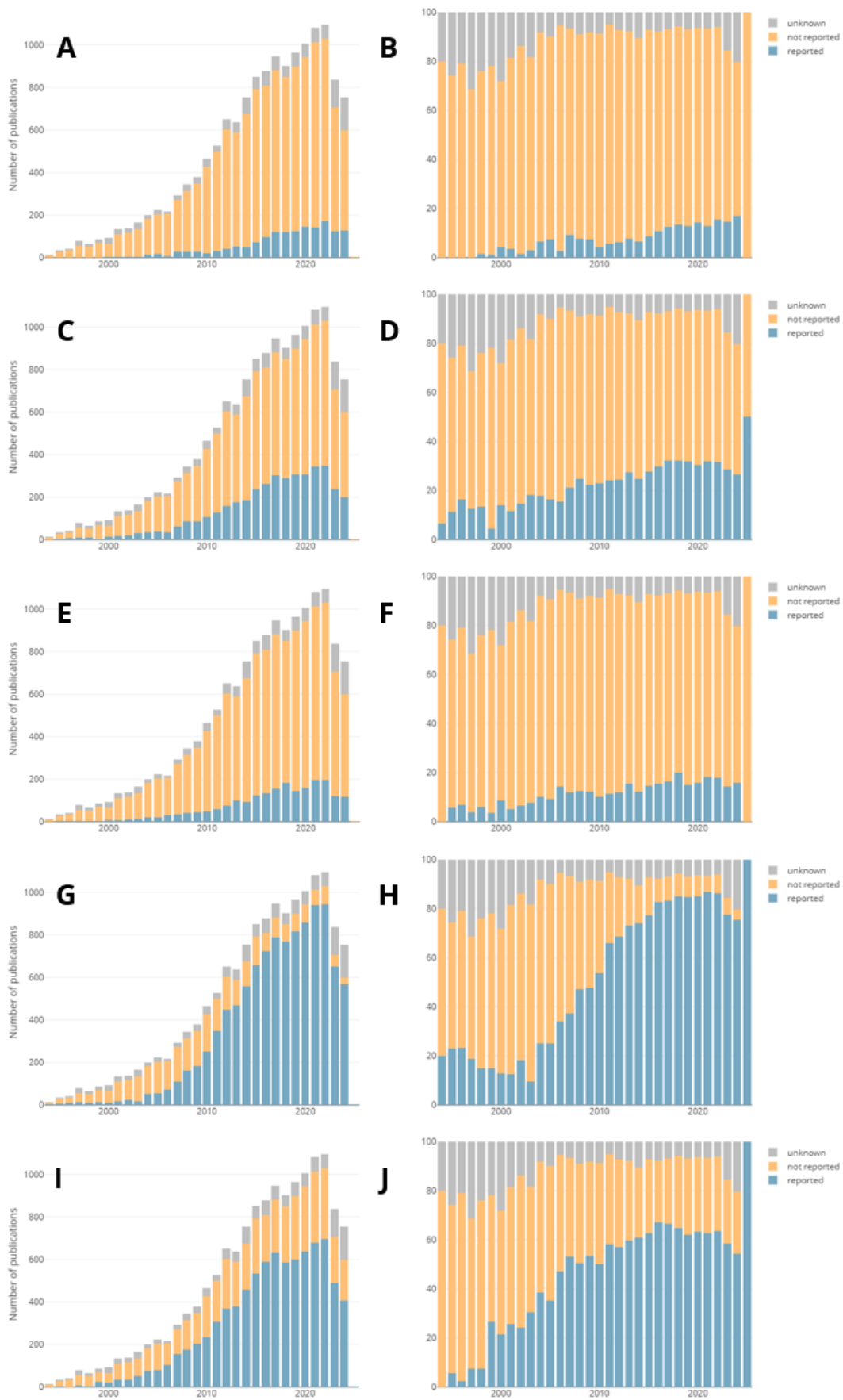


Figure 15: Publications tagged using risk of bias tool. Number and percent of publications tagged as reporting (a,b) random allocation of animals to groups, (c,d) blinded assessment of outcome, (e,f) reporting whether or not animals were excluded from analysis, (g,h) presence of a conflict of interests' statement (regardless of conflicts of interest), and (i,j) statement of compliance with animal welfare regulations, tagged by the natural language processing tool. An up-to-date interactive plot is available at:

<https://camarades.shinyapps.io/NDC-SOLES/>.

4.4.6 Web application

The systematic map can be accessed as a web application online at <https://camarades.shinyapps.io/NDC-SOLES/> (Figure 16). The web application contains information on how data is retrieved and includes interactive plots, similar to the ones shown above in this chapter, that are updated as new publications are included in dataset.

A key feature of the web application is the ability to search, filter, and download bibliographic information about publications. The app supports a basic search, where search keywords can be combined with AND or OR logic, and an advanced search, where search keywords can be combined with both AND and OR logic (Figure 17A-B). The application does not support exclusionary searches, e.g., using NOT logic. Search filters can be used on their own or in addition to the basic or advanced search (Figure 18). Bibliographic information for all publications, or publications retrieved via searching or filtering in the app, are displayed in a table, and can be downloaded in various formats (comma separated value (CSV), tab delimited text file (formatted for EndNote upload), SyRF upload, and Rayyan upload). Rayyan is a systematic review software similar to SyRF. Abstract text is not displayed or downloadable due to copyright restrictions on some abstracts.

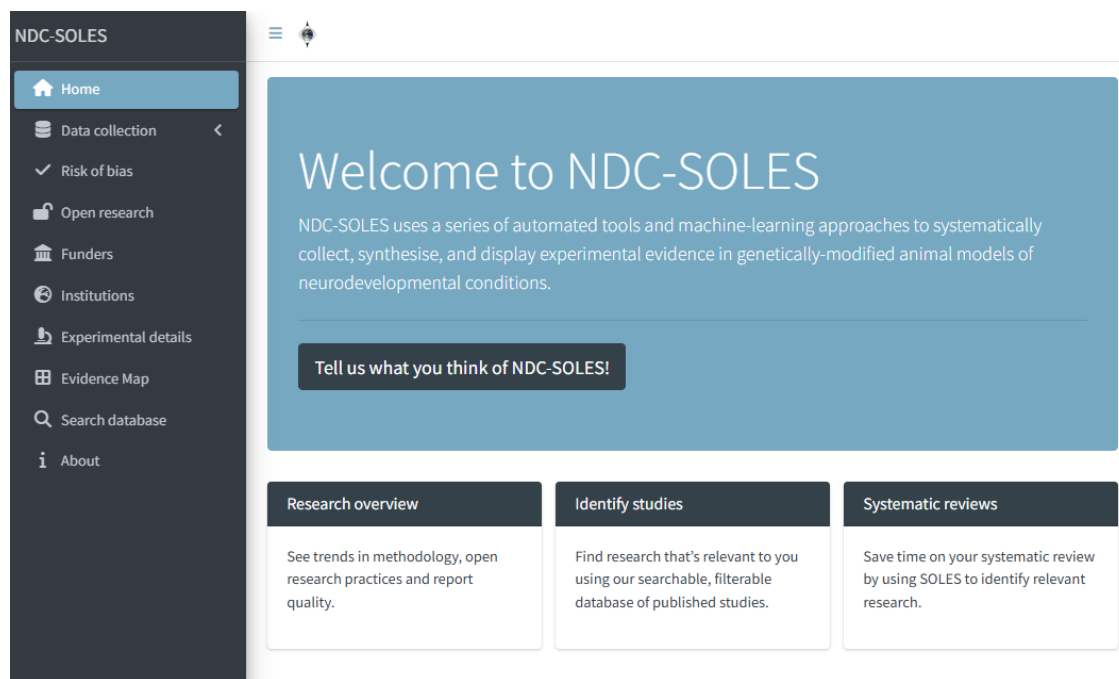


Figure 16: Screenshot of NDC-SOLES web application homepage. The web application can be accessed at <https://camarades.shinyapps.io/NDC-SOLES/>.

Figure 17: Screenshots of the (a) basic and (b) advanced search boxes. The web application can be accessed at <https://camarades.shinyapps.io/NDC-SOLES/>.

All 14867 citations loaded. Use the search box above or apply filters to identify relevant studies

Click to filter studies

Filter studies

Year Published

1994 1998 2002 2006 2010 2014 2018 2022 2025

Filter by model:

ADNP, ALDH5A1, ANK2, ANKRD11, AP2S1, ARID1B, ARX, ASH1L, ASXL3, AUTS2, E

Filter by species:

Armadillo, Cat, Chicken, Dog, Frog, Fruitfly, Guinea pig, Hamster, Llama, Mouse

Filter by sex:

Both sexes, Female, Male, Unknown Sex

Filter by outcome:

Active Avoidance, Barnes Maze or Reverse Barnes Maze, Chemotaxis, Cold Plate

Filter by funder:

"la Caixa" Foundation, Aard- en Levenswetenschappen, Nederlandse Organisatie

Filter by institution:

3M (United States), A.S. Watson (Netherlands), Aarhus University, Aarhus University

Filter by OpenAlex discipline:

Abdominal wall, Ablation, Abnormality, Abstinence, ABX test, Acceleration, Ace

Filter by language:

de, el, en, es, fr, it, ja, pl, ro, ru, Unknown

Show or hide retracted articles:

Not retracted, Retracted, Unknown

Apply filters

Figure 18: Screenshot of the available search filters. The web application can be accessed at <https://camarades.shinyapps.io/NDC-SOLES/>.

4.5 Discussion

Here, I succeeded in incorporating various software tools into a workflow to develop an automated systematic map of monogenic NDC animal research. Using the search strategy, trained machine learning classifier, and NER approach developed in the previous chapter, I was able to generate a dataset of relevant publications. As of 4 October 2024, 14,867 publications are included in the systematic map. Using scheduled scientific database API calls and code, I can keep this dataset up to date as new research is published. By making use of existing software, including ODDPub and a risk of bias reporting detection tool, and

existing data sources, including OpenAlex and ROR, I have been able to tag relevant publications with high-level metadata. Using R Shiny Apps, I have published a web application that allows the data to be interrogated, visualised, and searched.

4.5.1 Comparison with other approaches

The use of web applications is a novel approach to display evidence and other research information in an interactive way that may be more useful for decision-makers. The International Initiative for Impact Evaluation (3ie), produce web-based evidence gap maps on social sciences topics such as good security, agriculture, and climate change (Snilstveit et al., 2016). Although the 3ie documentation for producing these maps mentions the use of text mining approaches (Snilstveit et al., 2016), there doesn't appear to be updated documentation detailing how it may be used. Other maps in the climate sciences use machine-learning methods to handle and classify large numbers of potentially relevant records (Lück et al., 2025), but similarly do not use these methods to screen new literature on a regular basis. These systematic maps are static, snapshots of evidence at a point in time; but do use interactive visualisations such as bubble plots, heat maps, and bar charts to allow research users to interrogate the data of interest to them.

AD-SOLES, a systematic map of animal research in Alzheimer's disease, uses API searching and machine-learning to retrieve and process new literature, and displays results in an interactive web application (Hair et al., 2024). AD-SOLES uses the same underlying architecture and software code I have described here and in the previous chapter; the idea being that researchers from different domains can collaborate to build, maintain, and use a suite of interoperable software. However, despite building from the same starting point, there are also key differences between the work presented here and AD-SOLES. Some differences are due to advancements in the code available, such as integration of OpenAlex data. Others relate to differences in data visualisations to suit the needs of research users in that domain; for instance, AD-SOLES focuses on preclinical intervention-testing research, so therefore a sun-burst plot exploring the number of publications

testing different drug interventions, and a bubble plot exploring the number of publications testing specific drug interventions in different Alzheimer's models, are of particular relevance (Hair et al., 2024). Alternatively, as I was typically focusing on fundamental research, I found extracting drug names from texts was not particularly relevant to research users, as these drugs were typically used in normal experimental procedures such as anaesthetising an animal for surgery, and instead found it more relevant to display an interactive bubble plot exploring genetics, species, and experimental tests (<https://camarades.shinyapps.io/NDC-SOLES/>).

Overall, these are various approaches one might take to create a systematic map, but the processes used should be aligned with the project goal. While climate and social sciences research users may require highly curated datasets to make policy decisions, in the biomedical sciences, a systematic map may be useful as a discovery tool or a starting point for further analyses.

4.5.2 Strengths and limitations

A major strength of my approach is the use of code and software tools to automate the systematic map process and allow the continual identification and curation of relevant publications. Using automation approaches reduces the amount of manual resource required to keep the systematic map updated, which increases the feasibility and sustainability of the project. The broad focus of my systematic map means that publications investigating different monogenic NDC models can be identified in one resource, allowing connections to be made across different genetic conditions. Additionally, by sharing the data in an interactive and searchable web application, I hope to promote a culture of openness and that the resource can benefit research users' various needs relating to evidence synthesis.

A further strength of this work was my use of pre-made tools where they exist. There is a vibrant community of research software engineers developing evidence synthesis and open research related software tools in the R programming language, including ROpenSci (<https://ropensci.org/>) who developed many of the R

packages used here including roadoi (Jahn, 2017), openalexR (Aria et al., 2022), and rcrossref (Chamberlain et al., 2014). The data extraction tools used (ODDPub and the risk of bias reporting detection tool) have been validated in separate biomedical datasets (Riedel et al., 2020; Wang et al., 2022b). However, a limitation is that they may require additional validation. Similar to the limitations discussed in the previous chapter, the performance of data extraction tools is highly dependent on the text they are run on and how well different entities are reported. To acknowledge these limitations in my systematic map data, I have ensured that the app transparently describes where all data come from and links to publications validating each tools' performance.

I have also made use of existing research metadata where possible, including from the databases searched and additional data sources such as OpenAlex. OpenAlex is an online, open data source that curates and combines research metadata from sources including CrossRef and Retraction Watch as well as adding their own metadata such as language and discipline tags (Priem et al., 2022). OpenAlex is an invaluable resource, however, not infallible. 148 publications in the dataset were incorrectly labelled as closed / paywalled when in fact they were available open access. Funder information was only available for 36% of the dataset. According to the CrossRef documentation (OpenAlex retrieves funding information from CrossRef), this information should be deposited by journals in a particular format, and from inspection of the missing data it appears that often this information is not deposited. A small number of publications were missing a DOI, meaning full texts and metadata could not be retrieved at all using automated approaches, limiting the completeness of data extraction. Additionally, most databases do not provide easy linkage between preprints, conference proceedings, and peer-reviewed journal articles. OpenAlex takes publication type information from CrossRef, which combines the publication types "journal-article", "proceedings-article", and "posted-content" into one type labelled "article", making these publication types difficult to differentiate programmatically (Haupka et al., 2024).

I am not alone in considering the lack of complete and comprehensive metadata a major problem. The recent Barcelona Declaration on Open Research Information seeks to unlock research information that is currently held by propriety infrastructures to make open research information the default, work with services enabling open research information, and support infrastructure sustainability and the transition towards openness (Barcelona Declaration, 2024). The Declaration defines research information as *"information (sometimes referred to as metadata) relating to the conduct and communication of research. This includes, but is not limited to, (1) bibliographic metadata such as titles, abstracts, references, author data, affiliation data, and data on publication venues, (2) metadata on research software, research data, samples, and instruments, (3) information on funding and grants, and (4) information on organizations and research contributors. Research information is located in systems such as bibliographic databases, software archives, data repositories, and current research information systems."*, and open research information as *"research information that is free to access and free of restrictions on reuse"* (Barcelona Declaration, 2024).

Another key challenge was that some metadata is not static and needs to be constantly updated. An example of this type of metadata whether or not a publication is retracted. This information can be retrieved through OpenAlex and CrossRef, who recently acquired Retraction Watch, maintain the Retraction Watch database and allow free access to the full database. Rather than re-retrieving the data for every publication from OpenAlex, to check whether or not publications are retracted, it would make most sense to regularly retrieve and save the Retraction Watch database and check for new retractions.

Finally, I discovered that some metadata are not as reliable as others. Some OpenAlex metadata comes directly from publishers (e.g., funder information) and others are extracted using automated approaches, which, as I've discussed in my own work, have limitations. Language in OpenAlex is based on abstract text rather than full text, making it less useful to identify publications that, for instance, may have to be translated into English for text analysis processes to work. OpenAlex

uses a software called langdetect to guess publication abstract language (<https://docs.openalex.org/api-entities/works/work-object#language>). Similarly, disciplines (referred to in the OpenAlex documentation as topics) are assigned to publications using a large language model (LLM) (<https://docs.openalex.org/api-entities/topics/topic-object>). I found it difficult to use this data because I found that often disciplines were misassigned, for instance research on mouse ultrasonic vocalisations (the noises mice make to communicate with each other) were tagged as being related to music, likely due to the word vocalisation being misattributed by the large language model.

4.6 Conclusion

In conclusion, by incorporating various software tools into a workflow, I have developed a systematic online living evidence summary of published evidence of animal research used to study genetic neurodevelopmental conditions. This resource allows the identification and synthesis of relevant published research articles and is available online via an interactive and searchable web application (<https://camarades.shinyapps.io/NDC-SOLES/>). My systematic map incorporates pre-existing and custom-built software to retrieve, extract, and classify data, allowing the curation and visualisation of high-level data from publications. Further availability of open metadata and complete reporting of extractible features could increase the utility of such interactive datasets. Where automated approaches are used, especially in the extraction of information, it is important to make clear exactly where the data come from and how they were extracted.

Chapter 5: Systematic review of heterogeneity in *Syngap1* rodent models

5.1 Context

One of the key use cases of a systematic map is the ability to use it as a comprehensive resource for more detailed systematic reviews. I aimed to evaluate my systematic map of monogenic NDC animal research as a comprehensive resource through conducting a systematic review that was relevant to the wider research community.

I chose the topic for this systematic review by gathering opinions from laboratory researchers (PhD students, postdoctoral researchers, and principal investigators, total N=8) from the Simons Initiative for the Developing Brain at the University of Edinburgh, where I conducted my PhD. Using the information gathered, I consulted with my PhD supervisors (Professor Peter Kind and Professor Emily Sena) to determine the review topic.

As systematic reviews require all publications to be screened for relevance and data extracted by at least two independent reviewers, I aimed to conduct the systematic review collaboratively with other PhD students. Two other PhD students from SIDB (Chloe Henley and Lucy Pritchard) and one PhD student from CAMARADES (Francesca Tinsdeall) contributed to screening and annotation as second or reconciling reviewer. The data presented in this chapter have been dual screened and reconciled, however at this stage data is only single extracted by myself.

5.2 Introduction

SYNGAP1-related disorders, typically caused by *de novo* loss-of-function variations to the *SYNGAP1* (synaptic Ras GTPase-activating protein 1) gene, is one of the most common causes of non-syndromic intellectual disability (Satterstrom et al., 2020). Between 80-85% of people with SYNGAP1-related disorders have epilepsy and around half of those with epilepsy may respond to anti-epileptic treatments (Mignot et al., 2016). Co-occurring conditions include difficulties with eating and drinking (68%), sleep disturbances (62%), autism (53%), and ataxia or unsteady gait (51%) (Vlaskamp et al., 2019).

