The Landscape of Artificial Intelligence in Neurodegenerative Diseases: a systematic review

Walter Endrizzi[1,2], Flavio Ragni[1], Stefano Bovo[1], Avinash Chandra[3], Monica Moroni[1], Giuseppe Jurman[1,4], and Venet Osmani[5]

[1]Fondazione Bruno Kessler, Data Science for Health, Trento, 38123, Italy
[2]University of Trento, Department of Cellular, Computational and Integrative Biology, Trento, 38123, Italy
[3]Centre for Preventive Neurology, Wolfson Institute of Population Health, Queen Mary University of London, London EC1M 6BQ, United Kingdom
[4]Humanitas University, Department of Biomedical Sciences, Milan, 20072, Italy
[5]Digital Environment Research Institute, Queen Mary University of London, London, E1 1HH, United Kingdom
[*]Corresponding author: v.osmani@qmul.ac.uk

May 21, 2026

Abstract

Background: The rising global burden of neurodegenerative diseases underscores an urgent need for advanced research in diagnosis, prognosis, and treatment. Artificial Intelligence (AI) methods, particularly when applied to multimodal data, offer a powerful tool to address these challenges. However, a comprehensive overview and critique of the current landscape of AI methods is lacking.

Methods: 4,685 records of peer-reviewed primary research articles were screened and 1,956 articles reviewed in full text, yielding 1,186 included studies. For each included study, clinical objectives, disease focus, data modalities, modelling approach, evaluation strategy, and reporting practices were extracted.

Results: Fewer than 5% of studies integrated pharmacological treatments into their predictive models, limiting the extent to which models can directly inform clinical decision-making. Neuroimaging was the predominant input modality, while integration of other clinically relevant data types was relatively rare. Reproducibility rates remain critically low at 35%, and external validation practices fail to use geographically and demographically diverse datasets.

Conclusions: Overall, AI research in neurodegenerative diseases suffers from significant limitations in reproducibility, data inclusivity, and clinical translatability. We provide a set of recommendations that can be adopted to address these issues and improve reliability and downstream clinical utility.

Plain language summary

Neurodegenerative diseases such as Alzheimer's and Parkinson's are increasing as populations age. Researchers are using artificial intelligence (AI) to support early diagnosis, predict progression, and

improve care by learning from brain scans, medical records, and lab tests. We reviewed a large number of published studies to understand how close research is to real clinical use. Three gaps stand out: few models include information on medicines or treatments, limiting treatment-aware decisions; most rely mainly on brain imaging while other routine clinical data are used less often; and many studies lack sufficient data/code and independent testing on diverse populations, making results hard to reproduce and generalize. We offer practical recommendations to make future AI tools more reliable, inclusive, and clinically useful.

Introduction

Millions of individuals worldwide are affected by neurodegenerative diseases such as Alzheimer's disease (AD) and other types of dementia, Parkinson's disease (PD), Huntington's disease (HD), and amyotrophic lateral sclerosis (ALS) [9]. Significant emotional, physical and financial strain falls on the patients, their families and caregivers, constituting a rapidly growing global health challenge. In addition, a sharp increase in prevalence rates exerts immense pressure on healthcare systems due to the high costs of long-term care, diagnostics, and symptoms management [30]. AD and other dementias alone will cost the world economy $15 trillion by 2050 [3]. Lack of curative treatments for most neurodegenerative conditions and the projected increase in longevity globally [21] has created an urgent need towards cutting-edge research in diagnosis, prognosis, improved care and preventive strategies. Innovative solutions are imperative when considering an additional projected shortage of clinicians specialising in brain health [16][2]. Artificial Intelligence (AI) can be considered an umbrella term encompassing diverse computational paradigms, including rule-based or symbolic systems, conventional Machine Learning (ML) and Deep Learning (DL) approaches, and, more recently, foundation model–based methods. In the present review, however, the retrieved literature was composed almost exclusively of conventional ML and DL studies. These methods are particularly relevant to neurodegenerative research because they enable the extraction of complex patterns from large, heterogeneous, and multimodal datasets, making them well suited to address the multifaceted burden of these diseases [13]. Across healthcare fields, multimodal AI has demonstrated particular robustness relative to single modality approaches to precision medicine [15]. Several applications of AI methods to neurodegenerative conditions exist. AI algorithms may promote an earlier and more accurate diagnosis [17], with preliminary evidence suggesting that multimodal AI may improve the diagnostic efficacy of neurologists [29]. By analysing sources such as medical images (e.g. MRI and PET scans), genetic data, clinical records, and even subclinical symptoms like changes in speech patterns or gait, AI can aid with identifying disease onset well-before significant irreversible damage occurs, determining disease subtype grouping, or patient prognosis. This will then lead to formulation of personalised treatment strategies for patients suffering from neurodegenerative conditions.

However, tracking the rapid advancements is particularly challenging. Despite previous reviews addressing narrower scopes - e.g. applications of AI to PD [20, 19]) or AD [6, 28] - fewer studies synthesize AI methods across multiple neurodegenerative disorders. Additionally, we have identified a gap in the comprehensive investigation of several key factors that may influence translation to clinical practice, including data quality, methodological rigour and the research scope.

In response, we conducted the largest systematic review to date with 4685 articles screened and 1956 articles reviewed in detail, on applications of AI to a broad spectrum of neurodegenerative diseases. Here, we examine five specific areas: (1) *Clinical objectives* (such as diagnosis, prognosis, and treatment) addressed by AI research, (2) *Data modalities* (such as imaging, biomarkers and

omics) used and whether the heterogeneity found in clinical practice is reflected in the research literature (3) *AI methods* and the associated performance metrics, (4) *Reproducibility* of AI studies given transparency of methodology and dataset availability, and (5) *Translation to clinical practice*, with respect to external validation, generalisability in diverse patient populations as well as the interpretability of AI models. Our analysis reveals that fewer than 5% of studies incorporated pharmacological treatment information into their models. Neuroimaging was the predominant data source, while integration of other clinically relevant modalities was comparatively rare. Reproducibility remains low (35%), and external validation rarely uses geographically and demographically diverse cohorts. We further contrast these results with a subset of the most influential studies in our cohort, to investigate whether the forefront of the field overcomes the limitations observed in the broader ecosystem or merely replicates them. Finally, we provide a set of recommendations that can be adopted to address many of the limitations inherent in this research area. Doing so may help promote the clinical utility of AI for the diagnosis, prognosis, and treatment of neurodegenerative disorders.

Methods

The systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) guidelines [22].

Eligibility Criteria

The inclusion criteria applied during the screening process encompassed publication type, methods, subject and aim of the study. Eligible studies were restricted to peer-reviewed primary research articles published in English with full text available. The studies must address a clinical outcome related to at least one neurodegenerative disease between AD, PD, Multiple Sclerosis (MS), ALS, Prion Disease and Dementia, by means of at least one ML or DL method. Studies not involving human participants, as well as those deemed to be of insufficient methodological quality (e.g., very small retrospective samples, inadequate control groups, or suboptimal data analysis), were excluded.

Information Sources and Search Strategy

The literature search was conducted in June 2024 across the following databases, using the title and abstract fields: SCOPUS, Web of Science, IEEE Digital Library, and ACM Digital Library. The search was restricted to journal articles published in English between June 2019 and June 2024. The complete search terms for each database are reported in the Supplementary Materials, Table S1.

Selection Process

Records retrieved through the search strategy described above were de-duplicated using the Zotero [10] automation tool, followed by an additional manual screening to remove any remaining articles with duplicate titles. After de-duplication, studies underwent a two-stage screening process, performed by five investigators with varying levels of expertise.

In the first stage, studies were assigned to reviewers for eligibility assessment based on title and abstract. All studies meeting the inclusion criteria at this stage were then randomly redistributed among the reviewers for the second stage, which consisted of full-text evaluation to further refine the list of eligible studies. During this second stage, eligible records were additionally classified as reproducible or non-reproducible according to two criteria: (i) the use of publicly available datasets (either open-access or subscription-based), and (ii) the availability of code when novel architectures or algorithms were proposed. We note that reproducibility exists on a gradient, from re-running the same analysis with the same code and data to independent reproduction with re-implementation and/or independent cohorts. Our criteria represent a liberal and pragmatic definition, intended to capture whether studies provide sufficient elements to enable reproduction. Accordingly, throughout the manuscript we use the term 'reproducible' to mean reproducibility-enabled, and thus reproducibility estimates likely overstate the proportion of papers that would be reproducible under stricter, practice-oriented definitions. Detailed information on the selection process is provided in Table S2.

In both screening stages, uncertainties regarding the eligibility or reproducibility of a study were resolved through consensus, following discussion among all reviewers in recurrent internal meetings.

