Benefits, pitfalls, and future design of population-based registers in neurodegenerative disease

ABSTRACT

Population-based disease registers identify and characterize all cases of disease, including those that might otherwise be neglected. Prospective population-based registers in neurodegeneration are necessary to provide comprehensive data on the whole phenotypic spectrum and can guide planning of health services. With the exception of the rare disease amyotrophic lateral sclerosis, few complete population-based registers exist for neurodegenerative conditions. Incomplete ascertainment, limitations and uncertainty in diagnostic categorization, and failure to recognize sources of bias reduce the accuracy and usefulness of many registers. Common biases include population stratification, the use of prevalent rather than incident cases in earlier years, changes in disease understanding and diagnostic criteria, and changing demographics over time. Future registers are at risk of funding shortfalls and changes to privacy legislation. Notwithstanding, as heterogeneities of clinical phenotype and disease pathogenesis are increasingly recognized in the neurodegenerations, well-designed longitudinal population-based disease registers will be an essential requirement to complete clinical understanding of neurodegenerative diseases.

GLOSSARY

AD = Alzheimer disease; ALS = amyotrophic lateral sclerosis; EHR = electronic health record; FTD = frontotemporal dementia; PD = Parkinson disease.

The extent of heterogeneity in clinical phenotype and disease pathogenesis is increasingly recognized in the neurodegenerations, and population-based disease registers can provide high-quality longitudinal data that enable comparative analysis of demographics, phenotype, and outcome across different geographic regions.

Registers differ from other types of patient-related data such as electronic health records (EHRs), routine statistics held within national health services, and surveillance based on representative samples (e.g., the US National Health and Nutrition Examination Survey). Register data are actively collected (core data), and require rigorous standardized management. By contrast, EHRs are collected at the time of patient encounter and are coded based on the diagnosis at that time. While such records can be useful as source datasets in the construction of registers, EHRs alone are rarely sufficiently accurate or complete to enable the detailed surveillance associated with disease registers, as exemplified recently by comparison of an administrative database with a population-based register for amyotrophic lateral sclerosis (ALS) in France.

Restricted quality registers for neurodegeneration have been established in some regions, such as SveDem, an Internet-based quality registry that includes patients from memory clinics and primary care units with an estimated capture rate of 36%. Similarly, a French Dementia Register comprises 84% of memory clinics. These registers are not designed to recruit within the entire population, and while valuable in defining the characteristics of those enrolled, they
are by definition limited by their design and cannot provide an entire perspective of all cases with dementia.

Most population-based registers require a number of years of activity to ensure that early biases are resolved. Once fully functional, they can facilitate detailed analyses of disease heterogeneity, risk, health planning, and monitoring of disease burden in the population.6–14

The purpose of this review is to analyze the methods underpinning existing population-based registers for neurodegeneration, to provide insights into hidden biases that can confound analysis of register-based data, and to suggest methods to minimize errors of interpretation.

**EXISTING POPULATION-BASED REGISTERS IN NEURODEGENERATION** Few true population-based registers of adult neurodegenerative disease currently exist. Examples of successful registers include those developed for ALS6–8 and frontotemporal dementia (FTD).9–11 The European ALS registers can be viewed as gold standards for rare neurodegenerative diseases, as they have provided important insights into disease pathogenesis, clinical heterogeneity, and spatial distribution.6,12–14

Conversely, a relatively low number of true population-based registers for commoner conditions such as Parkinson disease (PD) and Alzheimer disease (AD) have been established, primarily in Scandinavia, the Netherlands, and some regions of the United States.9,13,16 The Nebraska Parkinson’s Disease Register17 has 98% ascertainment and uses capture-recapture analysis,15,18,19 which combines information from multiple ascertainment sources to estimate the number of cases that are likely to have been missed. Logistics of collection are aided through the support of state law, and the register ascertains from multiple sources with a verification process for all reported cases.15,17 This register is feasible in part due to the low population of Nebraska (1.8 million), which allows identification of all cases in a relatively labor-intensive manner. Given the high prevalence of PD in most countries (329.3/100,000 in Nebraska15), the costs of directly applying these methods in more populous regions would be considerable.