Researchers can investigate the biological mechanisms underpinning gene function by using genetically modified animals. Animal with alterations to the *Syngap1* gene modelling the human condition have been developed in several species, including the mouse and rat, and can be used to test anti-epileptic drugs (G. F. Marshall et al., 2021) to assess efficacy and safety before progressing to clinical trials. While genetic models including *Syngap1* models, can generally show high construct validity and may be used to develop personalised therapies (G. F. Marshall et al., 2021), there are still challenges to modelling epilepsy in animals. Seizures can be difficult to measure in animals, and frequency of seizure can vary (G. F. Marshall et al., 2021). Additionally, seizures can often be fatal; for instance, the *Syngap1*^{-/-} mice typically die within the first week of life (Kim et al., 2003). Differences in background strain have been shown to impact seizure onset in some genetic mouse models; this can impact the generalisability, construct, and face validity of models (G. F. Marshall et al., 2021). Conversely understanding these sources of heterogeneity can also give insights into epilepsy modifier genes (G. F. Marshall et al., 2021).

Systematic review is a structured research method to identify and evaluate all available evidence on a pre-defined research question; they can be followed by a meta-analysis to statistically combine evidence from published research (Soliman et al., 2020). In animal research, systematic reviews and meta-analyses can be

used to summarise knowledge in a particular area, identify gaps in the evidence, inform experimental design and identify sources of heterogeneity, inform measures to support the 3Rs (Replacement, Reduction and Refinement of animals in research), and inform decisions to justify clinical trials (Soliman et al., 2020).

5.2.1 Aims

I aim to conduct a systematic review and meta-analysis to answer the question: What factors account for heterogeneity in *Syngap* rodent seizure phenotype?

Secondarily, I aim to use this systematic review as a case study to evaluate the completeness of my systematic map as a comprehensive resource and starting point for systematic reviews, by comparing traditional search approaches with searching for relevant publications using my systematic map.

5.3 Methods

5.3.1 Protocol registration

A protocol detailing the rationale, aim, and methods planned for this systematic review was preregistered and is available on PROSPERO (<https://www.crd.york.ac.uk/PROSPERO/view/CRD42024596586>). Deviations from the protocol and this chapter are:

- Clarification that we intended to include any *Syngap* altered animals
- Publications without a wildtype control are included
- I used EndNote 20 to retrieve full text PDF files rather than R code
- Data has only been extracted by one reviewer for this thesis (and will be dual extracted and reconciled for future publication)
- The modifications to the SyRCLE risk of bias checklist are slightly different to account for (1) random allocation of animals to groups being relevant to some publications and not others and (2) publications not administering an intervention.

- I added checks of Retraction Watch and PubPeer to identify retracted or problematic publications.
- This chapter presents one analysis of data from publications identified from either or both of the traditional search or systematic map search rather than presenting two separate analyses. This is because meta-analysis was not possible and all the publications that could be quantitatively summarised were identified via both traditional and systematic map searches.

5.3.2 Inclusion and exclusion criteria

I used the following criteria to determine eligibility of publications for inclusion in my systematic review.

- Study design – I included experiments using animals to model human health conditions, published in peer-reviewed journal articles or as preprints. I excluded articles where no new data are generated, e.g., review articles, opinion pieces, letters to the editor; conference abstracts are also excluded as they typically do not report enough detail to be included in systematic reviews.
- Animal models – I included publications using rodents (mice or rats) with an alteration in the *Syngap* gene. I included publication where controls were wildtype littermates or age matched controls, or where no control animals were used. I excluded publications that did not use *Syngap* rodents.
- Outcome – I included publications that assessed seizures in *Syngap* altered rodents, such as seizure latency, spike frequency and amplitude, number of animals that developed a seizure, and seizure related animal mortality. I excluded publications that did not report seizure data.

I did not place any restrictions on publication date or language. Where the full text of a relevant article does not have an English version, I planned to translate the article into English using Google Translate. However, all full texts were in English.

5.3.3 Bibliographic search

I searched four electronic sources via their web interfaces: (1) PubMed (including MEDLINE); (2) Scopus; (3) Embase, and (4) Web of Science (Core Collection, BIOSIS Citation Index, KCI-Korean Journal Database, Preprint Citation Index, and SciELO Citation Index).

Search strategies were built from terms including (1) *Syngap1* or *Syngap*, and; (2) animals, rodents, rats or mice. Full search terms are given in Appendix 5A.

Due to the number of search results from this search strategy being manageable, I did not add further terms to limit the search, such as terms related to seizures.

5.3.4 Removal of duplicate publications

In Rstudio, I merged all search results from all four data sources and removed duplicate copies of search results using the automated and manual features in the ASySD tool (Hair, Bahor, et al., 2023). ASySD has previously been shown to identify more duplicates than EndNote's deduplication features (Hair, Bahor, et al., 2023).

5.3.5 Full text document retrieval

I used the find full text function in EndNote 20 to retrieve full text PDFs of all unique search results, and handsearching where EndNote 20 could not retrieve a full text (either due to no DOI being in the search result or publishers blocking automated full text retrieval. I planned to use interlibrary loans to retrieve full texts (only were included after screening, see below) that were not open access or I could not access via University of Edinburgh institutional subscriptions, however this was not necessary.

5.3.6 Screening

I conducted one round of full text screening to avoid erroneous exclusion of studies due to limited title and abstract reporting, as I have previously acknowledged is an issue in reviews of *in vitro* biomedical research (Wilson, Cruz, et al., 2023). Where full texts could not be accessed open access or via institutional

subscriptions, screening decisions were made on title and abstract. Full texts for these publications would only be sought through interlibrary loan if they met screening criteria 1-5 (below).

Publication metadata (title, abstract, authors, publication year, and full text PDF) were uploaded to SyRF (RRID: SCR_018907; (Bahor et al., 2021)) for screening and annotation. SyRF displays publications to reviewers in a random order, and reviewers are unaware of other reviewers' decisions.

Two reviewers independently screened each publication against the eligibility criteria and, where necessary, recorded reasons for exclusion. Each publication was screened by myself once (first screener) and by either Chloe Henley or Lucy Pritchard (second screener). Disagreements between first and second screener were reconciled by a third independent reviewer (Francesca Tinsdeall). The reconciler was unaware of which reviewer made which screening decision, but was aware they were reconciling a disagreement.

The following screening questions were used. Publications were only included if they answered "yes" to all six questions. Screening questions 2-6 were conditionally displayed if the reviewer answered "yes" to the previous question:

1. Related to *Syngap* (yes / no)
2. Related to *Syngap* altered animals (yes / no)
3. Related to *Syngap* altered rodents (yes / no)
4. Presenting original data from *Syngap* altered rodents (yes / no)
5. Assessing seizure phenotype data in *Syngap* altered rodents (yes / no)
6. Presenting seizure phenotype data from *Syngap* altered rodents (yes / no)

The screening guide is available in Appendix 5B.

5.3.7 Checking for retractions

I checked relevant publications against the Retraction Watch database and PubPeer website to identify retractions or comments regarding potential

problems. This was not a pre-planned analysis, and I did not find any retractions or PubPeer comments.

5.3.8 Data extraction

I single-extracted data from all included publications via the SyRF platform.

Publications were presented in a random order. I extracted the following data, if reported:

Animal details:

- Animal age (in days; convert if given in weeks / months and assume one month is 28 days)
- Animal weight (grams)
- Animal species (rats or mice)
- Animal sex (male, female, or unknown)
- Animal source (name of laboratory or company that supplied animals)
- Background strain (e.g., C57BL/6)
- Details of specific alternation to *Syngap* gene

Husbandry details:

- Number of cage mates (number)
- Presence of environmental enrichment (yes, no, or unknown)
- Light / dark cycle (e.g., 12 hr / 12 hr)

Outcome / seizure phenotype details:

- For each outcome:
 - Number of animals per group (number of animals)
 - Stimulation to induce seizure (spontaneous / no stimulation, acoustic stimulation, chemical)
 - Type of seizure (e.g., absence)
 - Qualitative description of motor behaviours (e.g., head bobbing)
- Any continuous outcomes, if conducted:

- Seizure latency, e.g., time between stimulation and seizure (seconds; mean and standard error of the mean (SEM) or standard deviation (SD))
- Electrophysiology analysis – spike frequency (spike/min; mean and SEM or SD) and amplitude (μ V mean and SEM or SD)
- Any dichotomous outcomes, if conducted:
 - Number of animals that developed seizures (number)
 - Animal mortality (number of deaths)

I extracted reporting of measures to reduce using the SYRCLE risk of bias checklist (Hooijmans et al., 2014), with slight modifications. Questions on random allocation of animals to groups were not applicable if the only groups used in relevant experiments differed only by genotype and hence could not be randomly allocated. Additionally, questions related to blinding, where investigators were unaware of intervention given to animals, were modified to mention investigators being unaware of animal group instead.

The data extraction guide, which clarifies exactly how risk of bias was assessed, is available in Appendix 5C.

5.3.9 Quantitative data synthesis

All analyses were conducted in R statistical software. I aimed to conduct a meta-analysis, using approaches described by Yang et al. (2023), if outcomes were reported in at least ten experiments. Unfortunately, due to lack of data this was not possible. Instead, I conducted synthesis without meta-analysis (SWiM) (M. Campbell et al., 2020) by visually examining forest plots for direction and magnitude of effects and the degree of overlap between 95% confidence intervals across individual publications. Additionally, due to the small number of included publications, I could not investigate sensitivity analyses, small study effects, or publication bias. Planned methods for these analyses are included in my preregistration on PROSPERO.

5.3.9.1 Effect sizes

Where there were enough experiments to analyse quantitative data (seizure latency only), I calculated effect sizes as the mean difference. I was able to use MD, rather than e.g., standardised or normalised mean difference, as outcomes were measured on the same scale (seizure latency in mice) using the same units (seconds from chemical induction to first clonus). Effect sizes are presented in a forest plot alongside their 95% confidence intervals.

5.3.9.2 Comparisons

The outcomes of interest were the frequency or magnitude of seizure related outcomes in *Syngap* rodents compared with wildtype rodents.

5.3.9.3 Heterogeneity and sensitivity analyses

I used the restricted maximum likelihood (REML) to estimate the between-study variance (τ^2) and the Hartung-Knapp method to adjust the confidence intervals of the effect sizes.

I planned to quantify heterogeneity using the τ^2 and the 95% prediction intervals of the effect sizes. However, due to lack of data, I could not perform subgroup analyses to identify sources of heterogeneity. In my preregistration, I planned to investigate other potential sources of heterogeneity with meta-regression, including conducting subgroup analyses to investigate whether the following variables account for heterogeneity for each outcome measure:

- Animal age
- Animal sex
- Animal background strain
- The specific *Syngap* modification
- Presence of environmental enrichment
- Blinded outcome assessment

5.3.10 Data presentation

I have presented a table of all included publications, along with the characteristics and risk of bias reporting extracted from each publication, and a qualitative

summary of those characteristics. Where I have been able to quantitatively summarise data, I have presented this data in a forest plot.

I planned to use a version of GRADE (Grading of recommendations assessment, development, and evaluation) adapted for animal research (Hooijmans et al., 2018) to evaluate the confidence in the summary of evidence. This evaluation considers the summary of the association (such as the magnitude and direction of the effects, imprecision and heterogeneity), potential threats to internal and external validity, potential biases in the systematic review process (also known as “meta-bias”) and reporting biases. However, this was not possible as I was unable to assess many of the GRADE criteria such as publication bias.

5.3.11 Validation of the systematic map

I used this systematic review as a test case to validate the useability of my systematic map for aiding systematic reviews. Systematic maps aim to curate publications from broad fields of research where there are high publications rates, and can use machine learning and text mining approaches to classify research and extract high level features which as study characteristics (Hair, Wilson, et al., 2023). NDC-SOLES is a purpose-built database of research using animals to model genetic neurodevelopmental conditions, as it updated on a regular basis to include newly published research. The data can currently be accessed via the web (<https://camarades.shinyapps.io/NDC-SOLES/>). Alongside my traditional systematic search described above, I searched NDC-SOLES via the web application using the following filters (SYNGAP1 (model), and mouse and rat (species)). I compared the number of relevant publications found or missed from the systematic map versus traditional searches. If the systematic map identified additional relevant publications that my traditional search missed, I planned to include these in my review.

5.4 Results

5.4.1 Systematic search results

PubMed, Embase, Scopus, and Web of Science searches were conducted on 30th October 2024. In total, 840 publications were retrieved (Figure 19). ASySD detected 516 publications were duplicate copies (512 via automated duplicate removal and 4 via manual duplicate removal) and these were removed leaving 324 unique publications. A Full text PDF document was retrieved for 285 of the unique publications.

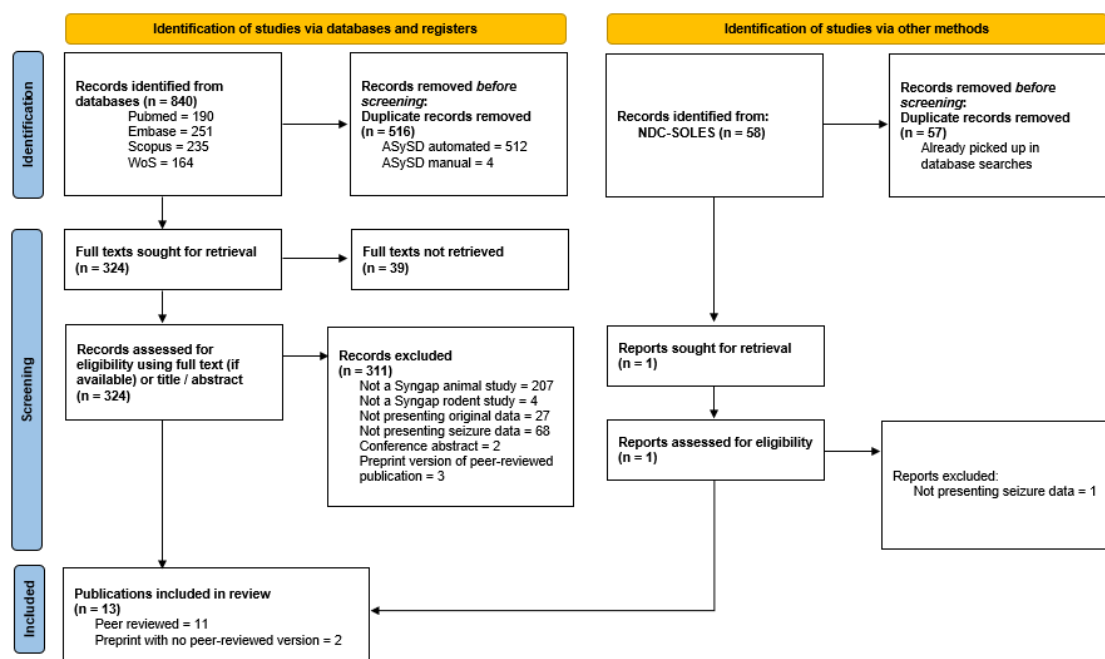


Figure 19: Adapted PRISMA ⁷ ow diagram for systematic review.

Diagram templates are shared under a Creative Commons By Attribution (CC-BY) license (<https://creativecommons.org/licenses/by/4.0/>) and available at <https://www.prisma-statement.org/prisma-2020-7-ow-diagram>.

All 324 unique publications were screened in SyRF. After reconciliation of screening decisions, 16 publications were determined to meet the inclusion criteria. Reviewers disagreed on screening decisions for 53 publications out of 324. Of the 16 publications meeting the inclusion criteria, three were preprints relating to a later published peer-reviewed journal article, so were excluded. 13 relevant publications were included in the review (11 peer-reviewed journal articles and

two preprints without a peer-reviewed journal article). On 8th January 2025, I checked all 13 included publications against the Retraction Watch database and PubPeer website and found no retractions or PubPeer comments.

Searching of my systematic map identified 58 publications. One publication identified via systematic map search was not picked up by traditional searches; however, I single-screened this publication and determined it was not relevant. 10 of the included publications were identified by the systematic map search. The three that were not identified were not included in the systematic map database and not picked up in systematic searches at all. For two of the publications (BioRxiv preprints, retrieved from Web of Science Preprint Index), this was because systematic map searches do not pick up all BioRxiv preprints; only National Institute of Health (NIH)-funded BioRxiv preprints seem to be indexed in PubMed and preprints are not included in the Web of Science Core Collection (NIH, 2024). The third publication was published in 2002. It's title and abstract did not mention autism, intellectual disability, or epilepsy, meaning it was not picked up in systematic map searches.

5.4.2 Narrative summary of included publications

The 13 included publications were published between 2002 and 2024 (mean publication year = 2019; median publication year = 2020) in a range of neuroscience and psychiatry related journals. 11 were published in peer-reviewed journals and two were preprinted on BioRxiv. Included publications are narratively summarised in Table 26 and in the sections below.