Data Collection Process

For each included study, one reviewer independently performed manual data extraction without the use of automated tools, using an Excel spreadsheet to record the information. Extracted data included the primary objective and any additional aims, the specific neurodegenerative disease investigated, and the type of data analysed. The reviewer also documented whether DL methods were employed and collected article-specific information such as the country of origin, defined by the institutional affiliation of the first author. For studies identified as reproducible during the second screening stage, additional information was collected. This included whether treatment-related information was reported, as well as detailed characteristics of the data source, such as the name of the dataset, its availability and access cost, the country of origin, whether the dataset was multicentric, and the presence of longitudinal data and follow-up information. The number of subjects and features included (where applicable) was also recorded. Information regarding the ML or DL methods was extracted, including the best-performing model, other models evaluated, and whether the study or model was multimodal. Evaluation strategies were also documented, including the presence of external validation, reported performance metrics, and the inclusion of algorithmic explainability. Finally, the availability of the code used in the study was recorded. A complete overview of the extracted features for both reproducible and non-reproducible studies is presented in Table S3. Citation counts for each article were retrieved on January 2026, using a custom Python script employing a hybrid retrieval strategy to ensure data completeness: DOIs were first queried against the Semantic Scholar Graph API [14], with a secondary fallback query against the OpenAlex API [24].

Risk of Bias Assessment

At the end of each screening stage, a report was generated to summarize the statistics of included and excluded articles for each reviewer. The report included the percentage of rejected papers, the distribution of these rejections across the different exclusion criteria, and the percentage of papers classified as reproducible during the second screening stage. These values were compared across reviewers to assess potential discrepancies in the selection process and to mitigate the risk biases during study selection. Reviewer-level summary reports used for this assessment can be found in the Supplementary Materials (Table S4).

Consistency of findings in high-influence studies

All main-text analyses were replicated on the most influential studies within the included cohort. We defined "influence" as the number of citations divided by the number of years since publication, even though we are aware of the limitations of using citation counts alone. This normalization enables fair comparison across publication years and reduce biases toward older studies that have had more time to accumulate citations. Although the choice of an "influence" metric is inherently debatable, citations per year provides a transparent, reproducible way of parametrizing a paper's impact on the broader literature.

Results

Literature search results

We identified 8203 publications, with 4682 unique abstracts screened, and 1953 studies reviewed in full text. After the full text review, 1186 publications met the inclusion criteria (see Figure 1 for a graphical representation), with a total of 421 studies identified as reproducible according to the criteria described in the Methods section. The study selection flow is reported in a PRISMA flowchart (Figure 2); the complete list of screened studies, their characteristics, references and reasons for exclusion are available at {*redacted for review*} repository.

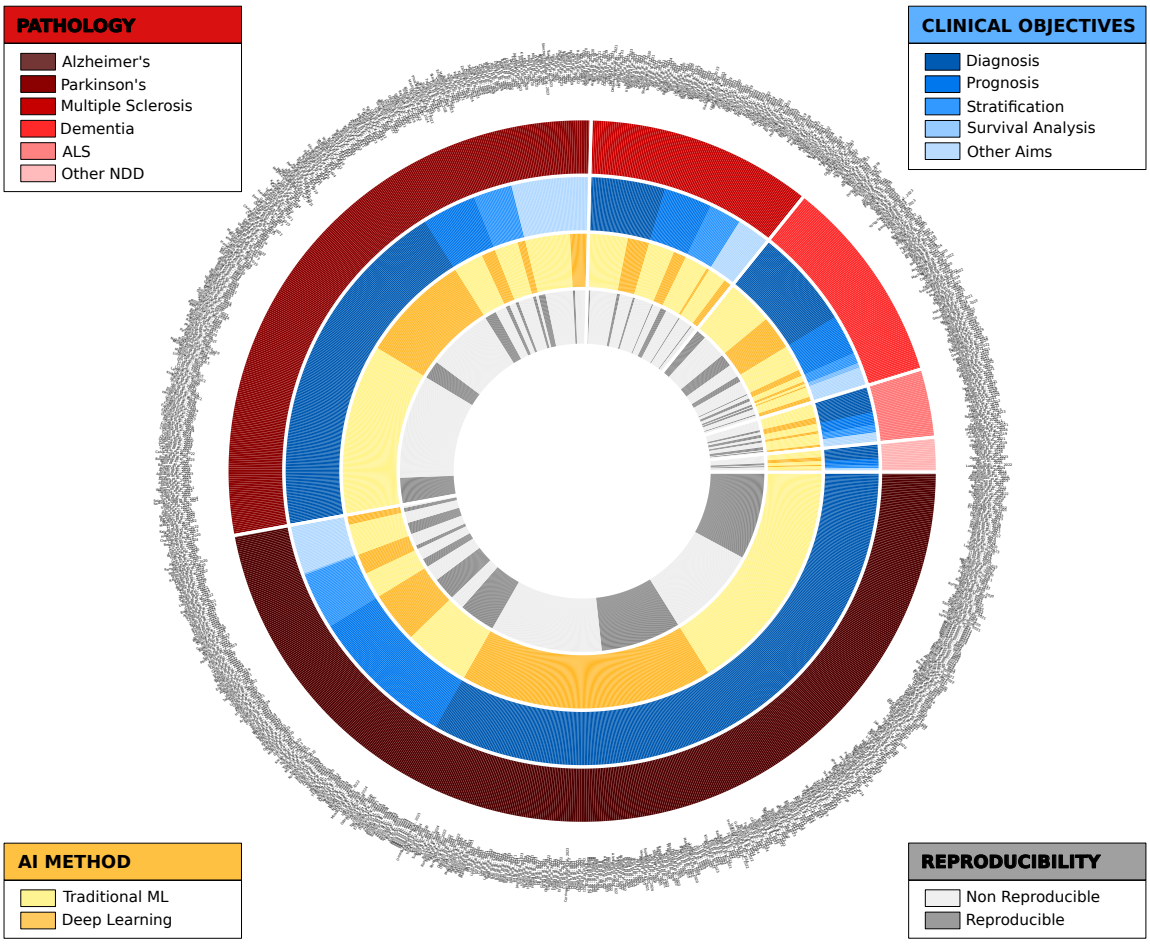

Figure 1: **Overview of the major findings on AI applications for neurodegenerative diseases literature review.** The outermost ring represents publications meeting the inclusion criteria after full-text review (n=1186), where each line corresponds to one included study, formatted as *First Author et al., Year*. Inner rings categorise publications by pathology (red); clinical objectives (blue); AI methods (yellow); and reproducibility (gray).

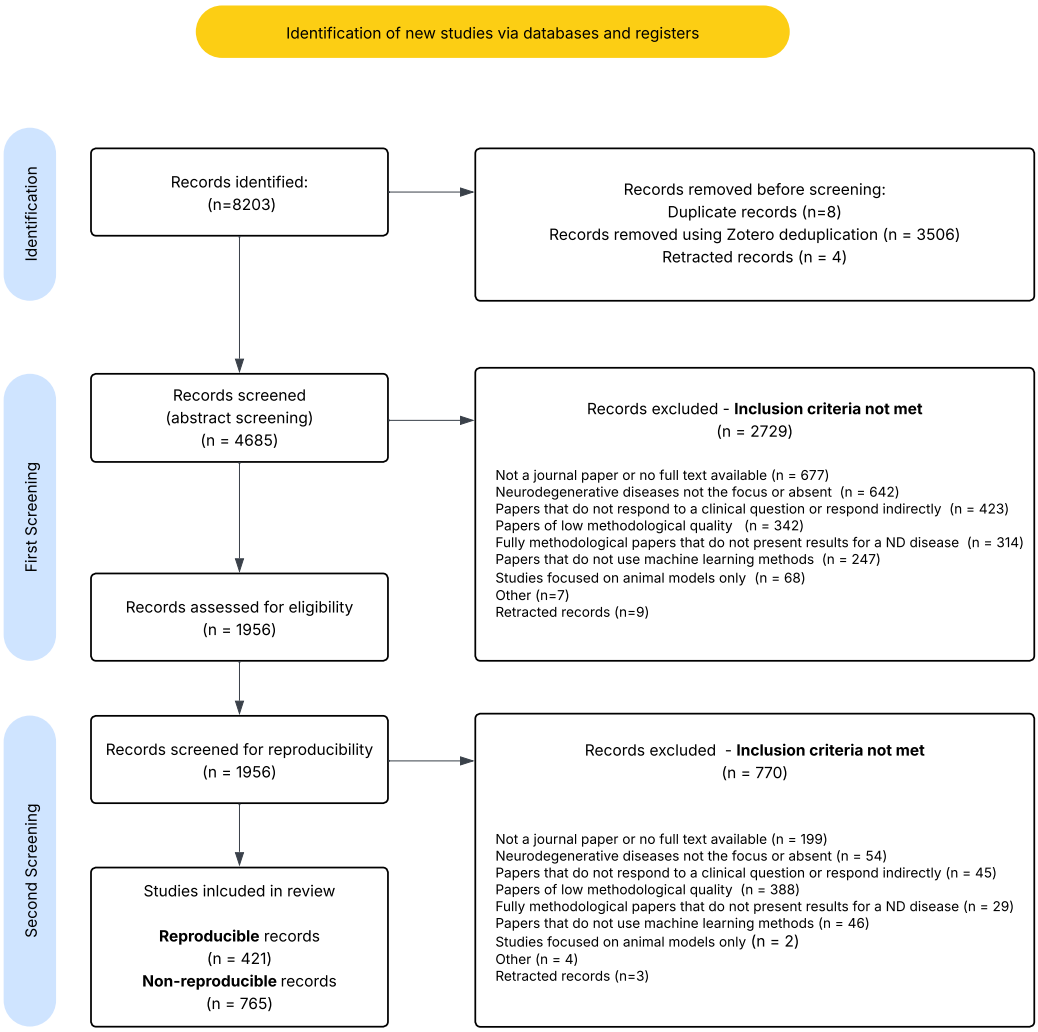

Figure 2: **PRISMA flowchart.** Detailed selection process for the systematic review. The figure outlines the number of records identified from databases, the subsequent screening of articles, and the final set of studies included in the qualitative synthesis. ND: neurodegenerative.