As demonstrated by the Nebraska register, registers require active ascertainment from multiple sources to exclude possible coding errors, and to identify those within the population who may not engage with specialist services (figure). Exclusive reliance on passive capture through existing records risks ascertainment bias. For example, an approach for PD in Thailand (population 67 million) using 3

-independent administrative sources achieved only 66% case ascertainment.20

Registers from Scandinavia,21 Canada,22 the United States,23 and Taiwan24 have ascertained cases using administrative EHRs. As exemplified by a prescription-based register for AD in Finland requiring stringent proof of diagnosis, such an approach can successfully exclude coding errors if administrative records are a priori designed to ensure reliability in ascertainment. However, use of EHRs that focus exclusively on drug prescription may also introduce other types of selection biases, as it is unlikely that all patients with neurodegenerative disease will use targeted medications.25 Even in countries with well-developed social health systems such as Finland, additional sources of ascertainment are required.

In general, studies that have ascertained cases only from specialist clinics, such as behavioral or old age psychiatry clinics for AD, are also likely to be biased,26 as referral patterns can vary based on age, income levels, and educational status (figure). There is evidence that specialist clinics are likely to exclude very early and very advanced cases,26–29 and those with a very aggressive course, as exemplified by data from the Irish ALS Register (table 1). Extrapolation from these types of datasets is limited by selective participation and attrition, particularly in dementia studies. As cohorts are more likely to be younger, interpretation of datasets from specialist memory clinics cannot be generalizable to the entirety of people with dementia, the majority of whom within the population are over 85 years of age.30 Because of this, biomarkers and prognostic indicators that are inferred from analysis of cohorts cannot be applied to the community at large. This high likelihood of a false-positive effect has been shown for mild cognitive impairment.29

**ALTERNATIVE APPROACHES TO POPULATION-BASED REGISTERS: POPULATION-BASED COHORT STUDIES** Given the known logistical challenges associated with full ascertainment of patients with common neurodegenerative disease, other methods based on complete follow-up of an entire or a randomly selected population in a small area have been used. All residents within a specific age range are regularly screened for the disease and exposures of interest, and are followed up regardless of the presence or absence of diagnosis. Examples of these types of studies include the Framingham Heart Study,31 the Swedish National Study on Aging and Care,32,33 and the Rotterdam Study in Holland. As exemplified by the Rotterdam study,33–35 all residents within a specific region and age category (over 55) are invited to participate in a longitudinal study of multiple health outcomes. These types of studies repeatedly evaluate the participating subject cohorts over many years, and
statistical assessment of associations between risk factors and disease incidence is determined. A major strength is that members of the population are regularly examined, including those who have not fully engaged with health services. New cases are identified at each repeated screening/visit that occurs within the target region. This differs from the traditional concept of a disease register, as the parent study is usually not restricted to a single diagnostic criterion for inclusion, but instead typically includes a range of diseases across specialties. However, by capturing information about diverse diseases within and across specialties, this method can be of great utility in collecting a true population-based estimate of common but heterogeneous conditions such as dementia or parkinsonism. It also has the advantage that it can capture ongoing changes in diagnostic criteria and can determine whether revised diagnoses predict outcomes. Another advantage of this approach is that the inclusion of healthy people within the cohort allows the rigorous assessment of both premorbid exposures and symptoms. This is valuable because the pathogenesis of some forms of neurodegeneration (e.g., dementia and PD) is known to be multifactorial, and utilization of a design that continuously samples and reassesses a population regardless of their diagnosis can facilitate large longitudinal studies that provide more accurate data pertaining to both clinical phenotype and risk. This approach has a further advantage of enabling recognition and evaluation of clinical conditions (comorbidities) and other changes that typically occur in elderly populations.