Study	Species	Background Strain(s)	Control Type	<i>Syngap1</i> Alteration(s)	Age / Weight	Sex	Husbandry	Seizure Recording	Outcome Measure(s)
(Komiyama et al., 2002); Journal of Neuroscience	Mouse	C57BL/6	Littermates	<i>Syngap</i> +/- (Komiyama et al., 2002)	Not reported	M + F	Not reported	Behavioural observation	Number of animals that developed spontaneous seizures
(Komiyama et al., 2002); Cell	Mouse	C57BL/6	Littermates	<i>Syngap</i> +/- (Kim et al., 2003)	17 days	M + F	Not reported	Flurothyl induced	Seizure latency
(Ozkan et al., 2014); Neuron	Mouse	C57BL/6	Littermates	<i>Syngap</i> +/- (Kim et al., 2003)	Not reported	Not reported	Not reported	Flurothyl induced	Seizure latency
(Creson et al., 2019); Elife	Mouse	C57BL/6	Littermates	<i>Syngap</i> +/- (Kim et al., 2003)	Not reported	M + F	4 per cage; 12:12 hr light dark cycle	Flurothyl induced	Seizure latency
(Kilinc et al., 2020); BioRxiv	Mouse	C57BL/6N & C57BL/6J	Littermates	<i>Syngap</i> PBM/PBM (Kilinc et al., 2020)	Not reported	M + F	4 per cage; Environmental enrichment	Flurothyl induced	Seizure latency

(Sullivan, Kipnis, et al., 2020); BioRxiv	Mouse	C57BL/6J + 129	No control	B6;129- <i>Syngap1</i> tm1Rlh/J (Jackson Laboratory)	7 days	M + F	Not reported	2 hr EEG	Spontaneous seizures per hour; Interictal spikes per hour
(Sullivan, Ammanuel, et al., 2020); Biological Psychiatry	Mouse	Not reported	No control	<i>Syngap</i> +/- (Sullivan, Ammanuel, et al., 2020)	46-120 days	M	1 per cage; 12:12 hr light dark cycle	24 hr EEG	Spontaneous seizure duration
(Buller-Peralta et al., 2022); Brain Communications	Rat	Long Evans- SG em2/PWC	Littermates	<i>Syngap</i> +/-Δ-GAP (Patrick Wild Centre)	112-119 days	M	Up to 4 per cage then single-housed post-surgery; 12:12 hr light dark cycle	6 hr EEG	Time in spike wave discharge

(Carreño-Muñoz et al., 2022); Brain	Mouse	C57Bl/6	Littermates	<i>Syngap</i> ^{+/-} (Komiyama et al., 2002)	80-120 days	M + F	2-5 per cage then single- housed post- surgery; 12:12 hr light dark cycle	2 hr EEG	Interictal spikes per minute
(Kilinc et al., 2022); Elife	Mouse	C57BL/6J & C57BL6/N	Littermates	<i>Syngap</i> PBM/PBM (Kilinc et al., 2020); <i>Syngap</i> ^{+/Td} (Spicer et al., 2017); <i>Syngap</i> B [*] /B [*] ((Kilinc et al., 2022), heterozygous also analysed)	Not reported	M + F	4 per cage; Environmental enrichment	Flurothyl induced	Seizure latency

(Verma et al., 2022); Experimental Brain Research	Mouse	C57BL/6J + 129	Littermates	B6;129- <i>Syngap1</i> tm1Rlh/J (Jackson Laboratory)	80 days	M	12:12 hr light dark cycle	Flurothyl induced seizures	Seizure latency
(Fenton et al., 2024); Translational Psychiatry	Mouse	C57BL/6J + 129	Littermates	B6;129- <i>Syngap1</i> tm1Rlh/J (Jackson Laboratory)	56-112 days; >20g	M	12:12 hr light dark cycle; Environmental enrichment	EEG (length of time unknown)	Number of spike trains
(Katsanevaki et al., 2024); Cell Reports	Rat	Long Evans- SGem2/PWC & Long Evans- SGem1/PWC	Littermates	<i>Syngap</i> +/ Δ -GAP (Patrick Wild Centre); <i>Syngap</i> +/- (Patrick Wild Centre and (Mastro et al., 2020))	84 days or older	M	12:12 hr light dark cycle	6 hr EEG	Whether animals had spike wave discharge; Number of spike wave discharges

Table 26: Summary characteristics extracted from included publications. M = Male; F = female.

5.4.2.1 Animal and genetics details

11 publications used mice to model *Syngap1* alterations while two publications used rats. The two publications using rats were recently published in 2022 and 2024. Eight publications reported animal age and five publications did not. Two publications used mice aged 7 and 17 days respectively, one publication used mice aged 80, and five gave age as a range from 46 to 120 days. Only one publication (Fenton et al., 2024) reported animal weight. Seven publications used male and female rodents, five used only males, and one did not report animal sex. The two publications using rats only used males.

Most publications only investigated one *Syngap* alteration, while Kilinc et al. (2022) investigated four alterations in mice and Katsanevaki et al. (2024) investigated two alterations in rats. Eleven publications reported using littermates as controls, and two publications (Sullivan, Ammanuel, et al., 2020; Sullivan, Kipnis, et al., 2020) did not report data from wildtype control animals. Five publications reported using the C57BL/6 background strain but did not specify C57BL/6J or C57BL/6N.

Encoded by the *Syngap1* gene, SYNGAP acts as a scaffolding protein, managing and organising molecules in synapses. It does so by binding to PDZ domains of PSD95, another scaffolding protein, and may regulate phase transition of the postsynaptic density (PSD); such phase transition is a key mechanism underlying synaptic plasticity (Zeng et al., 2016). While there are multiple isoforms of the SYNGAP protein, all isoforms contain a central calcium/lipid-binding (C2) and GTPase-activation protein (GAP) domain (Katsanevaki et al., 2024). These domains have enzymatic functions, including negatively regulating several small GTPases such as Ras, which activates the ERK pathway, and Rap, which negatively regulates the p38 pathway (Krapivinsky et al., 2004; Pena et al., 2008; Walkup et al., 2015). The ERK and p38 pathways play complex roles in many crucial biological processes, including cell proliferation, cell stress, and mRNA translation (Roux & Blenis, 2004).

Most publications used rodents with alterations between exons 7-12. In most models, this results in a full knock out of one copy of the gene. Two publications (Carreño-Muñoz et al., 2022; Komiyama et al., 2002) used *Syngap*^{+/-} mice which carry a deletion of *Syngap* exons 8-10. Three publications (Clement et al., 2012; Creson et al., 2019; Ozkan et al., 2014) used *Syngap*^{+/-} mice constructed according to Kim et al., 2003. These mice had an alteration that amplified a specific fragment between the C2 and GAP domains. Three publications (Fenton et al., 2024; Sullivan, Kipnis, et al., 2020; Verma et al., 2022) used *Syngap* KO mice (B6;129-*Syngap*¹tm1Rlh/J mice from the Jackson Laboratory) which have exons 7 and 8 replaced with a neomycin resistance cassette. Sullivan, Ammanuel, et al. (2020) does not report the source of their mice but states they used *Syngap*^{+/-} mice with deletion of exons 7 and 8.

Alternatively, both publications using rats (Buller-Peralta et al., 2022; Katsanevaki et al., 2024) used *Syngap*^{+/-}Δ-GAP rats (Long Evans-SG em2/PWC) which carry a selective deletion of *Syngap* exons 8–12, targeting only the C2 and GAP domains. Katsanevaki et al. (2024) also used *Syngap*^{+/-} rats (Long Evans-SGem1/PWC), sourced from Mastro et al. (2020) which carry a frameshift mutation in exon 8. The Both (Kilinc et al., 2020, 2022) used *Syngap*^{PBM/PBM} mice expressing endogenous *Syngap*-α1 isoforms carrying point mutations in the PDZ-binding motif (PBM), which impact cellular scaffolding. Additionally, Kilinc et al. (2022) also used *Syngap*^{+/-}Td mice (heterogenous Td/Td mice sourced from Spicer et al. (2017)) which carry alterations upregulating beta isoforms and disrupting alpha isoforms, and *Syngap*^{B*/B*} mice (bred in house) which carry alterations upregulating alpha isoforms and disrupting beta isoforms. In the same publication, *Syngap*^{B*/B*} mice were also bred with *Syngap* heterozygous mice and used in experiments. Alpha and beta isoforms are generated from alternative splicing of *Syngap* and may function differently in the body (Kilinc et al., 2022).

5.4.2.2 Animal husbandry details

Seven publications reported using a normal 12 hr:12 hr light dark cycle, and six publications did not report the light cycle used. Only three publications, all using mice, reported environmental enrichment. Of these, two publications reported using “pyramidal-shaped huts with two square opening on opposing sides of the hut” in cages for behavioural experiments for environmental enrichment and to help transfer rodents between home and behavioural testing cages (Kilinc et al., 2020, 2022) and one publication reported that each cage contained “a Nestlet square, shredded brown paper, and a cardboard tube” as well as standard bedding (Fenton et al., 2024). Six publications reported the number of animals housed per cage. Of these, three reported housing four rodents per cage (Creson et al., 2019; Kilinc et al., 2020, 2022), one reported single housing rodents (Sullivan, Ammanuel, et al., 2020), and two reported rodents were housed up to four and two to five per cage before surgeries and then single housed post-surgery (Buller-Peralta et al., 2022; Carreño-Muñoz et al., 2022).

5.4.2.3 Recording of seizures

One publication (Komiyama et al., 2002) reported that no seizures were detected in *Syngap1*^{+/-} mice but did not report the methods used to detect seizures. Carreño-Muñoz et al. (2022), using mice with the same genetic alteration, presented data on number of interictal spikes per minute detected during a 2hr EEG recording. Interictal spikes are electrical signals used as biomarkers for seizures (Carreño-Muñoz et al., 2022).

All three publications using *Syngap*^{+/-} constructed according to Kim et al. (2003) – (Clement et al., 2012; Creson et al., 2019; Ozkan et al., 2014) – used flurothyl to induce seizures and recorded seizure latency. (Kilinc et al., 2020, 2022) used flurothyl to induce seizures and recorded seizure latency from *Syngap*^{PBM/PBM, +/Td, B^{*}/B^{*}, +/B^{*}, and B^{*}/-} mice. Flurothyl is a fast-acting central nervous system stimulant widely used to induce seizures in animal models.

One of the publications using *Syngap* KO mice (Verma et al., 2022) used flurothyl to induce seizures and recorded seizure latency, while the other two measured number of spike trains (EEG spikes occurring in quick succession; the length of EEG recording was not reported) (Fenton et al., 2024) and spontaneous seizures per hour and interictal spikes per hour over a 2 hr EEG (Sullivan, Kipnis, et al., 2020). Sullivan, Ammanuel, et al. (2020) recorded spontaneous seizure duration over a 24 hr EEG. Both publications using rats measured spike wave discharges (SWDs); Katsanevaki et al. (2024) recorded whether animals had SWD and the number of SWDs during a 6 hr EEG, while Buller-Peralta et al. (2022) recorded time in SWD over a 6 hr EEG. SWDs tended to occur alongside behaviours associated with absence seizures such as head bobbing and freezing (Katsanevaki et al., 2024).

5.4.3 Quantitative summary of flurothyl-induced seizure data

Of the six publications that investigated flurothyl-induced seizure latency, five were quantitatively synthesised. One publication could not be included in quantitative synthesis as it measured outcomes using interquartile range and did not provide raw data to convert this into the mean and standard deviation. All five other publications reported the mean and standard error of the mean, which was converted to standard deviation. One publication investigated seizures in multiple *Syngap1* animal models and therefore in total, nine comparisons were included in the quantitative synthesis (Figure 20). All comparisons show a lower seizure latency in *Syngap* altered rodents compared to controls (mean difference = 41.96; 95% confidence intervals = -56.13 to 27.78; $p = 0.0015$), meaning the *Syngap* alteration causes flurothyl-induced seizures to occur faster.

Both I^2 and τ^2 are high (68.2% and 234.6, respectively), indicating heterogeneity in the data. Interestingly, we can see two clusters appear in the forest plot. The earlier publications (Clement et al., 2012; Creson et al., 2019; Ozkan et al., 2014) show a larger effect than the later publications (Kilinc et al., 2020, 2022). The earlier studies all used mice constructed according to (Kim et al., 2003) while the

later studies used mice constructed with other specific *Syngap* alterations (described in the section above). However, it is not possible to say if this is the source of heterogeneity we see in the forest plot. The later comparisons only represent two publications, both with the same first author, and some other aspect of experimental conduct may be influencing heterogeneity. Ideally, with more comparisons from more studies, I could perform subgroup analyses to better understand potential sources of heterogeneity.

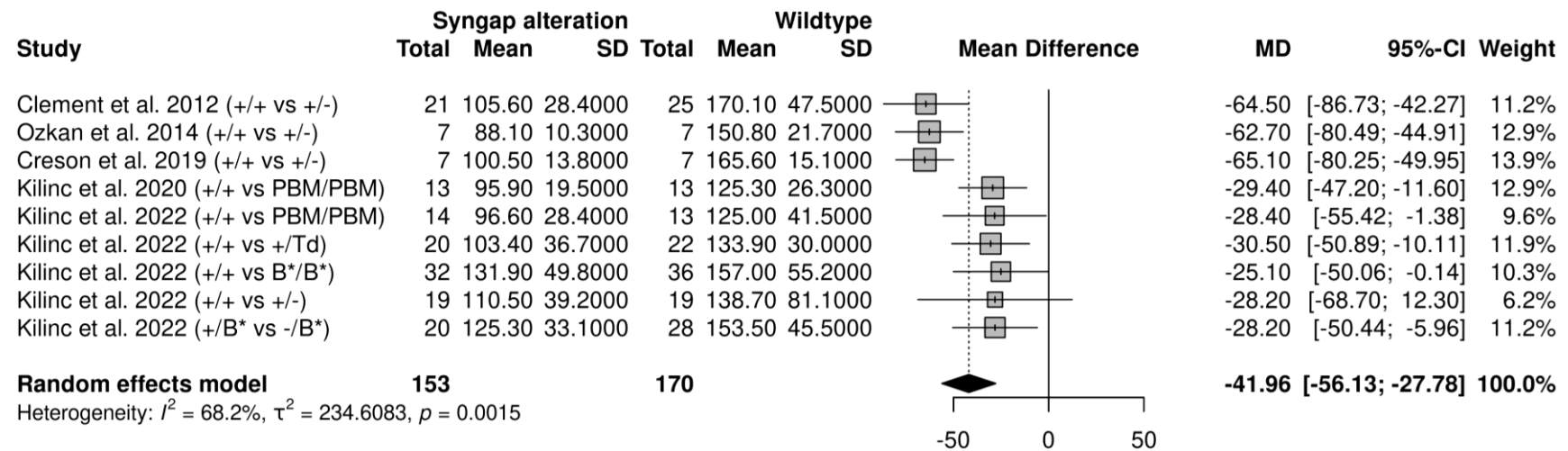


Figure 20: Forest plot of quantitative data

5.4.4 Reporting of measures to reduce risk of bias

Overall, publications scored either unclear or low risk of bias for all items except two publications on domain 9 (publication is free from selective reporting) where risk of bias was high (Figure 21). Domains 1 and 3 (relating to group allocation) were only relevant for four publications; the other nine publications either only compared groups distinguished by genotype only or did not use a wildtype control group. In the four publications where domains 1 and 3 were relevant, three of four reported randomisation was used to allocate animals to groups but did not state the method of randomisation used, and no publications reported allocation concealment. Only the four publications where domains 1 and 3 were relevant and one other publication reported groups were similar at baseline (domain 2). Only two publications reported animals were randomly housed within cages and no publications reported that cages were randomly located around animal house rooms (domain 4). 10 publications reported that caregivers or investigators were blinded to animal group during experiments (domain 5) and ten reported investigators were blinded to animal group during outcome assessment (domain 7). However, no publications reported methods for blinding and it was difficult to ascertain exactly who was blinded and at which particular stages. No publications reported that animals were randomly selected for outcome assessment (domain 6). Seven publications addressed incomplete data reporting (domain 8), eight publications appeared free from selective outcome reporting (domain 9), and 10 publications appeared free from other risks of bias. Two publications were judged to be at high risk of bias for domain 9 (selective outcome reporting) as they reported that some seizure data was not shown. Of the three publications judged to have unclear risk of bias for domain 10 (other risks of bias), one did not include a conflicts of interest statement, one did not include a funding statement, and one did not include both a conflicts of interest and a funding statement.

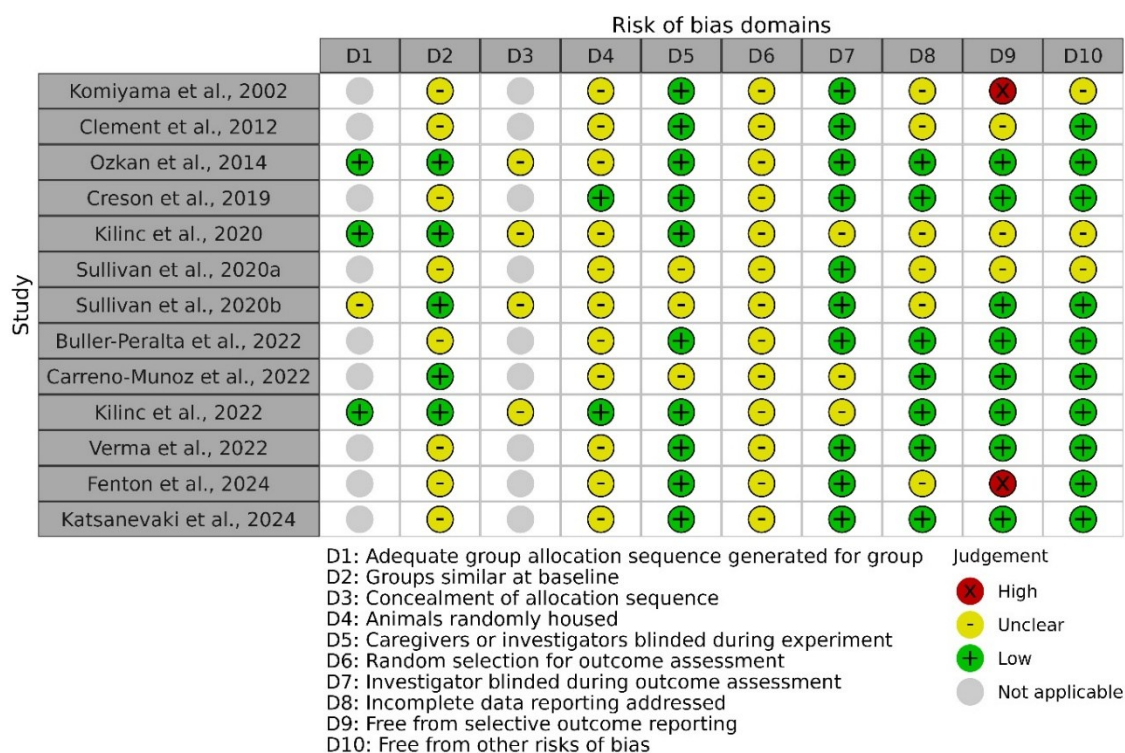


Figure 21: *TralE* light plot of the adapted SYRCLE risk of bias tool. Sullivan et al. 2020a refers to Sullivan, Kipnis, et al., 2020 and Sullivan et al., 2020b refers to Sullivan, Ammanuel, et al., 2020.

5.5 Discussion

Here I set out to conduct a systematic review of relevance to the research community, investigating potential sources of heterogeneity in *Syngap* rodent seizure phenotypes. I identified thirteen publications reporting data on seizures in *Syngap* rodents. Of these, eleven used mice and two used rats. I was able to provide a narrative summary of the *Syngap* models used and how seizures were measured in animals, typically via EEG recording of flurothyl-induced seizures. I was able to summarise quantitative data from five of the six publications measuring flurothyl-induced seizures, totalling nine experiments. Here, I found evidence of heterogeneity ($I^2 = 68.2\%$; $\tau^2 = 234.6$). Unfortunately, there were not enough publications measuring the same outcome to perform a meta-analysis that would allow investigation of the potential sources of this heterogeneity.

5.5.1 Sources of heterogeneity and experimental reporting

Reporting of factors which may influence heterogeneity (e.g., animal age, animal husbandry details, and reporting of measures to reduce the risk of bias) were poor, so even if there were more publications measuring the same outcome, such a meta-analysis may still have not been possible. For research publications to be best utilised in systematic reviews and meta-analyses, or any kind of research decision-making, experiments should be transparently reported.