Clinical objectives of AI application to neurodegenerative diseases

Our systematic review reveals that the landscape of AI applications in neurodegenerative diseases is multifaceted, covering a range of diverse pathologies and clinical objectives (Figure 3). The five most represented neurodegenerative diseases are AD, PD, Dementia all cause), MS and ALS. The clinical objectives are categorized into diagnosis, prognosis, stratification, survival analysis, and other, with diagnosis being the most frequently addressed objective across all disease types.

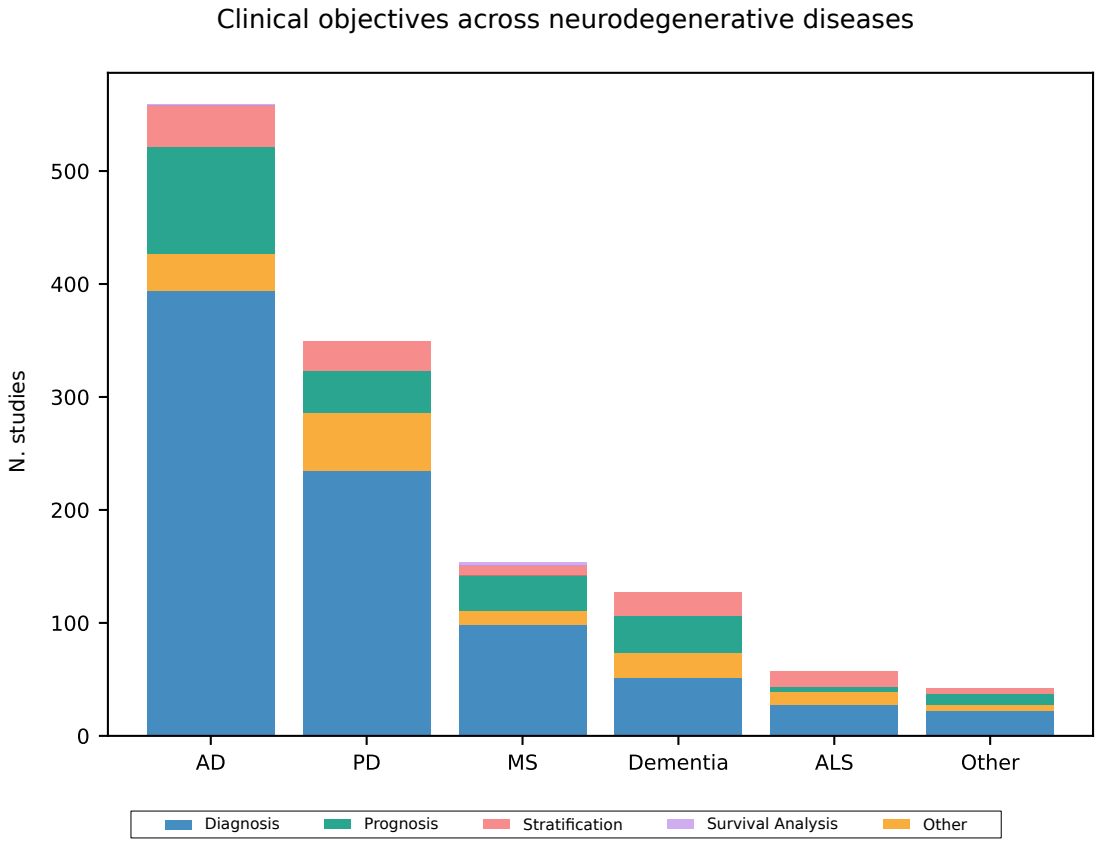

Figure 3: **Distribution of AI studies in neurodegenerative diseases across pathologies and clinical objectives.** Studies are categorised by the primary clinical objective, namely diagnosis, prognosis, stratification, survival analysis, or other. AD: Alzheimer's Disease, PD: Parkinson's Disease, MS: Multiple Sclerosis, ALS: Amyotrophic Lateral Sclerosis.

Across all reviewed manuscripts (n=1186), AD is the most represented condition with 559 studies, with major focus on diagnosis (n=394, 70.48%), followed by prognosis (n=94, 16.82%) and stratification (n=37, 6.62%). PD is represented with 349 studies, following a similar trend: diagnosis is the most common objective (n=235, 67.33%), followed by prognosis (n=37, 10.6%). For Dementia, MS and ALS, the volume of studies is lower, yet diagnostic applications remain predominant. Overall, while diagnosis is consistently the dominant clinical objective across all diseases, applications focusing on prognosis, stratification, and survival analysis are comparatively under-

represented, particularly in diseases with fewer studies.
Despite the broad application of AI to various clinical objectives, in depth review of the screened reproducible studies revealed that disease treatment remains an unexplored area, with only 21 manuscripts (4.9%) explicitly incorporating treatment-related information in their modelling approach. This remarkably low proportion highlights a significant gap in the neurodegenerative disease literature, where AI is primarily used for early-stage objectives such as diagnosis or short-term prognosis, but rarely extended to support treatment decision-making, response prediction or therapy optimization. This could reflect challenges in accessing longitudinal fine-grained treatment data, especially in publicly available datasets, which tend to prioritize imaging and clinical variables over detailed pharmacological histories. In addition, in some neurodegenerative diseases, the lack of effective treatment options or the presence of rapid disease progression may shift research priorities towards early detection and risk stratification rather than therapeutic modelling. Moreover, even when treatments exist, they may not act directly on the neurodegenerative component: for example, disease-modifying therapies for relapsing forms of MS target inflammatory processes in the central nervous system [23], which are particularly relevant to this condition. Therefore, modelling approaches should be careful in assuming a common degenerative framework. Nevertheless, given the growing interest in precision and personalised medicine, the limited integration of treatment information represents a missed opportunity for AI to have a more direct impact on patient care and to enhance its translational impact.

Integration of diverse data modalities

Given the complex aetiology and pathophysiology of neurodegenerative diseases, another important aspect to consider when assessing research studies is the integration of diverse data sources and the alignment with the inherent multimodality of clinical practice. Regarding the heterogeneity of data sources, the considered reproducible studies include imaging, Electronic Health Records (EHRs), omics, biomarkers, neuropsychological data, speech, waveforms, and wearables, among others.
In studies based on pre-existing datasets (Figure 4a), imaging data is by far the most commonly used modality (n=194), with AD and PD driving this trend. This suggests a strong dependence on well-established imaging repositories, particularly in AD research (i.e. the Alzheimer's Disease Neuroimaging Initiative - ADNI). Other widely adopted modalities include EHRs and omics, although to a much lesser extent, and mostly in AD and PD. Multimodal studies combining multiple data types are relatively limited (n=105). This could either indicate that publicly available datasets, despite their large volume of subjects, do not consistently provide sufficient multimodal information for a significant number of patients or, alternatively, that there is potential for a more effective exploitation of these datasets.
In contrast, studies based on original datasets (Figure 4b), defined here as data collected directly by the authors for the specific study, although fewer in number (n=36), exhibit greater diversity in both studied pathologies and data modalities. These studies more frequently include biomarker, wearables and EHRs data, with a more balanced distribution across diseases. They also show a higher frequency of multimodal combinations, indicating a more integrative approach to data collection.
Overall, the heterogeneity of data modalities used in clinical practice is only partially reflected in the selected studies, and is strongly influenced by the nature of the datasets employed. Large pre-existing datasets are predominantly imaging-focused, particularly in AD, whereas studies based on original datasets collected ad-hoc more frequently integrate non-imaging modalities, offering a