However, methods that depend on regular assessment of specified populations also have important limitations. They are not feasible for rare neurodegenerative conditions as it would be impossible to sample a sufficiently large population to enable detection of a large number of cases. Moreover, all cohorts are vulnerable to selective enrollment and attrition. For example, cognitive impairment strongly predicts study dropout, potentially leading to substantial underestimates of dementia incidence. Extremely fast-progressing cases may never be identified in cohorts, because a patient may be diagnosed and die during the interval between 2 assessment waves. This can result in misleading estimates of disease incidence.

Population-based registers must aim to reliably and consistently identify patients from all pathways, requiring government and institutional support, sufficient funding and manpower, and local expertise. Social, demographic, and clinical selection biases can influence the characteristics of patients following each pathway, and this can vary by disease and by health care system. As a result, research studies recruiting from single pathways will inherit these biases. Functioning population-based registers overcome these biases via inclusion of all pathways. Adapted from The Lancet (Brayne C, Davis D. Making Alzheimer’s and dementia research fit for populations. Lancet 2012;380:1441–1443), © 2012, with permission from Elsevier.
ESSENTIAL CONSIDERATIONS FOR NEW REGISTERS

Core content. Population-based registers in neurodegenerative diseases recognize the importance of defining in advance and collecting core content for all cases that fulfill the diagnostic criteria for ascertainment. Standardization permits accurate comparative analyses of different geographic populations and data pooling to enhance statistical power. The NIH/National Institute of Neurological Disorders and Stroke in the United States has recently defined detailed common data elements for a range of neurologic diseases, and the European Platform for Rare Disease Registries in Europe has proposed a more general set of common data elements for use in rare disease registries. A requirement for excessive detail increases participant burden and the likelihood of missing data, and for this reason extensive clinical data (such as would be required for care pathways) are best generated as additional nested studies associated with but not part of the core functioning of a prospective longitudinal register.

Recognizing ascertainment bias. The principal advantage of the population-based register is the minimization of sources of bias. Although elimination of all forms of ascertainment bias is not possible in the design and establishment of any new register, a recognition of the likely sources of bias is important. Factors that can assist the reduction of ascertainment biases...
bias, as does a stable population structure with limited migration patterns. Prospective design42 and longitudinal follow-up with the intention of complete ascertainment of all affected individuals can reduce known bias, as does a stable population structure with limited mobility, which minimizes the risk of bias associated with loss of follow-up.

As noted, no single data source is likely to identify all cases in population-based registers, and multiple sources of identification can improve reliability of ascertainment and ensure that the range and extent of disease heterogeneity is represented.7,42 The use of capture recapture analysis6,15,18,19 based on 2 or more independent source datasets permits evaluation of the degree of completeness of ascertainment, so that the population burden of cases can be estimated even if no source is complete.

Effective case identification strategies include referrals through networks of clinical professionals, death certification, and face-to-face or telephone-based interviews. Pitfalls that introduce bias include the use of single sources of ascertainment. These include collection of patients exclusively through specialist clinics as noted (tables 1 and 3), self-reporting of patients through online portals, or exclusive reliance on codes from billing data, which are subject to selection bias and diagnostic error.3,43,44 Death certification may also overestimate or underestimate disease incidence because of inaccurate recording and coding.45,46

### Bias of diagnostic uncertainty

Population-based registers require a high level of vigilance to exclude erroneous diagnoses and mimic syndromes in both rare and common diseases.47-49

Diagnostic certainty is problematic for both rare and common neurodegenerative diseases for different reasons. Rare diseases may not be recognized outside of specialist centers. Conversely, certainty of diagnosis of common diseases (e.g., dementias) can be limited by the absence of strict diagnostic criteria, and logistics surrounding the acquisition of sufficiently detailed clinical information to ensure clarity for inclusion purposes. This is a particular problem in the ascertainment of cognitive impairment that presages dementia. While memory clinics advocate regular screening, such an initiative is not practical within a population-based setting, and also risks the generation of false-positive diagnoses.28,29