As a systematic reviewer, I can only know what a researcher has done in an experiment if it's clearly reported in their publication. In my publication "Designing, conducting, and reporting reproducible animal experiments", my co-authors and I signpost animal researchers towards guidance and resources designed for transparent and reproducible reporting such as the ARRIVE 2.0 guidelines and voice a call to action for animal researchers engage with these guidelines (Wilson, Ramage, et al., 2023). When speaking to animal researchers about transparent reporting, many were unaware that the ARRIVE 2.0 guidelines require authors to report both if and if not a measure to reduce bias has been performed as well as the level to detail to give.

5.5.2 Validation of the systematic map

A secondary aim of this systematic review was to validate my systematic as a comprehensive resource that could be used to perform systematic reviews. Using the search filters in the NDC-SOLES web application, I was able to retrieve ten of the total 13 publications included in this systematic review.

Two of the missing publications were BioRxiv preprints. While building my systematic map, I assumed PubMed indexed all BioRxiv preprints, however, it appears they only index NIH-funded preprints (NIH, 2024). EuropePMC, on the other hand, indexes all of PubMed and all of BioRxiv, alongside other data sources, and has a free API and R Package (Jahn, 2016). Expanding the systematic map to pull in publications from EuropePMC could therefore improve its utility as a comprehensive resource.

One publication (Komiya et al., 2002) was missed by systematic map searches as it did not contain terms relevant to autism, intellectual disability, or epilepsy. I introduced these terms in the systematic map search strategy as without them I was retrieving far too many irrelevant publications. However, I did not need to use them in my traditional systematic review search as searching for gene name alone retrieved a manageable number of publications.

5.5.3 Strengths and limitations

A strength of this review was that I preregistered my methods, to reduce the influence of my own personal biases on my results. As this review was a small review with few included studies, it meant that traditional methods of searching and screening were fairly manageable. However, the small number of studies meant that a meta-analysis to explore factors that may influence heterogeneity was not possible. Different approaches to measuring seizure outcomes and lack of reporting of experimental conduct also meant that such analysis may not have been possible even with more publications. Additionally, the feasibility of using broad search terms in my traditional search limited how I could validate my systematic map as a comprehensive resource; most systematic reviews require at least population and outcome terms to be included in search strategies (in this case *Syngap1* and seizures), but I was able to search using only population search terms due to the low number of search results. However, my validation did highlight how systematic map searches could be expanded to capture more preprints.

5.6 Conclusion

In conclusion, I confirmed that there is a degree of heterogeneity in *Syngap* rodent seizure phenotype but was unable to investigate potential factors influencing this heterogeneity. My systematic map has potential as a comprehensive resource, with expansion of data sources to more comprehensively cover preprints,

however poor reporting in publication title and abstract may prevent publications from being picked up in systematic searches.

Chapter 6: General discussion and conclusion

6.1 Introduction

In this thesis, I set out to investigate how evidence synthesis methods and automation approaches can be combined to create a systematic map of research using animals to study monogenic NDCs and then use these methods to inform improvements to experimental design and reporting. In this final chapter, I discuss the feasibility of conducting systematic reviews of fundamental neuroscience research and building systematic maps at scale. I consider what actions might be required from systematic review methodologists, primary neuroscience researchers, research publishers, research database providers, and systematic map developers to further enable the summarisation and synthesis of this research, which in turn will allow further research into features influencing the reproducibility and reliability of research findings. I also highlight the need for interdisciplinary collaborations to effectively extract metadata from texts.

6.2 Systematic reviews of fundamental neuroscience research

Systematic reviews of animal research are vital to make sense of existing evidence, identify knowledge gaps, improve animal welfare, and optimise how experiments are designed, conducted, and reported (Soliman et al., 2020). In my introduction chapter, I discussed how research findings may be irreproducible due to many factors including design, conduct, analysis and reporting of experiments. I

discussed the importance of systematic review in combining information from all available published evidence to strengthen and build our knowledge, make decisions, identify gaps, and improve experimental design and animal welfare. Additionally, I discussed that despite the heterogeneity of NDCs, a small number of common molecular pathways are involved meaning research could benefit from a way to visualise and combine evidence across different NDCs. Finally, I discussed the possibility of this through combining evidence synthesis and automation approaches.

My first objective of this thesis was to understand the aims and reporting quality of previously conducted systematic reviews in this research area. I did this by conducting a review of systematic reviews, described in Chapter 2 and published in *Brain and Neuroscience Advances* as a Registered Report. I identified 12 publications and 17 preregistered but un-published reviews. Reviews had a wide variety of aims, from investigating model suitability to sensory differences and sleep disturbances, and reviews covered 11 of the 150 genes of interest listed in Appendix 1A. Reporting quality against the PRISA-Pre checklist was mixed, with many publications reporting clear titles, eligibility criteria, and full search strategies, but few publications reporting a protocol was preregistered, methods used for screening, or assessing risk of bias.

When writing this chapter, I suggested authors conducting in this area engage with resources and guidance developed for systematic reviews of animal data, including protocol templates, risk of bias checklists, and platforms such as SyRF designed for screening and data extraction. However, after conducting my own systematic review in Chapter 5 (and discussed in more detail later in this chapter) In Chapter 5, I discovered that I had to adapt many of the resources. This is because the resources available are designed specifically for systematic reviews of preclinical animal research, where the main aim is to assess the efficacy of an intervention in an animal model, rather than for fundamental animal research which can have much broader, diverse aims and experimental set ups.

Systematic review guidance for (preclinical) animal research has been adapted from guidance used in systematic reviews of clinical randomised controlled trials (RCTs) (Sena et al., 2014). From my own experience, I now believe that guidance and resources specific to fundamental animal research are required for reviews of this research type. This may involve adapting preclinical animal systematic review guidance or perhaps guidance from other fields such as toxicology or ecology.

6.3 Curating a corpus of research publications

My second objective was to combine evidence synthesis methods and automation approaches to build a corpus of research literature that can be systematically interrogated and updated as new research is published. I conducted this work over Chapters 3 and 4. In Chapter 3, I used several methods to identify relevant research publications and in Chapter 4 I used existing metadata and software tools to build upon this dataset of relevant publications and create a searchable, interrogable resource.

6.3.1 Searching for relevant research publications

Systematic search strategies are used to identify potentially relevant research publications for evidence synthesis projects. Search strategies combine terms and synonyms relevant to the research topics of interest and multiple databases are searched to capture as many available publications as possible. In Chapter 3, I automated the searching process using APIs for PubMed, Scopus, and Web of Science. These APIs are available free-to-use; I am unable to search Embase for instance because their API is too expensive for the University of Edinburgh to purchase. In Chapter 5, when using my systematic map as part of a systematic review, I found two BioRxiv preprints were missing. I mistakenly thought that PubMed indexed all BioRxiv preprints, however it appears they index only NIH-funded preprints (NIH, 2024). This issue could be resolved by searching BioRxiv

directly or EuropePMC, which does index all BioRxiv preprints; both databases have a free API.

Another publication was missed from my systematic map search in Chapter 5 as the title and abstract did not use enough terms to be picked up in my systematic map search strategy. In Chapter 3, when testing the search strategy, I found it performed with a recall of 0.93 and did acknowledge that older publications were more likely to be missed. Adherence to the ARRIVE guidelines by clearly reporting animal species and model in title and abstract text, would reduce the likelihood of publications being missed from searches.

Interestingly, while searches of my systematic map missed three publications relevant to my systematic review, an as-yet unpublished update of a systematic review of trace amine-associated receptor 1 (TAAR1) agonism for psychosis (Siafis et al., 2024) has found additional animal research publications through searching a systematic map of animal models of psychosis (Psychosis-SOLES) compared with traditional search strategies (correspondence with Malcolm Macleod, March 2025). These findings indicate that the use of systematic maps as a starting point for systematic reviews should be further investigated and may be dependent on the systematic review topic.

6.3.1.1 Systematic review question formulation

When setting out to conduct a systematic review in any research field, it is important to ensure that your review question is relevant to the field and, to avoid duplication of effort, that no one has previously conducted a review on the same topic. In Chapter 2, I identified existing reviews in this research field and in Chapter 5 I set out to conduct my own systematic review. I collaborated with primary fundamental neuroscience researchers to understand what questions were most relevant to them and thus conducted a review aiming to explore potential factors influencing heterogeneity in seizure phenotypes of *Syngap1* rodents. In total, I identified 13 relevant publications, from which only six experiments used the same method to induce and measure seizures. However, I was unable to explore

sources of heterogeneity in these publications as the systematic review and meta-analysis methods required to do so require a minimum ten experiments.

In Chapters 3 and 4 I developed NDC-SOLES, a systematic map aiming to provide a broad overview of published research. My systematic map was not fully developed until after I had begun my systematic review, so I could not use it to help decide my review question, however one of the use cases of systematic maps is as a starting point for systematic reviews. Traditional systematic review methods involve conducting systematic searches of multiple scientific databases. Through traditional searching, I identified 840 potentially relevant publications (following removal of duplicates). Using my systematic map, which I was only able to do following my traditional database searches, I identified 58 potentially relevant publications. If I had used the systematic map to understand the quantity of publications relevant to my review question, I may have at that point suspected that there may be too few publications to investigate sources of heterogeneity and may have chosen a more common behavioural outcome or more common animal model that is still of relevance to the research field.

6.3.1.2 Animal search filters

In Chapter 2, I used an animal filter in combination with terms related to systematic reviews and neurodevelopmental conditions to find systematic reviews of monogenic NDC animal research. However, I decided against using this animal filter in later chapters, including Chapter 3. After publishing my work described in Chapter 2, I discovered that the animal filter I had used (described as a “laboratory animal filter” for systematic reviews of “animal studies”) did not in fact contain terms for invertebrate species such as *Drosophila* and *Caenorhabditis elegans*. This means that I may have missed systematic reviews using fruit fly or roundworm models if they were not picked up by other search terms such as the more general term “animal”.

The publication describing the search filter claims that “only laboratory animals as defined in the EU directive 2010/63 are included” (van der Mierden et al., 2022).

This directive refers to the “protection of animals used for scientific purposes” (European Union, 2010) which includes conversation and agriculture research in addition to laboratory research. This explains why relevantly common invertebrate species were left out while vulnerable or endangered species such as the elephant, gorilla, hippopotamus, orangutan, panda, tiger, and whale appeared in the search filter. However, it does not explain why the search filter is described as a “laboratory animal filter”.

This discrepancy is the main reason I chose not to use this filter moving forward, and instead collaborated with subject matter experts to identify common animal model species to use as search terms (Chapter 3). Using the animal filter, containing the names of hundreds of animal species not routinely used in biological laboratories, would return many more irrelevant publications than using a more restricted set of animal model search terms.

6.3.2 Screening for relevant research publications

In evidence synthesis projects, all search results are screened for relevance. The resource required for screening publications can pose a massive barrier to the feasibility of conducting evidence synthesis (Hair, Wilson, et al., 2023). While the idea behind using a machine learning classifier to screen out irrelevant publications is that it can handle larger searches, classifiers struggle when there is a steep class imbalance (e.g., much more irrelevant publications than relevant in the search results).

The complexity of search terms required to identify monogenic NDC animal research meant that many search results were irrelevant; only about 10% of search results were relevant publications. While the classifier was successful in screening out around 90% of search results, I discovered while developing NER approaches that only a little over a quarter of publications included by the classifier were actually relevant. This result makes sense when you consider the class imbalance. Class imbalance such as this is also the reason screening for signs on cancer is not done in younger people; although younger people are still

susceptible to cancer, they are less likely to develop it than older people, and therefore screening young people would lead to a high number of false positive results which could burden healthcare systems and also lead to high levels of distress in patients.

Two potential solutions to ensure that publications included by a classifier are more likely to be relevant would be (1) manually screen all publications included by the classifier to ensure relevance or (2) further limit the search strategy to ensure there is a more equal balance of relevant and irrelevant publications. Option one would involve substantial resource for manual screening but is not completely unfeasible; this would count as a “human-in-the-loop” approach to automation. Option 2 could either significantly limit the recall of relevant publications or would require additional steps to filter irrelevant search results prior to screening with the classifier. Additionally, more transparent reporting in titles and abstracts of publications may make option 2 more feasible as less search terms would be required to capture publications without negatively impacting recall.

It's clear from my work here, and the work of others (Bannach-Brown et al., 2019, 2021; Hair et al., 2024) that machine learning classification can help reduce the resource burden of screening systematic search results. However, it is important that classification tools are adequately developed and validated, that class imbalance and dataset biases are addressed, and that there is appropriate guidance on how best to implement these approaches.

6.3.3 Making use of existing research metadata

In Chapter 4, I made use of existing metadata from database providers including OpenAlex, and ROR, CrossRef, and Retraction Watch. However, article metadata was of varying quality and quantity. A small number of publications were missing a DOI, meaning full texts and metadata could not be retrieved. Research artefacts referring to the same piece of research (e.g., journal articles, preprints, conference abstracts, protocols and other materials, data, and code) cannot be easily retrieved

and linked together. Funder information was only available for 36% of the dataset despite this appearing within the texts of most publications as a journal requirement. Meanwhile, some metadata was brilliantly accessible. Using data from OpenAlex and ROR, I was able to identify first author affiliations and pinpoint these institutions geographically on a world map. Expanding from this, it would be possible to identify collaborations between researchers and, alongside citation data, perform a citation network analysis to understand the structure of the overall research field.

It is my hope that through open research initiatives, the recent Barcelona Declaration on Open Research Information, and improved digital preservation of research artefacts, we can have more equitable and complete access to quality metadata. However, currently, metadata quantity and quality are barriers to effectively summarising research at scale.

6.3.4 Automated approaches to extracting research metadata

In Chapters 3 and 4, I used automated approaches to extract additional research publication metadata that was not available via database providers. I tool several different approaches: in Chapter 3 developing my own regular expressions dictionary-based approach to extract animal model details from article texts, and in Chapter 4 using pre-existing tools.

When developing NER approaches, I chose to prioritise recall meaning that several tags represent false positives. This decision was based on feedback from researchers who said they would prefer to screen out incorrectly tagged publications than potentially miss untagged publications.

ODDPub used text data mining to extract data and code availability statements from full texts. In a separate investigation of the ODDPub tool on biosciences articles published from researchers at the University of Edinburgh, colleagues and I found performs with low precision and is unable to differentiate between publications where code was used but not shared and where code is not used at

all (Deeb et al., 2024). An additional tool named `rtransparent` has been developed as an extension of the `ODDPub` tool, aiming to also extract information about preregistrations, the content of conflicts of interest statements, and research funders (Serghiou et al., 2021). However, updates to required R packages meant the code was out of date and as the code was not actively maintained, I could not use it without needing to rewrite the functions myself which I did not have capacity for during my studies due to limited documentation of the package.

The risk of bias tool used an NLP BERT transformer model to score publications on the likelihood of reporting measures to reduce the risk of bias. A limitation of this tool was that, as it was designed for preclinical rather than fundamental research, some of the items were less relevant. Randomisation of animals to experimental group is not possible when groups are distinct only by genotype; counterbalancing (ensuring animals are similar at baseline across groups) is more appropriate (Wilson, Ramage, et al., 2023) but not detected by the tool.

Machine learning and AI tools are not only used by evidence synthesis methodologists but are already used at scale by many database providers as well. As I mentioned in Chapter 4, OpenAlex use AI and LLMs to detect abstract text language and assign disciplines or topics to research artefacts, however this metadata wasn't always useful or consistent. Additionally, some disciplines were incorrect because they lacked the contexts certain words were used in, such as ultrasonic vocalisation research being tagged as music research. Where this type of metadata could be used in another context, say summarising research publication across disciplines, this could falsely give the impression that research was being published in more diverse fields than in actuality.

I identified further issues with some OpenAlex author metadata during a data science project I took as part of my studies, which I was not able to fully explore in my own dataset. As part of the course, I developed a simple programme where a user could input an ORCID (Open Researcher and Contributor Identifier) number and the programme would search OpenAlex for that author and output a basic, structured biography detailing the author's name, affiliation, and key research

interests. Through testing this programme using my own ORCID ID, I discovered OpenAlex had misattributed publications by other researchers sharing my name to my ORCID ID. This was not a problem with ORCID but rather a problem with OpenAlex attempting to fill in missing data where an author had published an article without using an ORCID ID. OpenAlex does have a feedback mechanism to report data errors such as this, but it relies on first being aware that the issue exists. Lack of transparency over which OpenAlex data is taken directly from sources such as ORCID and which is “guessed” to fill in missing data is unfortunately a major issue with this resource.

It's important that metadata extracted using automated approaches is clearly labelled and the performance and limitations of these approaches are clearly understood by users of the metadata. Metadata extracted using methods that prioritise recall over precision, such as my NER approach, may be best used as discovery tools, where incorrect tags can be excluded through manual screening. Where metadata may be incorrect for other reasons, such as metadata extracted from ODDPub, the risk of bias tool, or using AI approaches such as LLMs, all metadata may have to be additionally extracted manually, essentially treating the automatically extracted metadata as a reviewer 1 answer in a systematic review and manually extracted metadata as reviewer 2.

6.3.5 Black boxes and transparency

Where metadata are extracted using simple dictionary- or rule-based approaches, such as those used by my NER approach and by ODDPub, it is reasonably straightforward for users with technical knowledge to understand how and why the metadata is extracted: certain text strings either did or did not appear in a certain part of the text in a certain way.

On the other hand, BERT models, LLMs, and even machine learning classifiers can be considered “black box” decision-makers: it is not obvious to a user why and how metadata is extracted or classifications are made. Despite this limitation, AI such as BERT models and LLMs are increasingly used to extract more complex

metadata that cannot be extracted using more straightforward approaches. For instance, I chose to use the BERT model for risk of bias extraction as it had been proven to perform better than a regex approach (Bahor et al., 2017; Wang et al., 2022b). Additionally, some of my colleagues working to develop a systematic map of interventions to improve reproducibility in science (iRISE-SOLES) have used an LLM to assign research disciplines to included publications; their LLM approach seems to perform more accurately than the LLM approach OpenAlex uses to assign disciplines (Hair et al., 2025).

It is important, however, to consider the impact “black box” approaches to metadata extraction could have on the discoverability of research information and how this approach might conflict with efforts to improve the transparency and openness of research. Language models, including BERT and LLMs, can process natural language (e.g. text) but are not capable of natural language understanding (Bender et al., 2021).

Rather than reading words in sentences and paragraphs as a human reader would, language models consume (and in the case of generative AI, also output) text in tokens. Tokens in computer science can refer to individual words, sets of words of n-length (n-grams), or sentences. It’s the fact that LLMs cannot “see” words the same way human readers do that they may output incorrect or fabricated information (called hallucinations) or fail to correctly answer how many times the letter R appears in the word strawberry. How potential incorrect metadata might be mitigated or dealt with is an important consideration.