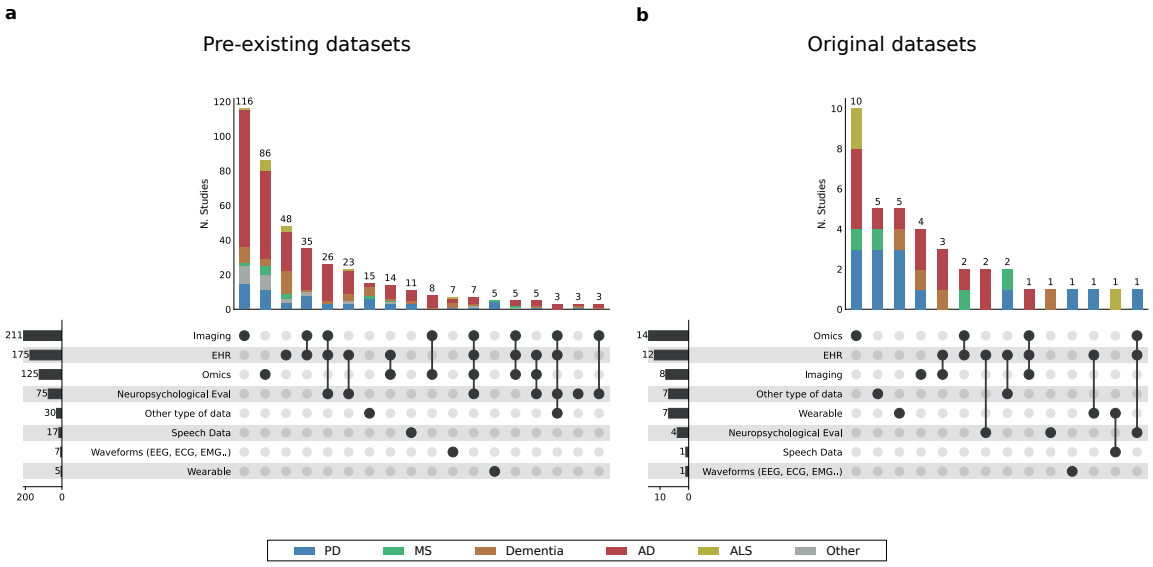

Figure 4: **Distribution of data modalities used in AI studies for neurodegenerative diseases.** UpSet plots show the data modalities and their combination, used in reproducible studies and stratified by disease. The left panel shows the studies using pre-existing datasets, while the right panel shows the studies that have collected data purposely for the study - original datasets. Only a small proportion of studies use multiple data modalities. AD: Alzheimer's Disease, PD: Parkinson's Disease, MS: Multiple Sclerosis, ALS: Amyotrophic Lateral Sclerosis

broader representation of real-world clinical practice.

AI modelling and evaluation

Accurate evaluation of AI models is crucial for their successful translation from research to clinical practice. This depends not only on developing effective algorithms, but also on selecting appropriate performance metrics, especially when dealing with challenges such as imbalanced datasets. To provide a clear overview of the current field, we focused on reproducible studies to understand which algorithms and corresponding performance metrics are most widely used and how they are applied. Since AI researchers often test multiple algorithms before selecting the optimal model, our analysis was divided into two parts: first, we considered all algorithms tested in each study and second, we focused only on those reported as the best performing.

When examining AI algorithms used across all reproducible studies (Figure 5a), several patterns emerge. Overall, the most frequently employed models are Support Vector Machines (SVM, 51.07%), Bagging Ensembles (39.67%), Logistic Regression (LR, 27.55%), Convolutional Neural Networks (CNN, 27.32%), Boosting Ensembles (23.75%), Deep Neural Networks (DNN, 17.34%), and k-Nearest Neighbors (kNN, 15.9%). A breakdown by pathology reveals that studies on AD, Dementia, and other neurodegenerative disorders exhibit similar trends, with SVM being the most commonly adopted algorithm, followed by Bagging Ensembles. PD studies follow a comparable pattern, with SVM and Bagging Ensembles used equally. In contrast, studies on MS and ALS demonstrate distinct algorithmic preferences. For MS, LR is most frequently employed, followed by Bagging Ensembles. For ALS, Bagging Ensembles are predominant, with SVM and CNN models used equally thereafter. However, given the limited number of studies addressing MS and ALS, these findings should be interpreted with caution. When considering exclusively top-performing models (Figure 5b), some shifts in both algorithm ranking and pathology-specific patterns emerge. The seven most frequently reported best-performing model types include SVM, CNN, Bagging Ensemble, Boosting Ensemble, LR, DNN, and attention- or transformer-based neural networks. For studies on AD and other neurodegenerative conditions, SVM is most often the top performing model, followed by CNN, and Bagging Ensembles. PD studies showed a similar distribution across these three architectures, while in Dementia studies SVM and Bagging Ensembles are equally represented as the leading models. Bearing in mind the previously mentioned caveat about studies numerosity, MS and ALS continue to display different trends. For MS, LR often achieve the highest performance, followed by Bagging Ensembles, ensemble models more generally, and CNNs. For ALS, Bagging Ensembles remain the predominant approach, with SVM and CNN models equally ranked as second-best. An interesting trend emerging from this comparison is that more recent and advanced AI techniques, such as attention-based methods, rank among the top-performing algorithms despite their limited adoption.

Moving to performance metrics, our review revealed that the most commonly reported metrics are: Area Under the Receiver Operating Characteristic Curve (AUROC, 65,80%), Accuracy (64,13%), Sensitivity (44,89%), Specificity (43,71%), F1-score (32,30%), Precision (26,84%), and Recall (19,71%) (Figure 5c). In studies focused on PD, AD, Dementia, and other neurodegenerative disorders, AUROC and Accuracy are the most frequently used metrics, each appearing in over 50% of cases, with a similar pattern across these pathologies. In contrast, studies on MS and ALS show a different trend: AUROC is markedly more prevalent, reported in more than 50% of studies, while Accuracy, Sensitivity, and Specificity are less common, appearing in only 16% to 33% of cases. These findings highlight a limited use of metrics capable of addressing class imbalance, such as balanced accuracy

or Matthews Correlation Coefficient (MCC), suggesting that a more fine-grained assessment of model errors may be necessary, and that caution should be exercised when interpreting the clinical impact of study results.

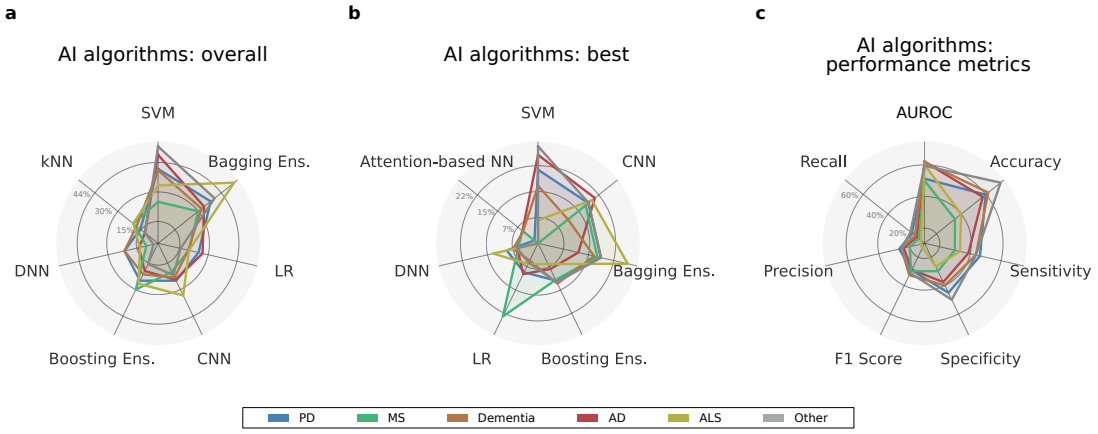

Figure 5: **Prevalence of AI algorithms and performance metrics in neurodegenerative AI research.** The charts show the distribution of the seven most common AI algorithms and the seven most common performance metrics across different neurodegenerative diseases research fields. Panels **a** and **b** show the prevalence of algorithms, with each spoke representing an algorithm and each coloured polygon corresponding to a specific pathology. Panel **a** illustrates the prevalence of all tested algorithms, while Panel **b** shows only the top-performing algorithms. Panel **c** displays the distribution of performance metrics, where each spoke represents a metric and each coloured polygon corresponds to a specific pathology. The values in all charts are normalised by the total number of algorithms or metrics reported for each disease, which allows for a direct comparison of methodological preferences under different reporting contexts. AD: Alzheimer's Disease, PD: Parkinson's Disease, MS: Multiple Sclerosis, ALS: Amyotrophic Lateral Sclerosis. SVM: Support Vector Machines, kNN: k-Nearest Neighbour, DNN: Deep Neural Network, Ens.: Ensemble, CNN: Convolutional Neural Networks, LR: Logistic Regression, AUROC: Area Under the Receiver Operating Characteristic Curve.

Towards clinical practice: reproducibility, generalisability and explainability of AI models

While good model performance represents a necessary condition for AI models to generate impact, it is not sufficient to ensure successful translation into clinical practice. Two key aspects are essential to move AI research closer to real-world clinical use.

The first, which applies to AI research broadly and is not limited to the clinical domain, is the reproducibility of the developed solutions. To provide a sufficiently comprehensive yet rigorous overview of the AI landscape in neurodegenerative diseases, we classified studies as reproducible where open data were used and code was made available for novel AI architectures, in line with previous works [7].