### Table 2 Essential design aspects for population-based registers

| Register variables should be selected carefully and a core content paradigm should be agreed upon in advance. |
| Core content should include clearly defined case definitions and these should be applied within a framework of specifically stated inclusion and exclusion criteria. |
| Capture methodology should be clearly defined in advance. |
| Case identification pathways should be agreed upon in advance and should include multiple sources for identification of incident cases. |
| Core content should facilitate international collaborative efforts or national merger of data in large countries using multiple registers to cover different regions is advisable for rare diseases. |
| Identification of data manager and data controller and clear definition of governance structures is necessary. |
| Compliance with local data protection legislation and ethical constraints is required. |

### Table 3 Common sources of bias and their resolution

<table>
<thead>
<tr>
<th>Bias source</th>
<th>Bias implication</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ascertainment from sources other than population</td>
<td>Incomplete data, very young, very old, and very severe and very mild cases likely to be excluded; economically disadvantaged/minorities excluded</td>
<td>Multiple sources with capture recapture</td>
</tr>
<tr>
<td>Prevalent cases only or mixed incident and prevalent cases recruited</td>
<td>Milder cases, incomplete clinical phenotypes, and longer natural history</td>
<td>Longitudinal incident-based inclusion</td>
</tr>
<tr>
<td>Diagnostic certainty</td>
<td>Misdiagnoses and mimic syndromes included</td>
<td>Quality control, review of source documentation, autopsy reports; excluding the most recent 1-2 years of data capture (or longer if the condition has long lag from initial diagnosis to inclusion on the register)</td>
</tr>
<tr>
<td>Startup bias</td>
<td>Mixture of incident and prevalent cases included; clinical phenotype and natural history data skewed in favor of milder cases with better survival</td>
<td>Recognition, employment of careful statistical analyses of data collected in the first 3-5 years to account for startup bias</td>
</tr>
<tr>
<td>Information creep</td>
<td>Subtle changes in inclusion criteria; differences in clinical phenotype between historical cases and those recently enrolled</td>
<td>Awareness, rigorous use of core data elements; addition of new fields with additional data; correction for missing data from earlier cohorts</td>
</tr>
<tr>
<td>Period and cohort effects</td>
<td>Reasons for true differences between different cohorts misattributed</td>
<td>Awareness; mathematical correction where possible; rigorous attention and correction for missing data</td>
</tr>
</tbody>
</table>
Some neurodegenerations have low clinicopathologic correlation, and a recent meta-analysis on diagnosis of PD showed that 1 out of 5 diagnoses of is wrong. The low clinicopathologic concordance is not intrinsic to register data, but reflects real ambiguity and controversy in the clinical diagnoses, the frequent overlap between different types of neurodegeneration (e.g., AD, vascular dementia, and Lewy body dementia), and imperfect correspondence between clinical symptoms and neuropathologic bases of disease in some instances. Moreover, diagnoses in neurodegeneration may shift with changing symptoms during the course of the disease, and some differential diagnoses (e.g., corticobasal degeneration vs progressive supranuclear palsy) can only be made after specific symptoms develop, or even at the time of autopsy. These limitations, although difficult to address at the time of ascertainment, should be anticipated in the design of population-based neurodegenerative disease registers by incorporating ongoing collection of clinical data at regular intervals and by regular quality control assessment by review of source documents, correlation with death certificates, and where possible, autopsy reports.

**Bias of including mixed incident and prevalent cohorts.** Inclusion of both prevalent and incident cases in prospective disease registers introduces bias. Collection of incident cases generally facilitates the characterization of phenotype throughout the disease course, including variability in progression when ascertained longitudinally over an extended period of time. Conversely, inclusion of prevalent cases can bias characterization of the natural history of disease because prevalent cases will on average overrepresent more slowly progressing and less severe cases. This bias is most likely to occur when patients are ascertained from clinic cohorts, which generally cater for a larger proportion of prevalent rather than incident cases. Clinical trial cohorts such as the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) database are also subject to this type of bias. Population-based registers of incident cases that have been in operation for long periods of time catalogue all incident and prevalent cases and effectively eliminate this source of bias.