To ensure reliability of metadata, next steps would include properly validating the use of pre-built automated approaches – including the BERT risk of bias model, ODDPub, and OpenAlex’s AI-assigned metadata – to gain a deeper understanding of how they perform on my dataset. A further step, which would require significant, sustained resource, would include providing “second reviewer” decisions on all automatically extracted metadata, ensuring reliability of all data. As a first step towards this, colleagues developing iRISE-SOLES are planning to implement a “feedback loop” allowing users to suggest edits to metadata or

suggest entire publications be added or removed from the systematic map based on eligibility criteria (correspondence with Sean Smith, March 2025). How much resource this will require to implement and maintain is unknown at this point.

Another critical limitation is that, as LLMs are trained on texts scrapped from the internet, they replicate the biases, discrimination, and harms that can be found on internet, including racism, misogyny, and ableism and extremism (Bender et al., 2021). As these training datasets are very large, they are impossible for the companies producing LLMs to properly curate (Bender et al., 2021). In their paper “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?”, Bender et al. hold LLM developers (such as Google and OpenAI) responsible and encourage a mindset shift towards equity, reliability, and accountability (Bender et al., 2021). I agree that such a mindset shift is needed before LLMs can be considered reliable tools for metadata extraction.

6.4 Examining research reproducibility and reliability

My third and fourth objectives were to identify features which may contribute to irreproducibility within the research and make recommendations to aid the improvement of reproducibility and reliability of evidence. I met objective three by conducting a systematic review aiming to identify factors that may influence heterogeneity in seizure phenotype in *Syngap* rodent models. Additionally, I was able to use this systematic review to assess the comprehensiveness of my systematic map as a resource (discussed earlier in this chapter). I was only partially successful in meeting objective 4 as there were too few publications, and too poor reporting of experimental procedures, to conduct a subgroup analysis to identify features which may contribute to heterogeneity. Instead, I was able to highlight how transparent and complete reporting of research planning, conduct and analysis is essential for research to be used in evidence synthesis.

6.5 Building blocks to more sustainable evidence synthesis

Throughout this thesis I have uncovered many barriers to effective identification, curation, and summarisation of fundamental neuroscience research. Overcoming these barriers will allow us to maximise how we can use published scientific research and is therefore a precursor to effective, feasible, and sustainable evidence synthesis at scale. Only by being able to effectively synthesis evidence will we be able to further research the features that may influence the reproducibility and reliability of research.

Throughout my work, I have identified key actions that are required from systematic review methodologists, primary researchers, research publishers, research database providers, and systematic map developers to overcome these challenges and make effective synthesis at scale a reality. I have outlined these actions below (Table 27), alongside the challenges they could help overcome. However, I am deeply aware that what I am suggesting is not trivial in any way. These actions represent a herculean effort requiring sustained and substantial input from multiple research users. Therefore, I have also listed key blockers to these actions that require consideration.

Research user group	Action	Challenges the action will resolve	Barriers to implementing action
Systematic review methodologists	Develop guidance specific to fundamental animal research	Researchers conducting reviews of fundamental animal research will have guidance tailored to their needs; could improve conduct and reporting of systematic reviews	Funding (systematic review methodologist salary) required to develop and test guidance and resources
Systematic review methodologists	Ensure it is clear (in title and abstract) what research type systematic review guidance is applicable to	Aid the discoverability of systematic review guidance and ensure guidance is used correctly; could improve conduct and reporting of systematic reviews.	Existing guidance is already published “as is”
Primary fundamental neuroscience researchers	Ensure that populations (experimental model and animal species) are clearly reported in title and abstract (part of the ARRIVE 2.0 essential 10 checklist)	Aid the discoverability of primary research; aid the identification of relevant research in systematic reviews and maps	Existing research is already published “as is”; relies on awareness of the ARRIVE 2.0 guidelines
Primary fundamental neuroscience researchers & Research publishers	Ensure that reporting of experiments is complete and transparent (uphold adherence to ARRIVE 2.0 guidelines)	Aid the discoverability of primary research; research can be more appropriately included in evidence synthesis; text data mining approaches	Existing research is already published “as is”; relies on awareness of and adherence of the ARRIVE 2.0 guidelines

		may be better able to extract relevant information about experimental conduct	
Research publishers & Database providers	Ensure that publication full texts are easily available and machine readable	Research can be more systematically included in systematic maps; text and data mining approaches will perform better with machine readable texts	Changes to current publishing systems where publishers retain copyright of scientific texts and prevent e.g., automated full text downloads; resource requirements for creating fully machine readable (XML, TXT, or HTML) versions of full texts; existing research is already published “as is”
Research publishers & Database providers	Addition of article metadata, accessible via API, including funder, conflicts of interest, data and code availability, study population (e.g., animal, human)	Aid the discoverability of primary research; research can be more appropriately included in evidence synthesis	Existing research metadata is available “as is”; access to some additional metadata may require removal of paywalls (to make freely accessible) and changes to API endpoints; some additional metadata may have to be extracted

			automatically using ML / AI approaches
Database providers & Systematic map providers	Provide clarity on where article metadata comes from, if ML / AI tools were used in the process, and how data were verified	Ensure trust in and reliability of the metadata; ensure transparency which can help make informed decisions on how metadata should be used	Funding (ML / AI developer and subject matter expert salary) to develop and verify ML / AI tools
Database providers & Systematic map providers	Ensure ML / AI tools are used appropriately and consider the implications, biases, or harms these tools may introduce and perpetuate through the evidence	For instance, ensure that AI / ML tools do not contribute to or exacerbate existing disparities in whose knowledge contributes to evidence (e.g., language or data source biases)	Many AI tools, such as LLMs, have these biases built-in

Table 27: Summary of actions, challenges, and barriers.

Evidence synthesis of biomedical animal research can be used to summarise knowledge in a particular area, identify gaps in the evidence, inform experimental design and identify sources of heterogeneity, inform measures to support the 3Rs (Replacement, Reduction and Refinement of animals in research), and inform decisions to justify clinical trials (Soliman et al., 2020). As I've discussed earlier, transparency and complete reporting of animal experiments will make it easier to extract information about research. However, to effectively summarise metadata at scale, there is a need for reliable and useful metadata and metadata extraction methods. To meet this aim, there also need to be effective interdisciplinary collaboration to ensure that knowledge and resources are shared and avoid duplication of effort.

Engaging with evidence synthesis methodologists in other fields outside of biomedical sciences is essential to sharing resources and developing tools that can be used in a multitude of projects. Throughout my studies, I have attended two Evidence Synthesis Hackathons, where I have collaborated with evidence synthesis methodologists, information scientists, and librarians across clinical research, social sciences, toxicology, and ecology to develop R packages that can be used in evidence synthesis projects.

There is a strong need to collaborate with computer scientists, both those with the technical skills to develop and implement automated approaches and those with the skills to evaluate the benefits and harms and ensure safety.

And finally, there is a crucial need to collaborate with qualitative researchers, such as historians and museum curators. Historians and museum curators for instance already use methods to digitise text information and extract metadata and are well versed in the challenges and opportunities in this area. While much of evidence synthesis is quantitative, processes of extracting metadata (either manually or automatically) involve extracting information from unstructured texts and therefore the learning from qualitative researchers will likely prove highly beneficial. During my studies, I attended a text and data analysis Summer School hosted by the Centre for Data Culture & Society at the University of Edinburgh.

Although I was the only researcher with a neuroscience background in attendance, I found working with and learning from historians, linguists, social scientists, and political scientists highly beneficial due to the overlap in the type of data we were working with.

6.6 Conclusion

Here I set out to combine evidence synthesis and automation approaches to create a systematic map that I could then use to inform improvements to experimental design and reporting of animal experiments, advancing our knowledge base on monogenic NDCs. I found that there is a need for more targeted guidance on conducting and reporting systematic reviews of fundamental animal research and that by combining evidence synthesis methods and automation approaches, it is possible to create a systematic map of monogenic NDC animal research. Crucially, I found that while using automation may aid the sustainability and feasibility of evidence synthesis, we must remember that “living” does not mean self-sustaining, and manual work is required to keep systematic maps running and ensure any automation approaches used are fit for purpose. And finally, I have identified action items and areas of interdisciplinary collaborations which can aid the continual improvement, utility, and transparency of such approaches.

Appendix

Supplementary materials

Supplementary materials from this thesis are shared under a Creative Commons By Attribution (CC-BY) license (<https://creativecommons.org/licenses/by/4.0/>) and have been made openly available on the Open Science Framework (OSF):

<https://doi.org/10.17605/OSF.IO/5VKRS>

- Appendix 1A: Full list of 150 relevant genes (.docx file; 6 pages)
- Appendix 2A: Chapter 2 protocol deviations (.docx file; 3 pages)
- Appendix 2B: Chapter 2 full search terms (.docx file; 23 pages)
- Appendix 2C: PRISMA-Pre reporting quality checklist (.docx file; 3 pages)
- Appendix 3A: Chapter 3 full search terms (.docx file; 4 pages)
- Appendix 3B: Initial and updated regex dictionary (.docx file; 18 pages)
- Appendix 5A: Chapter 5 full search terms (.docx file; 2 pages)
- Appendix 5B: Screening guide (.docx file; 1 page)
- Appendix 5C: Data extraction guide (.docx file; 4 pages)

Preregistrations

Preregistrations for Chapters 2 and 5 are available online at:

- Chapter 2: <https://doi.org/10.17605/OSF.IO/952QK>
- Chapter 3 (classifier) and 4: <https://doi.org/10.17605/OSF.IO/GFTZP>
- Chapter 3 (NER): <https://doi.org/10.17605/OSF.IO/5H4FB>
- Chapter 5: <https://www.crd.york.ac.uk/PROSPERO/view/CRD42024596586>

Data and code availability

Data and code used in Chapters 2- 5 are available at:

- Chapter 2 (MIT license): <https://github.com/emma-wilson/animal-sr-reporting-quality>
- Chapter 3 and 4 (GPL-3.0 license): <https://github.com/camaradesuk/ndc-soles>
- Chapter 5 (CC-BY license): <https://github.com/emma-wilson/syngap-sr>

The NDC-SOLES web application and data can be accessed here:

<https://camarades.shinyapps.io/NDC-SOLES/>

References

- Alamoudi, M. U., Hosie, S., Shindler, A. E., Wood, J. L., Franks, A. E., & Hill-Yardin, E. L. (2022). Comparing the Gut Microbiome in Autism and Preclinical Models: A Systematic Review. *Frontiers in Cellular and Infection Microbiology*, 12, 905841. <https://doi.org/10.3389/fcimb.2022.905841>
- American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders* (Fifth Edition). American Psychiatric Association. <https://doi.org/10.1176/appi.books.9780890425596>
- Antshel, K. M., & Russo, N. (2019). Autism Spectrum Disorders and ADHD: Overlapping Phenomenology, Diagnostic Issues, and Treatment Considerations. *Current Psychiatry Reports*, 21(5), 34. <https://doi.org/10.1007/s11920-019-1020-5>
- Aria, M., Le, T., & Choe, J. (2022). *openalexR: Getting Bibliographic Records from 'OpenAlex' Database Using 'DSL' API* (p. 1.4.0) [Dataset]. <https://doi.org/10.32614/CRAN.package.openalexR>
- Audunsdottir, K., Sartorius, A. M., Kang, H., Glaser, B. D., Boen, R., Nærland, T., Alaerts, K., Kildal, E. S. M., Westlye, L. T., Andreassen, O. A., & Quintana, D. S. (2024). The effects of oxytocin administration on social and routinized behaviors in autism: A preregistered systematic review and meta-analysis. *Psychoneuroendocrinology*, 167, 107067. <https://doi.org/10.1016/j.psyneuen.2024.107067>
- Bahor, Z., Liao, J., Currie, G., Ayder, C., Macleod, M., McCann, S. K., Bannach-Brown, A., Wever, K., Soliman, N., Wang, Q., Doran-Constant, L., Young, L., Sena, E. S., & Sena, C. (2021). Development and uptake of an online systematic review platform: The early years of the CAMARADES Systematic Review Facility (SyRF). *BMJ Open Science*, 5(1). <https://doi.org/10.1136/bmjos-2020-100103>

Bahor, Z., Liao, J., Macleod, M. R., Bannach-Brown, A., McCann, S. K., Wever, K. E., Thomas, J., Ottavi, T., Howells, D. W., Rice, A., Ananiadou, S., & Sena, E. (2017). Risk of bias reporting in the recent animal focal cerebral ischaemia literature. *Clinical Science*, 131(20), 2525–2532. <https://doi.org/10.1042/CS20160722>

Bannach-Brown, A., Hair, K., Bahor, Z., Soliman, N., Macleod, M., & Liao, J. (2021). Technological advances in preclinical meta-research. *BMJ Open Science*, 5(1). <https://doi.org/10.1136/bmjos-2020-100131>

Bannach-Brown, A., Przybyła, P., Thomas, J., Rice, A. S. C., Ananiadou, S., Liao, J., & Macleod, M. R. (2019). Machine learning algorithms for systematic review: Reducing workload in a preclinical review of animal studies and reducing human screening error. *Systematic Reviews*, 8(1), 23. <https://doi.org/10.1186/s13643-019-0942-7>

Barcelona Declaration. (2024). *Barcelona Declaration on Open Research Information*. barcelona-declaration.org

Basu, S. N., Kollu, R., & Banerjee-Basu, S. (2009). AutDB: A gene reference resource for autism research. *Nucleic Acids Research*, 37(suppl_1), D832–D836. <https://doi.org/10.1093/nar/gkn835>

Beadle-Brown, J., Leigh, J., Whelton, B., Richardson, L., Beecham, J., Baumker, T., & Bradshaw, J. (2016). Quality of Life and Quality of Support for People with Severe Intellectual Disability and Complex Needs. *Journal of Applied Research in Intellectual Disabilities*, 29(5), 409–421. <https://doi.org/10.1111/jar.12200>

Beller, E. M., Chen, J. K.-H., Wang, U. L.-H., & Glasziou, P. P. (2013). Are systematic reviews up-to-date at the time of publication? *Systematic Reviews*, 2(1), 36. <https://doi.org/10.1186/2046-4053-2-36>

Belter, C. (2020). *ScopusAPI* [Dataset]. <https://github.com/christopherBelter/scopusAPI>

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings*

of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623. <https://doi.org/10.1145/3442188.3445922>

Benoit, K., & Obeng, A. (2017). *readtext: Import and Handling for Plain and Formatted Text Files* (p. 0.91) [Dataset].

<https://doi.org/10.32614/CRAN.package.readtext>

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., Matsuo, A., & Lowe, W. (2015). *quanteda: Quantitative Analysis of Textual Data* (p. 4.2.0)

[Dataset]. <https://doi.org/10.32614/CRAN.package.quanteda>

Berrang-Ford, L., Sietsma, A. J., Callaghan, M., Minx, J. C., Scheelbeek, P. F. D., Haddaway, N. R., Haines, A., & Dangour, A. D. (2021). Systematic mapping of global research on climate and health: A machine learning review. *The Lancet Planetary Health*, 5(8), e514–e525. [https://doi.org/10.1016/S2542-5196\(21\)00179-0](https://doi.org/10.1016/S2542-5196(21)00179-0)

Berry-Kravis, E., Hagerman, R., Budimirovic, D., Erickson, C., Heussler, H., Tartaglia, N., Cohen, J., Tassone, F., Dobbins, T., Merikle, E., Sebree, T., Tich, N., Palumbo, J. M., & O’Quinn, S. (2022). A randomized, controlled trial of ZYN002 cannabidiol transdermal gel in children and adolescents with fragile X syndrome (CONNECT-FX). *Journal of Neurodevelopmental Disorders*, 14(1), 56.

<https://doi.org/10.1186/s11689-022-09466-6>

Bert, B., Heintz, C., Chmielewska, J., Schwarz, F., Grune, B., Hensel, A., Greiner, M., & Schönfelder, G. (2019). Refining animal research: The Animal Study Registry. *PLOS Biology*, 17(10), e3000463. <https://doi.org/10.1371/journal.pbio.3000463>

Birhane, A., & Guest, O. (2020). *Towards decolonising computational sciences* (Version 1). arXiv. <https://doi.org/10.48550/ARXIV.2009.14258>

Bölte, S., Neufeld, J., Marschik, P. B., Williams, Z. J., Gallagher, L., & Lai, M.-C. (2023). Sex and gender in neurodevelopmental conditions. *Nature Reviews Neurology*, 19(3), 136–159. <https://doi.org/10.1038/s41582-023-00774-6>

Botha, M., & Cage, E. (2022). "Autism research is in crisis": A mixed method study of researcher's constructions of autistic people and autism research. *Frontiers in Psychology*, 13, 1050897. <https://doi.org/10.3389/fpsyg.2022.1050897>

Box, G. E. P. (1976). Science and Statistics. *Journal of the American Statistical Association*, 71(356), 791–799. <https://doi.org/10.1080/01621459.1976.10480949>

Bradshaw, J., Pringle, J., Milton, D., & Beadle-Brown, J. (2024). A comparison of the sensory needs of autistic adults with and without intellectual disabilities: A short report. *Journal of Intellectual & Developmental Disability*, 49(1), 115–119. <https://doi.org/10.3109/13668250.2023.2231696>

Bramer, W. M., Giustini, D., De Jonge, G. B., Holland, L., & Bekhuis, T. (2016). De-duplication of database search results for systematic reviews in EndNote. *Journal of the Medical Library Association*, 104(3). <https://doi.org/10.5195/jmla.2016.24>

Brody, S., Loree, S., Sampson, M., Mensinkai, S., Coffman, J., Mueller, M., Askin, N., Hamill, C., Wilson, E., McAteer, M. B., Staines, H., & Best Practices for Searching During Public Health Emergencies Working Group. (2023). Searching for evidence in public health emergencies: A white paper of best practices. *Journal of the Medical Library Association*, 111(1/2), 566–578. <https://doi.org/10.5195/jmla.2023.1530>

Brunet, T., Jech, R., Brugger, M., Kovacs, R., Alhaddad, B., Leszinski, G., Riedhammer, K. M., Westphal, D. S., Mahle, I., Mayerhanser, K., Skorvanek, M., Weber, S., Graf, E., Berutti, R., Necpál, J., Havránková, P., Pavelekova, P., Hempel, M., Kotzaeridou, U., ... Wagner, M. (2021). *De novo* variants in neurodevelopmental disorders—Experiences from a tertiary care center. *Clinical Genetics*, 100(1), 14–28. <https://doi.org/10.1111/cge.13946>

Buiting, K., Williams, C., & Horsthemke, B. (2016). Angelman syndrome—Insights into a rare neurogenetic disorder. *Nature Reviews Neurology*, 12(10), 584–593. <https://doi.org/10.1038/nrneurol.2016.133>

Buller-Peralta, I., Maicas-Royo, J., Lu, Z., Till, S. M., Wood, E. R., Kind, P. C., Escudero, J., & Gonzalez-Sulser, A. (2022). Abnormal brain state distribution and network connectivity in a *SYNGAP1* rat model. *Brain Communications*, 4(6), fcac263.