Among the studies meeting these criteria, the average reproducibility rate of the reviewed studies

remains low (35.5%), with only a slight upward trend from 2019 to 2024 (Figure 6a). Unsurprisingly, the availability of open access datasets strongly affects the reproducibility rate, although it does not appear to be the only relevant factor, as indicated by the breakdown across different pathologies (Figure 6b). In the case of AD, the high reproducibility rate aligns with the availability of numerous public datasets and a substantial volume of published research. Despite having a similar number of available datasets, Dementia shows a higher percentage of reproducible studies. A similar pattern is observed for other diseases and MS: the former has over 50% reproducible studies, while MS has the lowest reproducibility rate. Despite being the least represented in terms of publicly available datasets, ALS studies demonstrate a relatively high level of reproducibility.
We also examined the choice of traditional ML or DL methods as a potential additional factor influencing reproducibility. Rates were similar between the two groups (35.35% for traditional ML studies and 35.71% for DL studies), suggesting that the use of DL, in itself, is not strongly associated with reproducibility outcomes.
To further explore potential trends, we examined the journals that published the studies included in this review (Figure 6c). The five most frequently represented open-access journals are *Scientific Reports*, *Frontiers in Aging Neuroscience*, *PLOS One*, *Sensors*, and *IEEE Access*. Analysis of the journals' thematic focus revealed three main categories: Multidisciplinary journals such as *Scientific Reports* and *PLOS One* are well represented, alongside journals with a stronger emphasis on methodology and technological innovation, including *Sensors* and *IEEE Access*. The largest group, however, comprises journals specializing in neuroscience and neurodegenerative diseases, such as *Frontiers in Neuroscience*, *Alzheimer's Research & Therapy*, and *npj Parkinson's Disease*. When assessing reproducibility across the top 15 journals, no clear trend emerged. Reproducibility rates vary considerably, ranging from 50% to 13.8%, indicating that reproducibility is not strongly linked with journal type or specialization.
Finally, we examined whether reproducibility is associated with the country of origin of studies (Figure 6d). The top 15 countries by number of publications have reproducibility rates ranging from 47.4% and 46.2% in Australia and the United States (U.S.), to 16.2% and 15.8% in Japan and Taiwan. However, no clear pattern or consistent relationship between reproducibility and country of origin was observed.
Overall, these results suggest that, despite increasing awareness about the importance of reproducibility and journals' efforts to promote it by requesting open data and code for publications, reproducibility rates remain low. Improving these rates is particularly difficult in the clinical domain, where health data protection regulations introduce an additional layer of complexity.

The second critical aspect in translating AI models to clinical practice relates to the challenge of deploying them in settings or populations that differ from those used during development. From a technical standpoint, variations in the distribution of population characteristics can lead to performance drops, limiting applicability. A good practice to assess AI solution's ability to generalize under such conditions is the inclusion of external validation cohorts in the study design. Beyond technical considerations, the clinical domain presents additional challenges related to ethics and inclusion. To mitigate the risk of bias and inequality, developed solutions should be applicable across diverse populations. To assess how well this is addressed in AI research for neurodegenerative diseases, we investigated how diverse populations are represented in AI research, both for model development and validation.

As illustrated in Figure 7a, research efforts are highly concentrated in a few countries, with North America (primarily the U.S.) and China leading in terms of the number of studies. Europe also contributes significantly, particularly through countries such as the United Kingdom (U.K.),

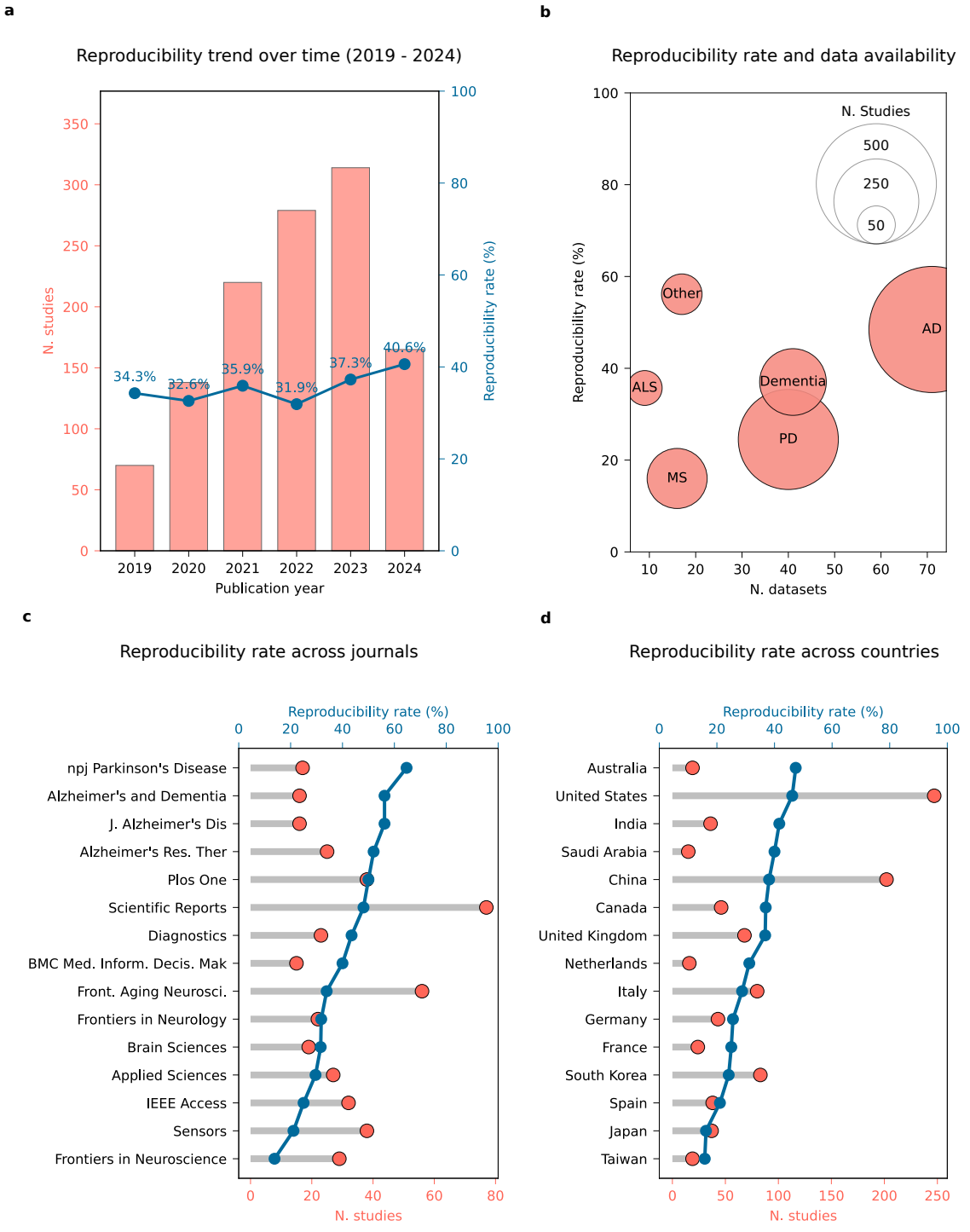

Figure 6: **Landscape of reproducibility of studies in AI methods for neurodegenerative diseases.** Panel **a** illustrates the trend of reproducibility from mid 2019 to mid 2024 (blue line), alongside a bar plot showing the total articles published each year. Panel **b** is a scatter plot that shows the relationship between reproducibility rate and data availability, as measured by the number of unique datasets per disease. Panel **c**, a dual-axis chart where the red lollipop markers are used to represent the distribution of published articles across the top 15 journals, while the blue line displays their corresponding reproducibility rates. Panel **d** uses a lollipop plot to show the total number of studies per country and the blue line plot to show the reproducibility rate. AD: Alzheimer's Disease, PD: Parkinson's Disease, MS: Multiple Sclerosis, ALS: Amyotrophic Lateral Sclerosis