**Startup bias.** Data captured in the early years of a register are unlikely to be of the same quality as subsequently captured data, as there is generally a bias towards enrollment of prevalent rather than incident cases. Also, most registers miss some cases in the early years. Such issues are inevitable in the initiation years of any disease register regardless of how well the methodology of capture of cases is planned. This creates a substantial bias that is rarely recognized, but that must be accounted for in later analyses. As the register develops, the multisource case ascertainment mechanism becomes more streamlined, data are collected prospectively, and the target population becomes aware of the register, reducing rates of missing values. Therefore, startup bias must be considered in interrogating registers of recent origin, especially when evaluating time to event outcomes (e.g., with survival analysis). This bias also applies to analysis of data from recent case ascertainment in longstanding registers, as there is usually a time lag between case ascertainment and quality control methods that ensure diagnostic accuracy. To minimize these types of bias, analyses should be conducted by excluding data acquired at the beginning of the register, and within a specified period (often the previous year) of recent case ascertainment.

**Changes of disease definition.** The changing understanding of different clinical diagnoses and their interrelationships presents challenges for registers. For example, there is now a recognition that the clinical condition commonly classified as dementia of the Alzheimer type is heterogeneous, with contributing vascular and inflammatory factors, and the value of detailed phenotype characterization is accordingly increasingly recognized. Similarly, ALS was traditionally considered to be a neurodegenerative disorder confined to the motor system. However, it is now recognized that cognitive and behavioral changes are intrinsic features of the condition. The overlap between ALS and FTD has been further recognized following the discovery of the C9orf72 hexanucleotide repeat expansion. Inclusion of cognitive and genetic status is now recognized as an important core element of ALS registers, necessitating adjustment of the existing core datasets, and effectively leading to a large volume of missing data for patients captured by ALS registers in the 1990s and early 2000s.

For these reasons, comparative analyses of register data collected over long periods of time can be problematic, both from the perspective of missing data and in relation to the effect that this new information has on the process of case ascertainment and inferences regarding risk. Moreover, although successful population-based registers may boast complete ascertainment, it is likely that the increased recognition of both comorbidities and subphenotypes can lead to subtle shifts in the types of patients who are included.

As is the case with diagnostic certainty, there is no mechanism to eliminate this type of bias. However, it is important to be aware that it exists, and that comparative analyses of data across time intervals will be subject to the bias of information creep.

In addition to the bias of evolving information and subtle changes in diagnostic categorizations, period
and cohort effects should also be recognized during comparative analyses across different time periods of an individual register, and across registers from different regions and of different durations. It cannot be assumed that populations ascertained by registers at different time intervals are the same, and these effects should be taken into account when performing comparative analyses. For these reasons, the use of historical controls in the evaluation of interventions introduces a bias that cannot be evaluated.

**CHALLENGES IN MANAGING LONGITUDINAL REGISTERS**

**Logistics and infrastructure.** Registers are extremely vulnerable to interruptions in funding. Disease registers are cost- and labor-intensive, and (unlike EHRs) require active management to ensure uniformity of ascertainment and ongoing quality control. The costs of establishing and maintaining a register will depend on (1) the size of the study population, (2) the type of disease (e.g., common or rare), and (3) the nature of the health service within which the register operates. In all cases, sustainable funding is challenging in the absence of a specific government/funding agency policy commitment, such as has occurred in the United States with the inception of the US ALS register, and in Denmark, where the is a strong historical tradition of supporting high-quality longitudinal registers with interlocking datasets that are open to interrogation. With these exceptions, disease registers continue to be viewed by many funding agencies as infrastructural resources rather than essential research instruments, and are excluded from funding as they do not of themselves generate research outputs. Indeed, in a recent survey of European rare disease registers, 23% of 202 registers were established with no specific source of funding. Unstable funding streams devalue the original infrastructure of registers and data should be codified to exclude personal identification. Clearly enunciated protocols should be in place to determine who can access the data, and the circumstances under which this access is permitted. Nevertheless, the goals of fostering access by qualified researchers and preventing re-identification are sometimes in competition. Recent efforts to enhance data-sharing throughout the health research community have fostered innovations in data security, and investment in adopting these protocols and updating as new technologies become available is important to ensure the confidentiality of participants while maximizing research applications.