<https://doi.org/10.1093/braincomms/fcac263>

Callaghan, M., Schleussner, C.-F., Nath, S., Lejeune, Q., Knutson, T. R., Reichstein, M., Hansen, G., Theokritoff, E., Andrijevic, M., Brecha, R. J., Hegarty, M., Jones, C., Lee, K., Lucas, A., Van Maanen, N., Menke, I., Pfliegerer, P., Yesil, B., & Minx, J. C. (2021). Machine-learning-based evidence and attribution mapping of 100,000 climate impact studies. *Nature Climate Change*, 11(11), 966–972.

<https://doi.org/10.1038/s41558-021-01168-6>

Campbell, F., Tricco, A. C., Munn, Z., Pollock, D., Saran, A., Sutton, A., White, H., & Khalil, H. (2023). Mapping reviews, scoping reviews, and evidence and gap maps (EGMs): The same but different— the “Big Picture” review family. *Systematic Reviews*, 12(1), 45.

<https://doi.org/10.1186/s13643-023-02178-5>

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Thomson, H. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. *BMJ*, l6890. <https://doi.org/10.1136/bmj.l6890>

Carreño-Muñoz, M. I., Chattopadhyaya, B., Agbogba, K., Côté, V., Wang, S., Lévesque, M., Avoli, M., Michaud, J. L., Lippé, S., & Di Cristo, G. (2022). Sensory processing dysregulations as reliable translational biomarkers in *SYNGAP1* haploinsufficiency. *Brain*, 145(2), 754–769. <https://doi.org/10.1093/brain/awab329>

Casajus, N., & Mouquet, N. (2022). *rwoslite: An R client for the WOS Lite API* [Dataset]. <https://github.com/FRBCesab/rwoslite>

Cascio, C. J. (2010). Somatosensory processing in neurodevelopmental disorders. *Journal of Neurodevelopmental Disorders*, 2(2), 62–69.

<https://doi.org/10.1007/s11689-010-9046-3>

Chalmers, I. (1993). The Cochrane Collaboration: Preparing, Maintaining, and Disseminating Systematic Reviews of the Effects of Health Care. *Annals of the New York Academy of Sciences*, 703(1), 156–165. <https://doi.org/10.1111/j.1749-6632.1993.tb26345.x>

Chalmers, I., & Haynes, B. (1994). Systematic Reviews: Reporting, updating, and correcting systematic reviews of the effects of health care. *BMJ*, 309(6958), 862–865. <https://doi.org/10.1136/bmj.309.6958.862>

Chamberlain, S., Zhu, H., Jahn, N., Boettiger, C., & Ram, K. (2014). *rcrossref: Client for Various "CrossRef" APIs* (p. 1.2.0) [Dataset]. <https://doi.org/10.32614/CRAN.package.rcrossref>

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2012). *shiny: Web Application Framework for R* (p. 1.10.0) [Dataset]. <https://doi.org/10.32614/CRAN.package.shiny>

Cheng, J., Schloerke, B., Karambelkar, B., & Xie, Y. (2015). *leaflet: Create Interactive Web Maps with the JavaScript "Leaflet" Library* (p. 2.2.2) [Dataset]. <https://doi.org/10.32614/CRAN.package.leaflet>

Chung, C. C. Y., Hue, S. P. Y., Ng, N. Y. T., Doong, P. H. L., Chu, A. T. W., & Chung, B. H. Y. (2023). Meta-analysis of the diagnostic and clinical utility of exome and genome sequencing in pediatric and adult patients with rare diseases across diverse populations. *Genetics in Medicine*, 25(9), 100896. <https://doi.org/10.1016/j.gim.2023.100896>

Clark, J., Glasziou, P., Del Mar, C., Bannach-Brown, A., Stehlik, P., & Scott, A. M. (2020). A full systematic review was completed in 2 weeks using automation tools: A case study. *Journal of Clinical Epidemiology*, 121, 81–90. <https://doi.org/10.1016/j.jclinepi.2020.01.008>

Clement, J. P., Aceti, M., Creson, T. K., Ozkan, E. D., Shi, Y., Reish, N. J., Almonte, A. G., Miller, B. H., Wiltgen, B. J., Miller, C. A., Xu, X., & Rumbaugh, G. (2012). Pathogenic SYNGAP1 Mutations Impair Cognitive Development by Disrupting

Maturation of Dendritic Spine Synapses. *Cell*, 151(4), 709–723.

<https://doi.org/10.1016/j.cell.2012.08.045>

Cobey, K. D., Ebrahimzadeh, S., Page, M. J., Thibault, R. T., Nguyen, P.-Y., Abu-Dalfa, F., & Moher, D. (2024). Biomedical researchers' perspectives on the reproducibility of research. *PLOS Biology*, 22(11), e3002870.

<https://doi.org/10.1371/journal.pbio.3002870>

Creson, T. K., Rojas, C., Hwaun, E., Vaissiere, T., Kilinc, M., Jimenez-Gomez, A., Holder, J. L., Tang, J., Colgin, L. L., Miller, C. A., & Rumbaugh, G. (2019). Re-expression of SynGAP protein in adulthood improves translatable measures of brain function and behavior. *eLife*, 8, e46752. <https://doi.org/10.7554/eLife.46752>

Curtis, J. S., Kennedy, S. E., Attarha, B., Edwards, L., & Jacob, R. (2021). Upper Gastrointestinal Disorders in Adult Patients with Intellectual and Developmental Disabilities. *Cureus*. <https://doi.org/10.7759/cureus.15384>

Damianidou, E., Mouratidou, L., & Kyrousi, C. (2022). Research models of neurodevelopmental disorders: The right model in the right place. *Frontiers in Neuroscience*, 16, 1031075. <https://doi.org/10.3389/fnins.2022.1031075>

De Vries, R. B. M., Hooijmans, C. R., Langendam, M. W., Van Luijk, J., Leenaars, M., Ritskes-Hoitinga, M., & Wever, K. E. (2015). A protocol format for the preparation, registration and publication of systematic reviews of animal intervention studies. *Evidence-Based Preclinical Medicine*, 2(1), 1–9. <https://doi.org/10.1002/ebm2.7>

De Vries, R. B. M., Wever, K. E., Avey, M. T., Stephens, M. L., Sena, E. S., & Leenaars, M. (2014). The Usefulness of Systematic Reviews of Animal Experiments for the Design of Preclinical and Clinical Studies. *ILAR Journal*, 55(3), 427–437.

<https://doi.org/10.1093/ilar/ilu043>

DeBoth, K. K., & Reynolds, S. (2017). A systematic review of sensory-based autism subtypes. *Research in Autism Spectrum Disorders*, 36, 44–56.

<https://doi.org/10.1016/j.rasd.2017.01.005>

Deeb, H., Creasey, S., De Ugarte, D. L., Strevens, G., Usman, T., Wong, H. Y., Kutzer, M. A. M., Wilson, E., Zieliński, T., & Millar, A. J. (2024). *The Rise of Open Data Practices Among Bioscientists at the University of Edinburgh*.

<https://doi.org/10.1101/2024.02.18.580901>

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Ozge Buzbas, E. (2021). The case for formal methodology in scientific reform. *Royal Society Open Science*, 8(3), rsos.200805, 200805. <https://doi.org/10.1098/rsos.200805>

DeWeerd, S. (2021). What studying worms, flies and fish says about autism. *Spectrum*. <https://doi.org/10.53053/QYKX3135>

Diemer, M. C., Gerstein, E. D., & Regester, A. (2022). Autism presentation in female and Black populations: Examining the roles of identity, theory, and systemic inequalities. *Autism*, 26(8), 1931–1946.

<https://doi.org/10.1177/13623613221113501>

Egger, M., Smith, G. D., & Altman, D. G. (Eds.). (2001). *Systematic Reviews in Health Care: Meta-Analysis in Context* (1st ed.). Wiley.

<https://doi.org/10.1002/9780470693926>

Eleftheriou, C., Giachetti, S., Hickson, R., Kamniti-Dumont, L., Templaar, R., Aaltonen, A., Tsoukala, E., Kim, N., Fryer-Petridis, L., Henley, C., Erdem, C., Wilson, E., Maio, B., Ye, J., Pierce, J. C., Mazur, K., Landa-Navarro, L., Petrović, N. G., Bendova, S., ... Kind, P. C. (2025). Better statistical reporting does not lead to statistical rigour: Lessons from two decades of pseudoreplication in mouse-model studies of neurological disorders. *Molecular Autism*, 16(1), 30.

<https://doi.org/10.1186/s13229-025-00663-3>

Elliott, J. H., Synnot, A., Turner, T., Simmonds, M., Akl, E. A., McDonald, S., Salanti, G., Meerpohl, J., MacLehose, H., Hilton, J., Tovey, D., Shemilt, I., Thomas, J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., Pestridge, C., ... Pearson, L. (2017). Living systematic review: 1. Introduction—the why, what, when, and how. *Journal of Clinical Epidemiology*, 91, 23–30.

<https://doi.org/10.1016/j.jclinepi.2017.08.010>

Elliott, J., Lawrence, R., Minx, J. C., Oladapo, O. T., Ravaut, P., Tendal Jeppesen, B., Thomas, J., Turner, T., Vandvik, P. O., & Grimshaw, J. M. (2021). Decision makers need constantly updated evidence synthesis. *Nature*, 600(7889), 383–385.
<https://doi.org/10.1038/d41586-021-03690-1>

Elsevier. (2025). *Elsevier Research Projects APIs* [Dataset]. <https://dev.elsevier.com/>

European Union. (2010). *Directive 2010/63/EU of the European Parliament and of the Council of 22 September 2010 on the protection of animals used for scientific purposes Text with EEA relevance*.
<https://web.archive.org/web/20250308092532/https://eur-lex.europa.eu/eli/dir/2010/63/oj>

Eve, M. P. (2014). *Open Access and the Humanities* (1st ed). Cambridge University Press.

Fallah, M. S., & Eubanks, J. H. (2020). Seizures in Mouse Models of Rare Neurodevelopmental Disorders. *Neuroscience*, 445, 50–68.
<https://doi.org/10.1016/j.neuroscience.2020.01.041>

Fenton, T. A., Haouchine, O. Y., Hallam, E. B., Smith, E. M., Jackson, K. C., Rahbarian, D., Canales, C. P., Adhikari, A., Nord, A. S., Ben-Shalom, R., & Silverman, J. L. (2024). Hyperexcitability and translational phenotypes in a preclinical mouse model of SYNGAP1-related intellectual disability. *Translational Psychiatry*, 14(1), 405.
<https://doi.org/10.1038/s41398-024-03077-6>

Fidler, F., & Wilcox, J. (2025). Reproducibility of Scientific Results. In *The Stanford Encyclopedia of Philosophy* (Winter 2025). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2025/entries/scientific-reproducibility/>

Glyph & Cog. (2025). *XpdfReader* [Dataset]. <https://www.xpdfreader.com/>.

Gu, S., Katyal, A., Zhang, Q., Chung, W., Franciosi, S., & Sanatani, S. (2023). The Association Between Congenital Heart Disease and Autism Spectrum Disorder: A

Systematic Review and Meta-Analysis. *Pediatric Cardiology*, 44(5), 1092–1107.

<https://doi.org/10.1007/s00246-023-03146-5>

Guastella, A. J., Gray, K. M., Rinehart, N. J., Alvares, G. A., Tonge, B. J., Hickie, I. B., Keating, C. M., Cacciotti-Saija, C., & Einfeld, S. L. (2015). The effects of a course of intranasal oxytocin on social behaviors in youth diagnosed with autism spectrum disorders: A randomized controlled trial. *Journal of Child Psychology and Psychiatry*, 56(4), 444–452. <https://doi.org/10.1111/jcpp.12305>

Hair, K. (2022). *Developing automated meta-research approaches in the preclinical Alzheimer's disease literature* [The University of Edinburgh].

<https://era.ed.ac.uk/handle/1842/38908>

Hair, K., Bahor, Z., Macleod, M., Liao, J., & Sena, E. S. (2023). The Automated Systematic Search Deduplicator (ASySD): A rapid, open-source, interoperable tool to remove duplicate citations in biomedical systematic reviews. *BMC Biology*, 21(1), 189. <https://doi.org/10.1186/s12915-023-01686-z>

Hair, K., Smith, S., Buljan, I., Hooijmans, C. R., Macleod, M., Marušić, A., Pejdo, D., Rackoll, T., Wever, K., Wendt, S., McCann, S. K., Sena, E., & iRISE Consortium. (2025). *iRISE-SOLES* [Dataset]. <https://camarades.shinyapps.io/irise-soles/>

Hair, K., Wilson, E., Maksym, O., Macleod, M. R., & Sena, E. S. (2024). A Systematic Online Living Evidence Summary of experimental Alzheimer's disease research. *Journal of Neuroscience Methods*, 409, 110209.

<https://doi.org/10.1016/j.jneumeth.2024.110209>

Hair, K., Wilson, E., Wong, C., Tsang, A., Macleod, M., & Bannach-Brown, A. (2023). Systematic online living evidence summaries: Emerging tools to accelerate evidence synthesis. *Clinical Science*, 137(10), 773–784.

<https://doi.org/10.1042/CS20220494>

Hardiman, R. L., & Bratt, A. (2016). Hypothalamic-pituitary-adrenal axis function in Fragile X Syndrome and its relationship to behaviour: A systematic review.

Physiology & Behavior, 167, 341–353.

<https://doi.org/10.1016/j.physbeh.2016.09.030>

Haupka, N., Culbert, J. H., Schniedermann, A., Jahn, N., & Mayr, P. (2024). *Analysis of the Publication and Document Types in OpenAlex, Web of Science, Scopus, Pubmed and Semantic Scholar* (Version 1). arXiv.

<https://doi.org/10.48550/ARXIV.2406.15154>

Heraty, S., Lautarescu, A., Belton, D., Boyle, A., Cirrincione, P., Doherty, M., Douglas, S., Plas, J. R. D., Van Den Bosch, K., Violland, P., Tercon, J., Ruigrok, A., Murphy, D. G. M., Bourgeron, T., Chatham, C., Loth, E., Oakley, B., McAlonan, G. M., Charman, T., ... Jones, E. J. H. (2023). Bridge-building between communities: Imagining the future of biomedical autism research. *Cell*, 186(18), 3747–3752.

<https://doi.org/10.1016/j.cell.2023.08.004>

Higgins, J., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (Second edition). Wiley Blackwell. <https://doi.org/10.1002/9781119536604>

Hooijmans, C. R., De Vries, R. B. M., Ritskes-Hoitinga, M., Rovers, M. M., Leeflang, M. M., IntHout, J., Wever, K. E., Hooft, L., De Beer, H., Kuijpers, T., Macleod, M. R., Sena, E. S., Ter Riet, G., Morgan, R. L., Thayer, K. A., Rooney, A. A., Guyatt, G. H., Schünemann, H. J., Langendam, M. W., & on behalf of the GRADE Working Group. (2018). Facilitating healthcare decisions by assessing the certainty in the evidence from preclinical animal studies. *PLOS ONE*, 13(1), e0187271.

<https://doi.org/10.1371/journal.pone.0187271>

Hooijmans, C. R., Rovers, M. M., De Vries, R. B., Leenaars, M., Ritskes-Hoitinga, M., & Langendam, M. W. (2014). SYRCLE's risk of bias tool for animal studies. *BMC Medical Research Methodology*, 14(1), 43. <https://doi.org/10.1186/1471-2288-14-43>

Hunniford, V. T., Montroy, J., Fergusson, D. A., Avey, M. T., Wever, K. E., McCann, S. K., Foster, M., Fox, G., Lafreniere, M., Ghaly, M., Mannell, S., Godwinska, K., Gentles, A., Selim, S., MacNeil, J., Sikora, L., Sena, E. S., Page, M. J., Macleod, M., ... Lalu, M. M. (2021). Epidemiology and reporting characteristics of preclinical systematic

reviews. *PLOS Biology*, 19(5), e3001177.

<https://doi.org/10.1371/journal.pbio.3001177>

Jahn, N. (2016). *europemc: R Interface to the Europe PubMed Central RESTful Web Service* (p. 0.4.3) [Dataset].

<https://doi.org/10.32614/CRAN.package.europemc>

Jahn, N. (2017). *roadoi: Find Free Versions of Scholarly Publications via Unpaywall* (p. 0.7.3) [Dataset]. <https://doi.org/10.32614/CRAN.package.roadoi>

Jiménez, J. A., & Zylka, M. J. (2021). Controlling litter effects to enhance rigor and reproducibility with rodent models of neurodevelopmental disorders. *Journal of Neurodevelopmental Disorders*, 13(1), 2. <https://doi.org/10.1186/s11689-020-09353-y>

Kat, R., Arroyo-Araujo, M., De Vries, R. B. M., Koopmans, M. A., De Boer, S. F., & Kas, M. J. H. (2022). Translational validity and methodological underreporting in animal research: A systematic review and meta-analysis of the Fragile X syndrome (Fmr1 KO) rodent model. *Neuroscience & Biobehavioral Reviews*, 139, 104722.

<https://doi.org/10.1016/j.neubiorev.2022.104722>

Katsanevaki, D., Till, S. M., Buller-Peralta, I., Nawaz, M. S., Louros, S. R., Kapgal, V., Tiwari, S., Walsh, D., Anstey, N. J., Petrović, N. G., Cormack, A., Salazar-Sanchez, V., Harris, A., Farnworth-Rowson, W., Sutherland, A., Watson, T. C., Dimitrov, S., Jackson, A. D., Arkell, D., ... Kind, P. C. (2024). Key roles of C2/GAP domains in SYNGAP1-related pathophysiology. *Cell Reports*, 43(9), 114733.

<https://doi.org/10.1016/j.celrep.2024.114733>

Kilinc, M., Arora, V., Creson, T. K., Rojas, C., Le, A. A., Lauterborn, J., Wilkinson, B., Hartel, N., Graham, N., Reich, A., Gou, G., Araki, Y., Bayés, À., Coba, M., Lynch, G., Miller, C. A., & Rumbaugh, G. (2022). Endogenous Syngap1 alpha splice forms promote cognitive function and seizure protection. *eLife*, 11, e75707.