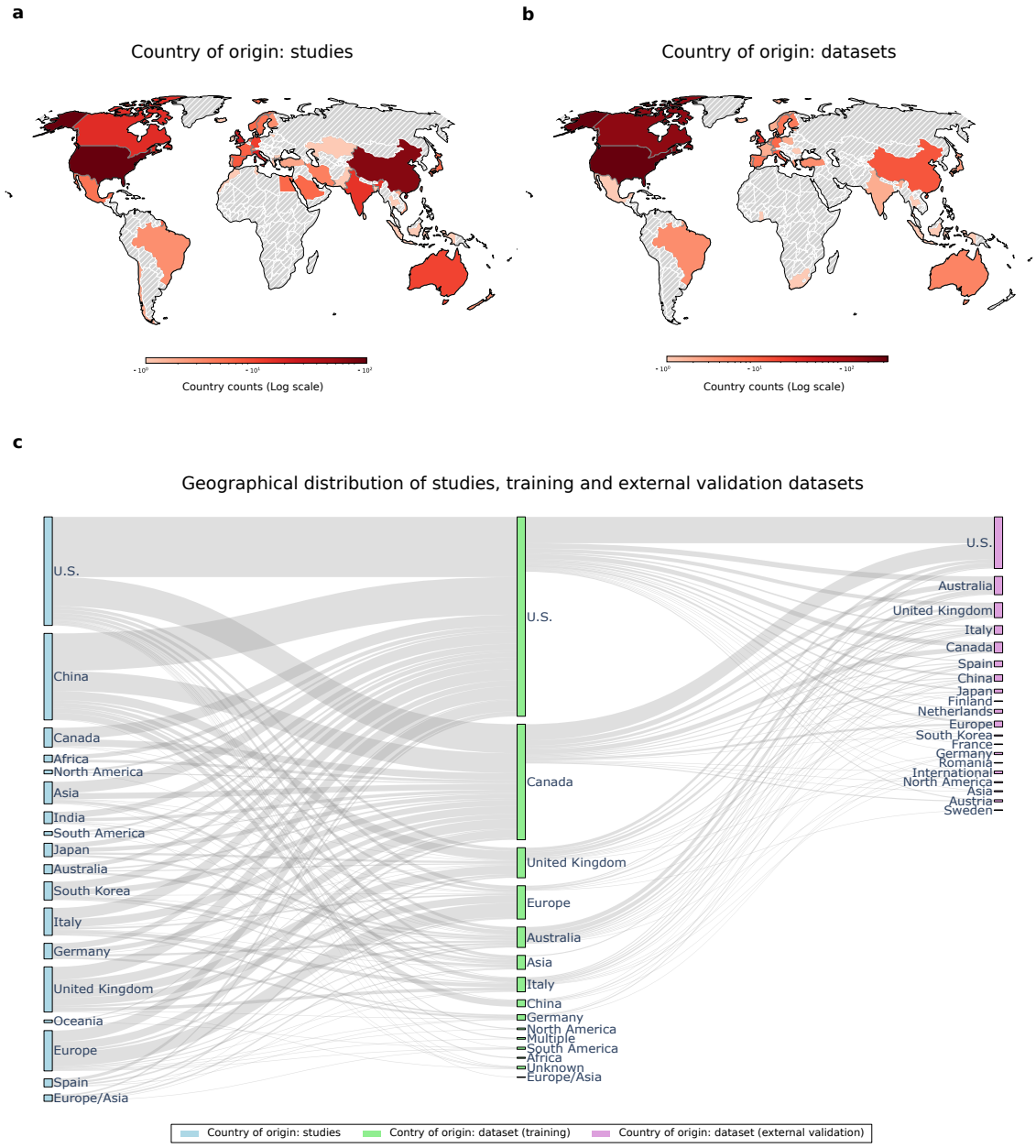

Figure 7: **Geographic bias and generalisability of AI studies for neurodegenerative diseases.** Panel **a** maps the countries of origin for all reviewed studies, while Panel **b** shows the countries from which datasets in reproducible studies were sourced. In both maps, each country was counted per appearance and displayed on a logarithmic scale to aid visualization. Panel **c** is a Sankey diagram tracing the flow from a study's country of origin to the countries of its training and external validation datasets, highlighting potential geographic disparities that may affect generalisability.

Germany, France, and Italy. When aggregated, European countries represent a substantial share of the global landscape (26.87%). Emerging contributions from the Middle East, notably Egypt, Saudi Arabia and Turkey, are beginning to appear; however, entire regions, including Africa, Southeast Asia (e.g., Indonesia), and large parts of Latin America, remain essentially absent.

These limitations in geographic diversity are further highlighted by the countries of origin of the datasets employed in these studies (Figure 7b). North American datasets, particularly large publicly available repositories such as ADNI and the Parkinson's Progression Markers Initiative (PPMI), dominate the field. European and Chinese datasets contribute to a lesser extent, but remain within the same narrow band of highly industrialized nations. There is virtually no dataset representation from low- and middle-income countries (LMICs), which raises concerns about generalizability and external validity of model predictions in these populations.

In addition, we evaluated the translational impact of the selected studies with respect to the use of external validation datasets. The first element of interest in examining the flow of the Sankey diagram (Figure 7c) is that across all reproducible studies, only a small proportion incorporates external validation in their workflow (75 studies, 17.81%). Most external validations are performed within the same geographic regions as the training datasets, with a particularly strong clustering around the U.S.. Studies originating in the U.S. often rely on U.S.-based datasets and validate models within U.S. populations. While some cross-national validation efforts are present (e.g. U.S. studies validating on data from U.K., and vice versa), these are still limited primarily to high-income countries with similar healthcare infrastructures. External validation on independent datasets representing diverse regions and population demographics, such as Africa, South Asia, or Latin America, is exceedingly rare.

Together, these findings reveal a clear limitation in the current research landscape in ensuring clinical translation and broad applicability of AI models across diverse patient populations. Most current research is based on datasets that reflect highly specific demographic and clinical characteristics, typically those of Western or East Asian cohorts with access to advanced imaging and diagnostic resources. As a result, models trained on these datasets risk poor performance in underrepresented populations, limiting their utility in real-world clinical settings.

Beyond dataset diversity and validation practices, another crucial element influencing the successful translation of AI models into clinical workflows is algorithmic explainability, i.e. the use of techniques aimed at making model decisions transparent and understandable to end-users. Across all reproducible screened studies, we found that only around half of them (46%) incorporated explainability methods. Interestingly, there appeared to be a relationship between explainability and the presence of an external validation dataset: 58.44% of studies with external validation also implemented some form of explainability, compared with 43.31% of studies without external validation (Figure 8). This might indicate that research placing stronger emphasis on robustness and generalizability may also be more likely to integrate elements of explainable AI. However, this trend must be interpreted with caution, given the relatively small number of studies with external validation. For all other examined variables, such as dataset type (pre-existing vs. original), AI methods (traditional ML vs. DL), modality (unimodal vs. multimodal), or country of origin, no meaningful associations with explainability were observed. In studies incorporating explainability, traditional ML models predominantly utilized feature importance measures and SHapley Additive exPlanations [18]. SHAP is a technique that uses game theory to fairly attribute each variable's contribution to a ML model's prediction, revealing how variables collectively influence the outcome. These findings highlight the need to expand the use of explainability methods to enhance interpretability: without transparent insights into the decision-making process, clinicians may struggle to trust or

appropriately interpret AI outputs, which could hinder adoption, especially in high-stakes diagnostic settings. Moreover, systematic integration of explainability methods would also enable broader model validation across different clinical environments, helping to uncover and address potential biases and ultimately improving the generalizability of AI solutions to real-world practice.

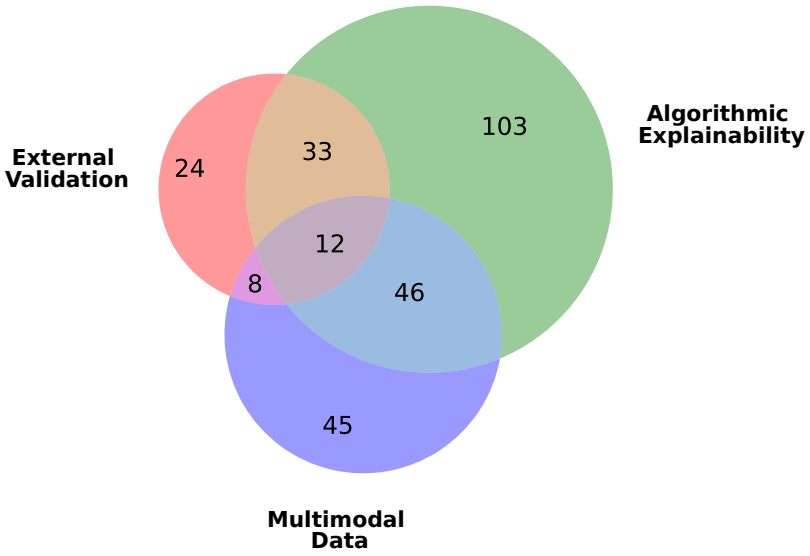

Figure 8: **Reproducibility, generalisability, and explainability in AI studies for neurodegenerative diseases.** Venn diagram visualising the distribution of studies with external validation (red), algorithmic explainability (green), and multimodal data (blue). The non-overlapping areas represent studies addressing only one aspect, while overlapping regions indicate studies incorporating multiple aspects.

Consistency of findings in high-influence studies

All previously described analysis were replicated in the subset of most influential studies (n=50). Overall, results in this cohort did not meaningfully differ from those observed in the full body of literature. Rather, these highly cited papers largely operate within the field's dominant paradigms: they prioritize diagnosis over treatment, rely almost exclusively on Western or high-income datasets, and largely adopt established architectures (e.g. CNNs and ensembles methods) rather than introducing radically new methodological paradigms. A detailed breakdown of results for the 50 top-cited researches is provided in the Supplementary Results.