<https://doi.org/10.7554/eLife.75707>

Kilinc, M., Creson, T. K., Rojas, C., Maity, S., Le, A. A., Lauterborn, J., Wilkinson, B., Hartel, N., Graham, N., Reich, A., Gou, G., Araki, Y., Bayés, À., Coba, M. P., Lynch, G., Miller, C. A., & Rumbaugh, G. (2020). *Splicing of the SynGAP Carboxyl-Terminus Enables Isoform-Specific Tuning of NMDA Receptor Signaling Linked to Cognitive Function*. <https://doi.org/10.1101/2020.01.18.911487>

Kilkenny, C., Browne, W. J., Cuthill, I. C., Emerson, M., & Altman, D. G. (2010). Improving Bioscience Research Reporting: The ARRIVE Guidelines for Reporting Animal Research. *PLoS Biology*, 8(6), e1000412. <https://doi.org/10.1371/journal.pbio.1000412>

Kim, J. H., Lee, H.-K., Takamiya, K., & Huganir, R. L. (2003). The Role of Synaptic GTPase-Activating Protein in Neuronal Development and Synaptic Plasticity. *The Journal of Neuroscience*, 23(4), 1119–1124. <https://doi.org/10.1523/JNEUROSCI.23-04-01119.2003>

Kirkham, J. J., Penfold, N. C., Murphy, F., Boutron, I., Ioannidis, J. P., Polka, J., & Moher, D. (2020). Systematic examination of preprint platforms for use in the medical and biomedical sciences setting. *BMJ Open*, 10(12), e041849. <https://doi.org/10.1136/bmjopen-2020-041849>

Komiyama, N. H., Watabe, A. M., Carlisle, H. J., Porter, K., Charlesworth, P., Monti, J., Strathdee, D. J. C., O'Carroll, C. M., Martin, S. J., Morris, R. G. M., O'Dell, T. J., & Grant, S. G. N. (2002). SynGAP Regulates ERK/MAPK Signaling, Synaptic Plasticity, and Learning in the Complex with Postsynaptic Density 95 and NMDA Receptor. *The Journal of Neuroscience*, 22(22), 9721–9732. <https://doi.org/10.1523/JNEUROSCI.22-22-09721.2002>

Kovalchik, S. (2011). *RISmed: Download Content from NCBI Databases* (p. 2.3.0) [Dataset]. <https://doi.org/10.32614/CRAN.package.RISmed>

Krapivinsky, G., Medina, I., Krapivinsky, L., Gapon, S., & Clapham, D. E. (2004). SynGAP-MUPP1-CaMKII Synaptic Complexes Regulate p38 MAP Kinase Activity and NMDA Receptor-Dependent Synaptic AMPA Receptor Potentiation. *Neuron*, 43(4), 563–574. <https://doi.org/10.1016/j.neuron.2004.08.003>

Kumar, A., Wadhawan, R., Swanwick, C. C., Kollu, R., Basu, S. N., & Banerjee-Basu, S. (2011). Animal model integration to AutDB, a genetic database for autism. *BMC Medical Genomics*, 4(1), 15. <https://doi.org/10.1186/1755-8794-4-15>

Kundap, U. P., Paudel, Y. N., & Shaikh, Mohd. F. (2020). Animal Models of Metabolic Epilepsy and Epilepsy Associated Metabolic Dysfunction: A Systematic Review. *Pharmaceuticals*, 13(6), 106. <https://doi.org/10.3390/ph13060106>

Langendam, M. W., Magnuson, K., Williams, A. R., Walker, V. R., Howdeshell, K. L., Rooney, A. A., & Hooijmans, C. R. (2021). Developing a database of systematic reviews of animal studies. *Regulatory Toxicology and Pharmacology*, 123, 104940. <https://doi.org/10.1016/j.yrtph.2021.104940>

Lazic, S. E., Clarke-Williams, C. J., & Munafò, M. R. (2018). What exactly is 'N' in cell culture and animal experiments? *PLOS Biology*, 16(4), e2005282. <https://doi.org/10.1371/journal.pbio.2005282>

Leng, G., & Ludwig, M. (2016). Intranasal Oxytocin: Myths and Delusions. *Biological Psychiatry*, 79(3), 243–250. <https://doi.org/10.1016/j.biopsych.2015.05.003>

Leser, U., & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings in Bioinformatics*, 6(4), 357–369. <https://doi.org/10.1093/bib/6.4.357>

Liao, J. (2021). *AutoAnnotation* [Dataset]. <https://github.com/camaradesuk/AutoAnnotation>

Lück, S., Callaghan, M., Borchers, M., Cowie, A., Fuss, S., Gidden, M., Hartmann, J., Kammann, C., Keller, D. P., Kraxner, F., Lamb, W. F., Mac Dowell, N., Müller-Hansen, F., Nemet, G. F., Probst, B. S., Renforth, P., Repke, T., Rickels, W., Schulte, I., ... Minx, J. C. (2025). Scientific literature on carbon dioxide removal revealed as much larger through AI-enhanced systematic mapping. *Nature Communications*, 16(1), 6632. <https://doi.org/10.1038/s41467-025-61485-8>

Lyons-Warren, A. M., Herman, I., Hunt, P. J., & Arenkiel, B. R. (2021). A systematic-review of olfactory deficits in neurodevelopmental disorders: From mouse to

human. *Neuroscience & Biobehavioral Reviews*, 125, 110–121.

<https://doi.org/10.1016/j.neubiorev.2021.02.024>

Macleod, M. R., O'Collins, T., Howells, D. W., & Donnan, G. A. (2004). Pooling of Animal Experimental Data Reveals Influence of Study Design and Publication Bias. *Stroke*, 35(5), 1203–1208. <https://doi.org/10.1161/01.STR.0000125719.25853.20>

Macleod, M. & the University of Edinburgh Research Strategy Group. (2022). Improving the reproducibility and integrity of research: What can different stakeholders contribute? *BMC Research Notes*, 15(1), 146. <https://doi.org/10.1186/s13104-022-06030-2>

Marshall, G. F., Gonzalez-Sulser, A., & Abbott, C. M. (2021). Modelling epilepsy in the mouse: Challenges and solutions. *Disease Models & Mechanisms*, 14(3), dmm047449. <https://doi.org/10.1242/dmm.047449>

Marshall, I. J., Kuiper, J., & Wallace, B. C. (2016). RobotReviewer: Evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1), 193–201. <https://doi.org/10.1093/jamia/ocv044>

Mason, D., McConachie, H., Garland, D., Petrou, A., Rodgers, J., & Parr, J. R. (2018). Predictors of quality of life for autistic adults. *Autism Research*, 11(8), 1138–1147. <https://doi.org/10.1002/aur.1965>

Mastro, T. L., Preza, A., Basu, S., Chattarji, S., Till, S. M., Kind, P. C., & Kennedy, M. B. (2020). A sex difference in the response of the rodent postsynaptic density to synGAP haploinsufficiency. *eLife*, 9, e52656. <https://doi.org/10.7554/eLife.52656>

McCann, S. K., Cramond, F., Macleod, M. R., & Sena, E. S. (2016). Systematic Review and Meta-Analysis of the Efficacy of Interleukin-1 Receptor Antagonist in Animal Models of Stroke: An Update. *Translational Stroke Research*, 7(5), 395–406. <https://doi.org/10.1007/s12975-016-0489-z>

McKenzie, K., Milton, M., Smith, G., & Ouellette-Kuntz, H. (2016). Systematic Review of the Prevalence and Incidence of Intellectual Disabilities: Current Trends and

Issues. *Current Developmental Disorders Reports*, 3(2), 104–115.

<https://doi.org/10.1007/s40474-016-0085-7>

Mignot, C., Von Stülpnagel, C., Nava, C., Ville, D., Sanlaville, D., Lesca, G., Rastetter, A., Gachet, B., Marie, Y., Korenke, G. C., Borggraefe, I., Hoffmann-Zacharska, D., Szczepanik, E., Rudzka-Dybała, M., Yiş, U., Çağlayan, H., Isapof, A., Marey, I., Panagiotakaki, E., ... Depienne, C. (2016). Genetic and neurodevelopmental spectrum of *SYNGAP1* -associated intellectual disability and epilepsy. *Journal of Medical Genetics*, 53(8), 511–522. <https://doi.org/10.1136/jmedgenet-2015-103451>

Mueller, K. F., Briel, M., Strech, D., Meerpohl, J. J., Lang, B., Motschall, E., Gloy, V., Lamontagne, F., & Bassler, D. (2014). Dissemination Bias in Systematic Reviews of Animal Research: A Systematic Review. *PLoS ONE*, 9(12), e116016.

<https://doi.org/10.1371/journal.pone.0116016>

Munafò, M. R., Chambers, C., Collins, A., Fortunato, L., & Macleod, M. (2022). The reproducibility debate is an opportunity, not a crisis. *BMC Research Notes*, 15(1), 43. <https://doi.org/10.1186/s13104-022-05942-3>

Nakai, N., Takumi, T., Nakai, J., & Sato, M. (2018). Common Defects of Spine Dynamics and Circuit Function in Neurodevelopmental Disorders: A Systematic Review of Findings From in Vivo Optical Imaging of Mouse Models. *Frontiers in Neuroscience*, 12, 412. <https://doi.org/10.3389/fnins.2018.00412>

Natri, H. M., Chapman, C. R., Heraty, S., Dwyer, P., Walker, N., Kapp, S. K., Dron, H. A., Martinez-Agosto, J. A., Mikkola, L., & Doherty, M. (2023). Ethical challenges in autism genomics: Recommendations for researchers. *European Journal of Medical Genetics*, 66(9), 104810. <https://doi.org/10.1016/j.ejmg.2023.104810>

NIH. (2024). *NIH Preprint Policy*.

<https://web.archive.org/web/20250216213238/https://pmc.ncbi.nlm.nih.gov/about/nihpreprints/>

- Nimbley, E., Golds, L., Sharpe, H., Gillespie-Smith, K., & Duffy, F. (2022). Sensory processing and eating behaviours in autism: A systematic review. *European Eating Disorders Review*, 30(5), 538–559. <https://doi.org/10.1002/erv.2920>
- Noel-Storr, A., Dooley, G., Elliott, J., Steele, E., Shemilt, I., Mavergames, C., Wisniewski, S., McDonald, S., Murano, M., Glanville, J., Foxlee, R., Beecher, D., Ware, J., & Thomas, J. (2021). An evaluation of Cochrane Crowd found that crowdsourcing produced accurate results in identifying randomized trials. *Journal of Clinical Epidemiology*, 133, 130–139. <https://doi.org/10.1016/j.jclinepi.2021.01.006>
- Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, 348(6242), 1422–1425. <https://doi.org/10.1126/science.aab2374>
- O'Collins, V. E., Macleod, M. R., Donnan, G. A., Horky, L. L., Van Der Worp, B. H., & Howells, D. W. (2006). 1,026 Experimental treatments in acute stroke. *Annals of Neurology*, 59(3), 467–477. <https://doi.org/10.1002/ana.20741>
- O'Mara-Eves, A., Thomas, J., McNaught, J., Miwa, M., & Ananiadou, S. (2015). Using text mining for study identification in systematic reviews: A systematic review of current approaches. *Systematic Reviews*, 4(1), 5. <https://doi.org/10.1186/2046-4053-4-5>
- Ooms, J. (2016). *pdfutils: Text Extraction, Rendering and Converting of PDF Documents* (p. 3.5.0) [Dataset]. <https://doi.org/10.32614/CRAN.package.pdfutils>
- Ozkan, E. D., Creson, T. K., Kramár, E. A., Rojas, C., Seese, R. R., Babayan, A. H., Shi, Y., Lucero, R., Xu, X., Noebels, J. L., Miller, C. A., Lynch, G., & Rumbaugh, G. (2014). Reduced Cognition in Syngap1 Mutants Is Caused by Isolated Damage within Developing Forebrain Excitatory Neurons. *Neuron*, 82(6), 1317–1333. <https://doi.org/10.1016/j.neuron.2014.05.015>

- Pankevich, D. E., Wizemann, T. M., Altevogt, B. M., & Institute of Medicine (U.S.) (Eds.). (2013). *Improving the utility and translation of animal models for nervous system disorders: Workshop summary*. National Academies Press.
- Panzenhagen, A. C., Cavalcanti, A., Stein, D. J., de Castro, L. L., Vasconcelos, M., Abreu, M. B., Almeida, R. F., Bertoglio, L. J., & Hermann, A. P. (2022). Behavioral manifestations in rodent models of autism spectrum disorder: A systematic review and meta-analyses. *Laboratory Animals. Abstracts of the 15th FELASA Congress 2022*, 56(S1). <https://doi.org/10.1177/0023677222110395>
- Parenti, I., Rabaneda, L. G., Schoen, H., & Novarino, G. (2020). Neurodevelopmental Disorders: From Genetics to Functional Pathways. *Trends in Neurosciences*, 43(8), 608–621. <https://doi.org/10.1016/j.tins.2020.05.004>
- Peall, K. J., Owen, M. J., & Hall, J. (2024). Rare genetic brain disorders with overlapping neurological and psychiatric phenotypes. *Nature Reviews Neurology*, 20(1), 7–21. <https://doi.org/10.1038/s41582-023-00896-x>
- Pena, V., Hothorn, M., Eberth, A., Kaschau, N., Parret, A., Gremer, L., Bonneau, F., Ahmadian, M. R., & Scheffzek, K. (2008). The C2 domain of SynGAP is essential for stimulation of the Rap GTPase reaction. *EMBO Reports*, 9(4), 350–355. <https://doi.org/10.1038/embor.2008.20>
- Percie Du Sert, N., Hurst, V., Ahluwalia, A., Alam, S., Avey, M. T., Baker, M., Browne, W. J., Clark, A., Cuthill, I. C., Dirnagl, U., Emerson, M., Garner, P., Holgate, S. T., Howells, D. W., Karp, N. A., Lazic, S. E., Lidster, K., MacCallum, C. J., Macleod, M., ... Würbel, H. (2020). The ARRIVE guidelines 2.0: Updated guidelines for reporting animal research. *PLOS Biology*, 18(7), e3000410. <https://doi.org/10.1371/journal.pbio.3000410>
- Pieper, D., & Rombey, T. (2022). Where to prospectively register a systematic review. *Systematic Reviews*, 11(1), 8. <https://doi.org/10.1186/s13643-021-01877-1>
- Pilla, B., Jordan, Z., Christian, R., Kynoch, K., McInerney, P., Cooper, K., Wu, Y., Porritt, K., Lockwood, C., & Munn, Z. (2022). JBI series paper 4: The role of

collaborative evidence networks in promoting and supporting evidence-based health care globally: reflections from 25 years across 38 countries. *Journal of Clinical Epidemiology*, 150, 210–215. <https://doi.org/10.1016/j.jclinepi.2022.04.009>

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of OA: A large-scale analysis of the prevalence and impact of Open Access articles. *PeerJ*, 6, e4375. <https://doi.org/10.7717/peerj.4375>

Plint, A. C., Moher, D., Morrison, A., Schulz, K., Altman, D. G., Hill, C., & Gaboury, I. (2006). Does the CONSORT checklist improve the quality of reports of randomised controlled trials? A systematic review. *Medical Journal of Australia*, 185(5), 263–267. <https://doi.org/10.5694/j.1326-5377.2006.tb00557.x>

Priem, J., Piwowar, H., & Orr, R. (2022). *OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts* (arXiv:2205.01833). arXiv. <https://doi.org/10.48550/arXiv.2205.01833>

R Special Interest Group on Databases (R-SIG-DB), Wickham, H., & Müller, K. (2001). *DBI: R Database Interface* (p. 1.2.3) [Dataset]. <https://doi.org/10.32614/CRAN.package.DBI>

Ramirez, F. D., Motazedian, P., Jung, R. G., Di Santo, P., MacDonald, Z. D., Moreland, R., Simard, T., Clancy, A. A., Russo, J. J., Welch, V. A., Wells, G. A., & Hibbert, B. (2017). Methodological Rigor in Preclinical Cardiovascular Studies: Targets to Enhance Reproducibility and Promote Research Translation. *Circulation Research*, 120(12), 1916–1926. <https://doi.org/10.1161/CIRCRESAHA.117.310628>

Repke, T., Tinsdeall, F., Schmidt, L., Danilenko, D., Thomas, J., Graziosi, S., Valkenhoef, G. V., & Müller-Hansen, F. (2025). *Don't stop me now, 'cause I'm having a good me screening: Evaluation of stopping methods for safe use of priority screening in systematic maps and reviews*. <https://doi.org/10.22541/au.175074486.66241013/v1>

- Riedel, N., Kip, M., & Bobrov, E. (2020). ODDPub – a Text-Mining Algorithm to Detect Data Sharing in Biomedical Publications. *Data Science Journal*, 19(1), 42. <https://doi.org/10.5334/dsj-2020-042>
- Ritskes-Hoitinga, M., & Van Luijk, J. (2019). How Can Systematic Reviews Teach Us More about the Implementation of the 3Rs and Animal Welfare? *Animals*, 9(12), 1163. <https://doi.org/10.3390/ani9121163>
- Roux, P. P., & Blenis, J. (2004). ERK and p38 MAPK-Activated Protein Kinases: A Family of Protein Kinases with Diverse Biological Functions. *Microbiology and Molecular Biology Reviews*, 68(2), 320–344. <https://doi.org/10.1128/MMBR.68.2.320-344.2004>
- Sander, T., Ghanawi, J., Wilson, E., Muhammad, S., Macleod, M., & Kahlert, U. D. (2022). Meta-analysis on reporting practices as a source of heterogeneity in in vitro cancer research. *BMJ Open Science*, 6(1). <https://doi.org/10.1136/bmjos-2021-100272>
- Sanders, S. J., Campbell, A. J., Cottrell, J. R., Moller, R. S., Wagner, F. F., Auldridge, A. L., Bernier, R. A., Catterall, W. A., Chung, W. K., Empfield, J. R., George, A. L., Hipp, J. F., Khwaja, O., Kiskinis, E., Lal, D., Malhotra, D., Millichap, J. J., Otis, T. S., Petrou, S., ... Bender, K. J. (2018). Progress in Understanding and Treating SCN2A-Mediated Disorders. *Trends in Neurosciences*, 41(7), 442–456. <https://doi.org/10.1016/j.tins.2018.03.011>
- Saran, A., & White, H. (2018). Evidence and gap maps: A comparison of different approaches. *Campbell Systematic Reviews*, 14(1), 1–38. <https://doi.org/10.4073/cmdp.2018.2>
- Satterstrom, F. K., Kosmicki, J. A., Wang, J., Breen, M. S., De Rubeis, S., An, J.-Y., Peng, M., Collins, R., Grove, J., Klei, L., Stevens, C., Reichert, J., Mulhern, M. S., Artomov, M., Gerges, S., Sheppard, B., Xu, X., Bhaduri, A., Norman, U., ... Walters, R. K. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. *Cell*, 180(3), 568–584.e23. <https://doi.org/10.1016/j.cell.2019.12.036>

Scior, K., Hamid, A., Hastings, R., Werner, S., Belton, C., Laniyan, A., Patel, M., Groce, N., & Kett, M. (2016). Consigned to the margins: A call for global action to challenge intellectual disability stigma. *The Lancet Global Health*, 4(5), e294–e295.

[https://doi.org/10.1016/S2214-109X\(16\)00060-7](https://doi.org/10.1016/S2214-109X(16)00060-7)

Scott, A. M., Glasziou, P., & Clark, J. (2023). We extended the 2-week systematic review (2weekSR) methodology to larger, more complex systematic reviews: A case series. *Journal of Clinical Epidemiology*, 157, 112–119.