Discussion

This systematic review that included 1186 articles provides a detailed overview and critique of AI methods used for a wide spectrum of neurodegenerative diseases. Our work provides a global perspective on these diseases, allowing for a comparison of trends and an assessment of how various factors — such as disease characteristics, biomarker modalities, data availability, AI methods, and

study reproducibility — influence research outcomes. This paper provides a timely synthesis on the technological advancements in AI applied to neurodegeneration and potential shortcomings found in the research literature.

Overall, the prevalence of diseases represented in the reviewed studies aligns with findings from the Global Burden of Diseases study [26]. AD and other dementias are the most prevalent, with 694 cases per 100,000 people, followed by PD (139 per 100,000), MS (30 per 100,000), and ALS (9.9 per 100,000). We observed similar trends for research examining AI applications in neurodegenerative conditions. This may highlight priority disease areas and neurological conditions that may benefit from further research using ML and DL methodology. Across all examined diseases, the most common research objective was diagnosis. Given the retrospective nature of the majority of studies, this was relatively unsurprising. This finding also supports growing interest in utilising AI as a tool to drive 'clinical decision support systems', focused on maximising accuracy of neurodegenerative diseases diagnoses [31]. This is followed by prognosis and patient stratification. Survival analysis is the least studied objective, likely because it may require longitudinal follow-up data, which is not always available or very limited in retrospective datasets. Given its vast potential clinical utility - including patient outcome prediction and treatment guidance [12] - further efforts to interrogate the usefulness of AI for survival analysis may be warranted.

Surprisingly, fewer than 5% of studies controlled for or integrated pharmacological treatments in their AI models. This means the vast majority of research relied solely on non-pharmacological clinical variables, which is a significant issue. Since treatments can substantially alter disease trajectory, excluding them could potentially compromise the robustness of predictive models. This gap may be due to publicly available datasets often lacking comprehensive pharmacological histories, a common limitation even in purely clinical settings, where accurate medication records are difficult to obtain [11]. However, with the increasing focus on personalised medicine, this omission represents a missed opportunity for AI to have a more direct and meaningful impact on patient care and to improve its translational potential. In accordance, a previous review demonstrated the ability of AI to aid in drug discovery and clinical trial design, but specifically in Dementia [8]. However, these authors note that open science practices may be limited in this field due to proprietary ownership of drug development pipelines, particularly by pharmaceutical companies.

Clinical decision-making related to diagnosis, prognosis and treatment of neurodegenerative diseases commonly involve a wide range of data, including imaging, biomarkers, omics, and neuropsychological evaluations. Despite this, our analysis shows that most studies overwhelmingly concentrate on imaging data for predictive modeling, with other modalities being underutilised. While imaging is a valuable data source and particularly well-suited for DL applications [1], this narrow focus fails to provide a holistic view of the patient's condition. This limitation may be a consequence of large-scale pre-existing datasets lacking sufficient multimodal data, and a potential over-reliance on neuroimaging. However, this finding highlight an important opportunity to improve existing or create new datasets that more accurately reflect the complexity of neurodegenerative diseases.

We also observed that studies using original datasets collected independently by the authors, while fewer in number, tend to incorporate a more diverse range of non-imaging data, more accurately representing real-world clinical practice. A further reason for the limited use of such data may lie in the methodological challenges associated with integrating diverse data types. As multimodal methods continue to advance, including development of foundation models, datasets must evolve accordingly to enable a comprehensive view of the full spectrum of disease characteristics, from omics to neuropsychological state, thereby supporting the development of more robust models. In this context, foundation models may partially alleviate current constraints by enabling the learn-

ing of generalizable representations from large-scale unlabeled data, which can then be adapted to multiple downstream tasks with relatively limited annotation. Recent work in brain MRI provides early support for this direction, showing improved performance in low-data and few-shot settings, as well as greater robustness across prediction and segmentation tasks [27]. However, these models remain at an early stage: training is still largely modality-specific, integration with non-imaging clinical variables is not yet mature, and their real-world value will depend on external validation and successful incorporation into clinical workflows. Thus, foundation models are perhaps best understood, at present, as a promising research framework rather than an already established route to clinical translation. These data limitations may also reflect shortcomings in the practical implementation and efficiency of research data management systems [25].

Strong model performance is a prerequisite for AI to achieve clinical impact, but it is not sufficient to ensure successful translation into practice. Our analysis show that reproducibility - a critical factor in this process - remains limited, with only 35.5% of studies meeting this standard. This figure has not substantially improved for the past five years (2019-2024), which is the time period considered by our systematic review. While the availability of public datasets is a contributing factor, it may not be not the only one. For instance, studies on ALS show relatively high reproducibility despite having the fewest available datasets, whereas MS studies have the lowest reproducibility rates. This suggests that other, non-data-related factors are at play, which may include features unique to differential neurodegenerative classifications. Additionally, the journals' target audience, whether specialised or general, does not appear to have a strong correlation with reproducibility, with rates varying widely from 14% to 50%.

Our findings also indicate that the studies reviewed studies did not consistently use appropriate evaluation metrics. In particular, overreliance on AUROC, typical in medicine, biostatistics and healthcare in general, may in some context provide an overly optimistic assessment of performance [4]. This issue becomes more pronounced in studies with a markedly uneven distribution of observations across categories, where more sensitive metrics such as balanced accuracy or the MCC are useful in dealing with potential bias [5]. Moreover, strong discrimination does not guarantee that a model's predicted probabilities reflect the true likelihood of the outcome of interest. A well-calibrated model provides clinically meaningful probabilities rather than only binary labels, however only 26 studies used calibration methods. This suggests a need for a more detailed analysis of model errors and warrants caution when interpreting the clinical significance of these findings.

The current body of published studies is highly concentrated geographically. The U.S. and China produce the highest number of studies, followed by European countries — specifically the U.K., Germany, France, and Italy. While emerging contributions are noted from Middle Eastern countries like Egypt, Saudi Arabia, and Turkey, it is concerning that entire regions, including Africa, Southeast Asia, and large parts of Latin America, are absent from this research landscape. This geographic imbalance is also reflected in publicly available datasets, which are dominated by North American, European, and Chinese sources. Crucially, no datasets from LMICs were identified. As a result, current research relies on datasets reflecting narrow demographic and clinical characteristics, most often those of Western or East Asian cohorts with access to advanced imaging and diagnostic resources. This represents a major concern for the generalisability and validity of AI models when applied to under-represented populations. Unmeasured factors such as ethnicity, geographic diversity, and socioeconomic conditions may influence model robustness. Generalisability of developed AI solutions might be further limited by current external validation practices. For example, studies originating from the U.S. frequently use U.S.-based datasets for external validation. Cross-national validations are infrequent and largely confined to other high-income countries, thereby excluding

more demographically diverse regions, such as Africa, South Asia, or Latin America. These gaps may also reflect limitations of our review, as we included only English-language, open-access studies - an approach that may have limited the identification of datasets from LMICs. Moreover, differences in national regulatory frameworks could further shape the extent to which AI solutions are translated into clinical practice.

Recommendations

Synthesising the evidence from this review, we outline three key recommendations for the application of AI in neurodegenerative research and clinical practice.

Improving data quality and inclusivity. Studies should integrate pharmacological treatment information into their models where relevant and available, as interventions can significantly alter disease trajectories and impact predictive robustness - though the applicability of treatment-aware modelling depends heavily on the condition, disease stage, and the effectiveness and availability of interventions. Researchers should also move beyond imaging alone and incorporate broader data modalities such as biomarkers, omics, and neuropsychological evaluations to capture a more holistic view of the patient's condition; existing and future public datasets must be deliberately designed to reflect this multimodal breadth alongside a sufficiently large and demographically diverse patient base.

Enhancing methodological rigour and reproducibility. To provide a more accurate and robust evaluation of model performance - particularly with imbalanced datasets - researchers should employ metrics such as balanced accuracy or the Matthews correlation coefficient (MCC) alongside standard measures. Improving reproducibility, which currently remains low across the field, is a fundamental prerequisite for clinical translation. The adoption of explainability methods is equally important, as interpretable insights into model decision-making and errors are essential for clinical implementation.

Broaden research scope and impact. There is a critical need to develop and publicly release datasets from low- and middle-income countries (LMICs) and under-represented regions such as Africa, Southeast Asia, and Latin America, to address geographic, ethnic, and socioeconomic disparities that limit model generalisability. Complementing this, studies should conduct external validation on independent datasets from diverse demographics and regions to confirm the validity and broader applicability of their models.

Conclusion

AI research in neurodegeneration is currently hindered by three systemic bottlenecks: poor reproducibility, limited data inclusivity, and a lack of clinical translatability. Many models rely on proprietary code and siloed datasets, making independent verification nearly impossible. Furthermore, a reliance on narrow, Western-centric demographics has the potential to create biased algorithms that fail diverse global populations. We proposed a set of recommendations, whose adoption will push forward AI from the academic bench into a reliable, equitable tool with genuine clinical utility.

Data Availability Statement

The data supporting the findings of this study are available within the paper and its Supplementary Information. The analysis code, the complete list of included studies, and all raw and processed data have been deposited in a GitHub repository (https://github.com/endrwalter/ai-neurodegen-systematic-review-materials).