<https://doi.org/10.1016/j.jclinepi.2023.03.007>

Sena, E. S., Currie, G. L., McCann, S. K., Macleod, M. R., & Howells, D. W. (2014). Systematic Reviews and Meta-Analysis of Preclinical Studies: Why Perform Them and How to Appraise Them Critically. *Journal of Cerebral Blood Flow & Metabolism*, 34(5), 737–742. <https://doi.org/10.1038/jcbfm.2014.28>

Serghiou, S., Contopoulos-Ioannidis, D. G., Boyack, K. W., Riedel, N., Wallach, J. D., & Ioannidis, J. P. A. (2021). Assessment of transparency indicators across the biomedical literature: How open is open? *PLOS Biology*, 19(3), e3001107.

<https://doi.org/10.1371/journal.pbio.3001107>

SFARI Gene. (2024a). *Animal Model Database* [Dataset].

<https://gene.sfari.org/database/animal-models/genetic-animal-models/>

SFARI Gene. (2024b). *Human Gene Database* [Dataset]. <https://gene.sfari.org>

Shaw, K. A., Bilder, D. A., McArthur, D., Williams, A. R., Amoakohene, E., Bakian, A. V., Durkin, M. S., Fitzgerald, R. T., Furnier, S. M., Hughes, M. M., Pas, E. T., Salinas, A., Warren, Z., Williams, S., Esler, A., Grzybowski, A., Ladd-Acosta, C. M., Patrick, M., Zahorodny, W., ... Maenner, M. J. (2023). Early Identification of Autism Spectrum Disorder Among Children Aged 4 Years—Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2020. *MMWR. Surveillance Summaries*, 72(1), 1–15. <https://doi.org/10.15585/mmwr.ss7201a1>

Shemilt, I., Noel-Storr, A., Thomas, J., Featherstone, R., & Mavergames, C. (2022). Machine learning reduced workload for the Cochrane COVID-19 Study Register:

Development and evaluation of the Cochrane COVID-19 Study Classifier.

Systematic Reviews, 11(1), 15. <https://doi.org/10.1186/s13643-021-01880-6>

Siafis, S., Chiocchia, V., Macleod, M. R., Austin, C., Homiar, A., Tinsdeall, F., Friedrich, C., Ramage, F. J., Kennett, J., Nomura, N., Maksym, O., Rutigliano, G., Vano, L. J., McCutcheon, R. A., Gilbert, D., Ostinelli, E. G., Stansfield, C., Dehdarirad, H., Juma, D. O., ... Leucht, S. (2024). Trace amine-associated receptor 1 (TAAR1) agonism for psychosis: A living systematic review and meta-analysis of human and non-human data. *Wellcome Open Research*, 9, 182.

<https://doi.org/10.12688/wellcomeopenres.21302.1>

Silverman, J. L., Thurm, A., Ethridge, S. B., Soller, M. M., Petkova, S. P., Abel, T., Bauman, M. D., Brodtkin, E. S., Harony-Nicolas, H., Wöhr, M., & Halladay, A. (2022). Reconsidering animal models used to study autism spectrum disorder: Current state and optimizing future. *Genes, Brain and Behavior*, 21(5), e12803.

<https://doi.org/10.1111/gbb.12803>

Smith, A. J., Clutton, R. E., Lilley, E., Hansen, K. E. A., & Brattelid, T. (2018). PREPARE: Guidelines for planning animal research and testing. *Laboratory Animals*, 52(2), 135–141. <https://doi.org/10.1177/0023677217724823>

Snilstveit, B., Vojtkova, M., Bhavsar, A., Stevenson, J., & Gaarder, M. (2016). Evidence & Gap Maps: A tool for promoting evidence informed policy and strategic research agendas. *Journal of Clinical Epidemiology*, 79, 120–129.

<https://doi.org/10.1016/j.jclinepi.2016.05.015>

Snoeijs-Schouwenaars, F. M., Young, C., Rowe, C., Van Ool, J. S., Schelhaas, H. J., & Shankar, R. (2021). People with epilepsy and intellectual disability: More than a sum of two conditions. *Epilepsy & Behavior*, 124, 108355.

<https://doi.org/10.1016/j.yebeh.2021.108355>

Sohn, E. (2017, March 15). Meet the newest contestant in the scientific rat race. *The Transmitter*. <https://www.thetransmitter.org/spectrum/meet-newest-contestant-scientific-rat-race/>

Soliman, N., Haroutounian, S., Hohmann, A. G., Krane, E., Liao, J., Macleod, M., Segelcke, D., Sena, C., Thomas, J., Vollert, J., Wever, K., Alaverdyan, H., Barakat, A., Barthlow, T., Bozer, A. L. H., Davidson, A., Diaz-delCastillo, M., Dolgorukova, A., Ferdousi, M. I., ... Rice, A. S. C. (2021). Systematic review and meta-analysis of cannabinoids, cannabis-based medicines, and endocannabinoid system modulators tested for antinociceptive effects in animal models of injury-related or pathological persistent pain. *Pain*, 162(1), S26–S44.

<https://doi.org/10.1097/j.pain.0000000000002269>

Soliman, N., Rice, A. S. C., & Vollert, J. (2020). A practical guide to preclinical systematic review and meta-analysis. *Pain*, 161(9), 1949–1954.

<https://doi.org/10.1097/j.pain.0000000000001974>

Spicer, T. P., Hubbs, C., Vaissiere, T., Colli, D., Rojas, C., Kilinc, M., Vick, K., Madoux, F., Baillargeon, P., Shumate, J., Martemyanov, K. A., Page, D. T., Puthanveetil, S., Hodder, P., Davis, R., Miller, C. A., Scampavia, L., & Rumbaugh, G. (2017). Improved Scalability of Neuron-Based Phenotypic Screening Assays for Therapeutic Discovery in Neuropsychiatric Disorders. *Complex Psychiatry*, 3(3), 141–150.

<https://doi.org/10.1159/000481731>

Steward, R., Crane, L., Mairi Roy, E., Remington, A., & Pellicano, E. (2018). “Life is Much More Difficult to Manage During Periods”: Autistic Experiences of Menstruation. *Journal of Autism and Developmental Disorders*, 48(12), 4287–4292.

<https://doi.org/10.1007/s10803-018-3664-0>

Stubben, C. (2019). *tidypmc: Parse Full Text XML Documents from PubMed Central* (p. 2.0) [Dataset]. <https://doi.org/10.32614/CRAN.package.tidypmc>

Sullivan, B. J., Ammanuel, S., Kipnis, P. A., Araki, Y., Huganir, R. L., & Kadam, S. D. (2020). Low-Dose Perampanel Rescues Cortical Gamma Dysregulation Associated With Parvalbumin Interneuron GluA2 Upregulation in Epileptic Syngap1+/- Mice. *Biological Psychiatry*, 87(9), 829–842.

<https://doi.org/10.1016/j.biopsych.2019.12.025>

- Sullivan, B. J., Kipnis, P. A., Ammanuel, S. G., & Kadam, S. D. (2020). *GABAergic signaling promotes early-life seizures in epileptic SYNGAP1^{+/-} mice*. <https://doi.org/10.1101/2020.11.14.381749>
- Summers, J., Shahrami, A., Cali, S., D'Mello, C., Kako, M., Palikucin-Reljin, A., Savage, M., Shaw, O., & Lunsky, Y. (2017). Self-Injury in Autism Spectrum Disorder and Intellectual Disability: Exploring the Role of Reactivity to Pain and Sensory Input. *Brain Sciences*, 7(11), 140. <https://doi.org/10.3390/brainsci7110140>
- Sysoeva, O. V., Smirnov, K., & Stroganova, T. A. (2020). Sensory evoked potentials in patients with Rett syndrome through the lens of animal studies: Systematic review. *Clinical Neurophysiology*, 131(1), 213–224. <https://doi.org/10.1016/j.clinph.2019.11.003>
- Thapar, A., Cooper, M., & Rutter, M. (2017). Neurodevelopmental disorders. *The Lancet Psychiatry*, 4(4), 339–346. [https://doi.org/10.1016/S2215-0366\(16\)30376-5](https://doi.org/10.1016/S2215-0366(16)30376-5)
- Thawley, A. J., Veneziani, L. P., Rabelo-da-Ponte, F. D., Riederer, I., Mendes-da-Cruz, D. A., & Bambini-Junior, V. (2022). Aberrant IL-17 Levels in Rodent Models of Autism Spectrum Disorder: A Systematic Review. *Frontiers in Immunology*, 13, 874064. <https://doi.org/10.3389/fimmu.2022.874064>
- The Dutch-Belgian Fragile X Consortium. (1994). Fmr1 knockout mice: A model to study fragile X mental retardation. *Cell*, 78(1), 23–33.
- Thomas, J., Noel-Storr, A., Marshall, I., Wallace, B., McDonald, S., Mavergames, C., Glasziou, P., Shemilt, I., Synnot, A., Turner, T., Elliott, J., Agoritsas, T., Hilton, J., Perron, C., Akl, E., Hodder, R., Pestrige, C., Albrecht, L., Horsley, T., ... Pearson, L. (2017). Living systematic reviews: 2. Combining human and machine effort. *Journal of Clinical Epidemiology*, 91, 31–37. <https://doi.org/10.1016/j.jclinepi.2017.08.011>
- Trevisan, D. A., Parker, T., & McPartland, J. C. (2021). First-Hand Accounts of Interoceptive Difficulties in Autistic Adults. *Journal of Autism and Developmental Disorders*, 51(10), 3483–3491. <https://doi.org/10.1007/s10803-020-04811-x>

Tufanaru, C., Surian, D., Scott, A. M., Glasziou, P., & Coiera, E. (2024). The 2-week systematic review (2weekSR) method was successfully blind-replicated by another team: A case study. *Journal of Clinical Epidemiology*, 165, 111197.

<https://doi.org/10.1016/j.jclinepi.2023.10.013>

Understanding Animal Research. (2025). *Animal research species*.

<https://www.understandinganimalresearch.org.uk/using-animals-in-scientific-research/animal-research-species>

van der Mierden, S., Hooijmans, C. R., Tillema, A. H., Rehn, S., Bleich, A., & Leenaars, C. H. (2022). Laboratory animals search filter for different literature databases: PubMed, Embase, Web of Science and PsycINFO. *Laboratory Animals*, 56(3), 279–286. <https://doi.org/10.1177/00236772211045485>

Van Der Naald, M., Chamuleau, S. A. J., Menon, J. M. L., De Leeuw, W., De Haan, J., Duncker, D. J., & Wever, K. E. (2022). Preregistration of animal research protocols: Development and 3-year overview of preclinicaltrials.eu. *BMJ Open Science*, 6(1). <https://doi.org/10.1136/bmjos-2021-100259>

Van Der Worp, H. B., Howells, D. W., Sena, E. S., Porritt, M. J., Rewell, S., O'Collins, V., & Macleod, M. R. (2010). Can Animal Models of Disease Reliably Inform Human Studies? *PLoS Medicine*, 7(3), e1000245.

<https://doi.org/10.1371/journal.pmed.1000245>

Verkerk, A. J. M. H., Pieretti, M., Sutcliffe, J. S., Fu, Y.-H., Kuhl, D. P. A., Pizzuti, A., Reiner, O., Richards, S., Victoria, M. F., Zhang, F., Eussen, B. E., Van Ommen, G.-J. B., Blonden, L. A. J., Riggins, G. J., Chastain, J. L., Kunst, C. B., Galjaard, H., Thomas Caskey, C., Nelson, D. L., ... Warren, S. T. (1991). Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, 65(5), 905–914.

[https://doi.org/10.1016/0092-8674\(91\)90397-H](https://doi.org/10.1016/0092-8674(91)90397-H)

Verma, V., Kumar, M. J. V., Sharma, K., Rajaram, S., Muddashetty, R., Manjithaya, R., Behnisch, T., & Clement, J. P. (2022). Pharmacological intervention in young adolescents rescues synaptic physiology and behavioural deficits in Syngap1+/-

mice. *Experimental Brain Research*, 240(1), 289–309.

<https://doi.org/10.1007/s00221-021-06254-x>

Vernes, S. C., Newbury, D. F., Abrahams, B. S., Winchester, L., Nicod, J., Groszer, M., Alarcón, M., Oliver, P. L., Davies, K. E., Geschwind, D. H., Monaco, A. P., & Fisher, S. E. (2008). A Functional Genetic Link between Distinct Developmental Language Disorders. *New England Journal of Medicine*, 359(22), 2337–2345.

<https://doi.org/10.1056/NEJMoa0802828>

Vlaskamp, D. R. M., Shaw, B. J., Burgess, R., Mei, D., Montomoli, M., Xie, H., Myers, C. T., Bennett, M. F., XiangWei, W., Williams, D., Maas, S. M., Brooks, A. S., Mancini, G. M. S., Van De Laar, I. M. B. H., Van Hagen, J. M., Ware, T. L., Webster, R. I., Malone, S., Berkovic, S. F., ... Scheffer, I. E. (2019). *SYNGAP1* encephalopathy: A distinctive generalized developmental and epileptic encephalopathy. *Neurology*, 92(2). <https://doi.org/10.1212/WNL.0000000000006729>

Walkup, W. G., Washburn, L., Sweredoski, M. J., Carlisle, H. J., Graham, R. L., Hess, S., & Kennedy, M. B. (2015). Phosphorylation of Synaptic GTPase-activating Protein (synGAP) by Ca²⁺/Calmodulin-dependent Protein Kinase II (CaMKII) and Cyclin-dependent Kinase 5 (CDK5) Alters the Ratio of Its GAP Activity toward Ras and Rap GTPases. *Journal of Biological Chemistry*, 290(8), 4908–4927.

<https://doi.org/10.1074/jbc.M114.614420>

Wang, Q., Liao, J., Lapata, M., & Macleod, M. (2022a). PICO entity extraction for preclinical animal literature. *Systematic Reviews*, 11(1), 209.

<https://doi.org/10.1186/s13643-022-02074-4>

Wang, Q., Liao, J., Lapata, M., & Macleod, M. (2022b). Risk of bias assessment in preclinical literature using natural language processing. *Research Synthesis Methods*, 13(3), 368–380. <https://doi.org/10.1002/jrsm.1533>

Wickham, H., Hester, J., & Ooms, J. (2015). *xml2: Parse XML* (p. 1.3.7) [Dataset].

<https://doi.org/10.32614/CRAN.package.xml2>

Wijetunge, L. S., Chattarji, S., Wyllie, D. J. A., & Kind, P. C. (2013). Fragile X syndrome: From targets to treatments. *Neuropharmacology*, 68, 83–96.
<https://doi.org/10.1016/j.neuropharm.2012.11.028>

Wilde, M., Constantin, L., Thorne, P. R., Montgomery, J. M., Scott, E. K., & Cheyne, J. E. (2022). Auditory processing in rodent models of autism: A systematic review. *Journal of Neurodevelopmental Disorders*, 14(1), 48.
<https://doi.org/10.1186/s11689-022-09458-6>

Wiley. (2025). *Wiley Text and Data Mining API* [Dataset].
<https://onlinelibrary.wiley.com/library-info/resources/text-and-datamining>

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. 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). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1), 160018.
<https://doi.org/10.1038/sdata.2016.18>

Wilson, E., Cruz, F., Maclean, D., Ghanawi, J., McCann, S. K., Brennan, P. M., Liao, J., Sena, E. S., & Macleod, M. (2023). Screening for *in vitro* systematic reviews: A comparison of screening methods and training of a machine learning classifier. *Clinical Science*, 137(2), 181–193. <https://doi.org/10.1042/CS20220594>

Wilson, E., Ramage, F. J., Wever, K. E., Sena, E. S., Macleod, M. R., & Currie, G. L. (2023). Designing, conducting, and reporting reproducible animal experiments. *Journal of Endocrinology*, 258(1), e220330. <https://doi.org/10.1530/JOE-22-0330>

Wong, C., Gregory, J. M., Liao, J., Egan, K., Vesterinen, H. M., Ahmad Khan, A., Anwar, M., Beagan, C., Brown, F. S., Cafferkey, J., Cardinali, A., Chiam, J. Y., Chiang, C., Collins, V., Dormido, J., Elliott, E., Foley, P., Foo, Y. C., Fulton-Humble, L., ... Macleod, M. (2023). Systematic, comprehensive, evidence-based approach to identify neuroprotective interventions for motor neuron disease: Using systematic reviews to inform expert consensus. *BMJ Open*, 13(2), e064169.
<https://doi.org/10.1136/bmjopen-2022-064169>

Wright, C. F., Campbell, P., Eberhardt, R. Y., Aitken, S., Perrett, D., Brent, S., Danecek, P., Gardner, E. J., Chundru, V. K., Lindsay, S. J., Andrews, K., Hampstead, J., Kaplanis, J., Samocha, K. E., Middleton, A., Foreman, J., Hobson, R. J., Parker, M. J., Martin, H. C., ... Firth, H. V. (2023). Genomic Diagnosis of Rare Pediatric Disease in the United Kingdom and Ireland. *New England Journal of Medicine*, 388(17), 1559–1571. <https://doi.org/10.1056/NEJMoa2209046>

Yang, Y., Macleod, M., Pan, J., Lagisz, M., & Nakagawa, S. (2023). Advanced methods and implementations for the meta-analyses of animal models: Current practices and future recommendations. *Neuroscience & Biobehavioral Reviews*, 146, 105016. <https://doi.org/10.1016/j.neubiorev.2022.105016>

Zeng, M., Shang, Y., Araki, Y., Guo, T., Huganir, R. L., & Zhang, M. (2016). Phase Transition in Postsynaptic Densities Underlies Formation of Synaptic Complexes and Synaptic Plasticity. *Cell*, 166(5), 1163-1175.e12. <https://doi.org/10.1016/j.cell.2016.07.008>

Zhang, X., Lin, J., & Spruyt, K. (2021). Sleep problems in Rett syndrome animal models: A systematic review. *Journal of Neuroscience Research*, 99(2), 529–544. <https://doi.org/10.1002/jnr.24730>

Zhang, X.-Y., & Spruyt, K. (2022). Disrupted sleep in rett syndrome animal models. *Sleep Medicine*, 100, S178. <https://doi.org/10.1016/j.sleep.2022.05.479>

Zhao, X., & Bhattacharyya, A. (2018). Human Models Are Needed for Studying Human Neurodevelopmental Disorders. *The American Journal of Human Genetics*, 103(6), 829–857. <https://doi.org/10.1016/j.ajhg.2018.10.009>

Zoghbi, H. Y., & Bear, M. F. (2012). Synaptic Dysfunction in Neurodevelopmental Disorders Associated with Autism and Intellectual Disabilities. *Cold Spring Harbor Perspectives in Biology*, 4(3), a009886–a009886. <https://doi.org/10.1101/cshperspect.a009886>