Funding Declaration

This work was partially funded under the National Plan for Complementary Investments to the NRRP, project "D34H—Digital Driven Diagnostics, prognostics and therapeutics for sustainable Health care" (project code: PNC0000001), Spoke 2: "Multilayer platform to support the generation of the Patients' Digital Twin", CUP: B53C22006170001, funded by the Italian Ministry of University and Research.
A.C. is funded by the National Institute for Health and Care Research (NIHR; ref: NIHR203373).

Author contributions statement

V.O. conceived and supervised the study. W.E., F.R., S.B., M.M. contributed to the methodology; W.E., F.R., S.B., M.M. A.C. and V.O. conducted the formal analyses; W.E., F.R., S.B., M.M. wrote the initial draft of the manuscript; W.E., F.R., A.C. and V.O. reviewed the final version of the manuscript; F.R., M.M. and W.E. designed and curated plots and images; G.J. and V.O. provided scientific guidance and oversight. All authors reviewed, edited, and approved the final manuscript.

Competing Interests

The authors declare no competing interests.

References

[1] Robin J Borchert, Tiago Azevedo, AmanPreet Badhwar, Jose Bernal, Matthew Betts, Rose Bruffaerts, Michael C Burkhart, Ilse Dewachter, Helena M Gellersen, Audrey Low, et al. Artificial intelligence for diagnostic and prognostic neuroimaging in dementia: A systematic review. *Alzheimer's & Dementia*, 19(12):5885–5904, 2023.

[2] Adrian Burton. How do we fix the shortage of neurologists? *The Lancet Neurology*, 17(6): 502–503, 2018.

[3] Simiao Chen, Zhong Cao, Arindam Nandi, Nathaniel Counts, Lirui Jiao, Klaus Prettner, Michael Kuhn, Benjamin Seligman, Daniel Tortorice, Daniel Vigo, et al. The global macroeconomic burden of alzheimer's disease and other dementias: estimates and projections for 152 countries or territories. *The Lancet Global Health*, 12(9):e1534–e1543, 2024.

[4] Davide Chicco and Giuseppe Jurman. The matthews correlation coefficient (mcc) should replace the roc auc as the standard metric for assessing binary classification. *BioData Mining*, 16(1):4, 2023.

[5] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14(1):13, 2021.

[6] Sofia De la Fuente Garcia, Craig W Ritchie, and Saturnino Luz. Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review. *Journal of Alzheimer's Disease*, 78(4):1547–1574, 2020.

[7] Abhyuday Desai, Mohamed Abdelhamid, and Nakul R Padalkar. What is reproducibility in artificial intelligence and machine learning research? *AI Magazine*, 46(2):e70004, 2025.

[8] Thomas Doherty, Zhi Yao, Ahmad Al Khleifat, Hanz Tantiangco, Stefano Tamburin, Chris Albertyn, Lokendra Thakur, David J Llewellyn, Neil P Oxtoby, Ilianna Lourida, et al. Artificial intelligence for dementia drug discovery and trials optimization. *Alzheimer's & Dementia*, 19 (12):5922–5933, 2023.

[9] Valery L Feigin, Theo Vos, Emma Nichols, Mayowa O Owolabi, William M Carroll, Martin Dichgans, Günther Deuschl, Priya Parmar, Michael Brainin, and Christopher Murray. The global burden of neurological disorders: translating evidence into policy. *The Lancet Neurology*, 19(3):255–265, 2020.

[10] Corporation for Digital Scholarship. Zotero. Computer software, 2024. URL https://www.zotero.org/. Accessed: 2024-08-06.

[11] Henry Halapy and Heather Kertland. Ascertaining problems with medication histories. *The Canadian journal of hospital pharmacy*, 65(5):360, 2012.

[12] Junyong In and Dong Kyu Lee. Survival analysis: part ii-applied clinical data analysis. *Korean J Anesthesiol*, 72(5):441–457, 2019.

[13] Fariha Khaliq, Jane Oberhauser, Debia Wakhloo, and Sameehan Mahajani. Decoding degeneration: the implementation of machine learning for clinical detection of neurodegenerative disorders. *Neural Regeneration Research*, 18(6):1235–1242, 2023.

[14] Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Shannon Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023. URL https://api.semanticscholar.org/CorpusID:256194545.

[15] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *NPJ digital medicine*, 5(1):171, 2022.

[16] Paula E Lester, TS Dharmarajan, and Eleanor Weinstein. The looming geriatrician shortage: ramifications and solutions. *Journal of aging and health*, 32(9):1052–1062, 2020.

[17] Yuru Li, Xiaowei Chang, Jianlin Wu, Yuchen Liu, Hailu Wang, and Yiyin Zhang. Machine learning in early diagnosis of neurological diseases: Advancing accuracy and overcoming challenges. *Brain Network Disorders*, 2025.

[18] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[19] Maria Chiara Malaguti, Lorenzo Gios, and Giuseppe Jurman. The third wheel or the game changer? how ai could team up with neurologists in parkinson's care. *Parkinsonism & Related Disorders*, 134:107797, 2025.

[20] Jordi Martorell-Marugan, Marco Chierici, Sara Bandres-Ciga, Giuseppe Jurman, and Pedro Carmona-Saez. Machine learning applications in the study of parkinson's disease: A systematic review. *Current Bioinformatics*, 18(7):576–586, 2023.

[21] Atta Naqvi, GBD 2021 Europe Life Expectancy Collaborators, et al. Changing life expectancy in european countries 1990-2021: a subanalysis of causes and risk factors from the global burden of disease study 2021. *The Lancet Public Health*, 10(3):e172–e188, 2025.

[22] Matthew J Page, Joanne E McKenzie, Patrick M Bossuyt, Isabelle Boutron, Tammy C Hoffmann, Cynthia D Mulrow, Larissa Shamseer, Jennifer M Tetzlaff, Elie A Akl, Sue E Brennan, Roger Chou, Julie Glanville, Jeremy M Grimshaw, Asbjørn Hróbjartsson, Manoj M Lalu, Tianjing Li, Elizabeth W Loder, Evan Mayo-Wilson, Steve McDonald, Luke A McGuinness, Lesley A Stewart, James Thomas, Andrea C Tricco, Vivian A Welch, Penny Whiting, and David Moher. The prisma 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*, 372, 2021. doi: 10.1136/bmj.n71. URL https://www.bmj.com/content/372/bmj.n71.

[23] Gabriel Pardo and David E Jones. The sequence of disease-modifying therapies in relapsing multiple sclerosis: safety and immunologic considerations. *Journal of neurology*, 264(12):2351–2374, 2017.

[24] Jason Priem, Heather Piwowar, and Richard Orr. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts, 2022. URL https://arxiv.org/abs/2205.01833.

[25] Anna Schweinar, Franziska Wagner, Carsten Klingner, Sven Festag, Cord Spreckelsen, and Stefan Brodoehl. Simplifying multimodal clinical research data management: Introducing an integrated and user-friendly database concept. *Applied Clinical Informatics*, 15(02):234–249, 2024.

[26] Jaimie D Steinmetz, Katrin Maria Seeher, Nicoline Schiess, Emma Nichols, Bochen Cao, Chiara Servili, Vanessa Cavallera, Ewerton Cousin, Hailey Hagins, Madeline E Moberg, et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the global burden of disease study 2021. *The Lancet Neurology*, 23(4):344–381, 2024.

[27] Divyanshu Tak, Biniam A Garomsa, Anna Zapaishchykova, Tafadzwa L Chaunzwa, Juan Carlos Climent Pardo, Zezhong Ye, John Zielke, Yashwanth Ravipati, Suraj Pai, Sri Vajapeyam, et al. A generalizable foundation model for analysis of human brain mri. *Nature Neuroscience*, pages 1–12, 2026.

[28] Vimbi Viswan, Noushath Shaffi, Mufti Mahmud, Karthikeyan Subramanian, and Faizal Hajamohideen. Explainable artificial intelligence in alzheimer's disease classification: A systematic review. *Cognitive Computation*, 16(1):1–44, 2024.

[29] Chonghua Xue, Sahana S Kowshik, Diala Lteif, Shreyas Puducheri, Varuna H Jasodanand, Olivia T Zhou, Anika S Walia, Osman B Guney, J Diana Zhang, Serena Poésy, et al. Ai-based differential diagnosis of dementia etiologies on multimodal data. *Nature Medicine*, 30(10):2977–2989, 2024.

[30] Salim Yılmaz, Canser Boz, Salih Haluk Özsarı, Faruk Yılmaz, Buse Fidan Türkön, and Anı Hande Mete. Effects of neurological disorders on health expenditure and economic output: Dynamic panel analysis for oecd countries. *Systems*, 13(7):521, 2025.

[31] Yabin Zhang, Lei Yu, Yuting Lv, Tiantian Yang, and Qi Guo. Artificial intelligence in neurodegenerative diseases research: a bibliometric analysis since 2000. *Frontiers in Neurology*, 16:1607924, 2025.