**Ethics and data protection legislation.** The sustainability of registers in the longer term is challenged by ongoing limitations on the types of data that disease registers are likely to be permitted to record. Data protection legislation has been strengthened within the last 20 years, and inclusion of data regarding living patients without their expressed informed consent in the absence of specific legislation is now in breach of data protection laws in Europe. This has the potential to introduce bias.

These problems could be addressed by legislation providing a derogation for population-based registers from the full stricures of data protection legislation as in the case of the Nebraska Parkinson’s Disease Register, while maintaining strict control in the types of access permitted. This requires an understanding and recognition by the public of the important potential societal benefits of population-based epidemiologic research, and in particular the potential public health benefit of identifying and communicating de-identified data regarding regional variations in disease incidence, prevalence, and survival. This benefit, which is implicit in the case of notifiable infectious and communicable diseases, is juxtaposed with the right to privacy of the individual. While some countries created registers that are now core research tools (e.g., the US ALS registry, the Danish registers, and cancer registers in many countries), there remains a limited recognition within Europe of the public health and societal benefits of such an approach for most noncancer and noncommunicable disease from the dual perspectives of research and public health/policy. The issues of privacy and data protection must be clearly addressed in registers and data should be codified to exclude personal identification. Clearly enunciated protocols should be in place to determine who can access the data, and the circumstances under which this access is permitted. Nevertheless, the goals of fostering access by qualified researchers and preventing re-identification are sometimes in competition. Recent efforts to enhance data-sharing throughout the health research community have fostered innovations in data security, and investment in adopting these protocols and updating as new technologies become available is important to ensure the confidentiality of participants while maximizing research applications.

**DISCUSSION**

Acknowledgment, recognition, and rectification of the inherent biases within population-based datasets is necessary for accurate demographic and deep phenotypic characterization of neurodegenerative diseases. However, some biases within registers can never be fully avoided. Notwithstanding, appropriate interrogation of register data can build our understanding of clinical and biological heterogeneity of neurodegenerative diseases within and across different populations. Well-designed population-based registers are essential to the provision of accurate epidemiologic and phenotypic characterization that will drive understanding of the heterogeneous disease mechanisms underpinning rare and common neurodegenerative diseases.

**AUTHOR CONTRIBUTIONS**

O.H. conceived the study and contributed to the writing of first and subsequent drafts. J.P.K.R. wrote the first draft, performed the literature search, and prepared tables. K.T. prepared the figure and contributed to the writing of the first draft. C.B., M.M.G., and G.L. provided expert review of the first draft and contributed to subsequent drafts. All authors approved the final manuscript.

**STUDY FUNDING**

J.P.K.R., K.T., and O.H. are funded by the Health Research Board (Ireland).
DISCLOSURE
J. Rooney was funded by the Health Research Programme Clinical Fellowship Programme. C. Brayne reports no disclosures relevant to the manuscript. K. Tobin is funded by the HRB Interdisciplinary Capacity Enhancement Fellowship Programme. She reports no conflicts of interest. G. Logroscino serves as associate editor for the Neuropneumology Journal, Karger. He received founding from Ministry of Health, Ministry of University and Research, regione Puglia Special program for research. M. Glymour is funded by the NIH, the Robert Wood Johnson Foundation, and the University of California, San Francisco. She reports no conflicts of interest. O. Hardiman is funded by the Health Research Board Clinician Scientist Programme and Science Foundation Ireland (15/SPI/3244 and 16/ERC/3854). Prof. Hardiman has received speaking honoraria from Novartis, Biogen Idec, Sanofi Aventis, and Merck-Serono. She has been a member of advisory panels for Biogen Idec, Allergen, Ono Pharmaceuticals, Novartis, Cytokinetics, and Sanofi Aventis. She serves as Editor-in-Chief of Amyotrophic Lateral Sclerosis and Frontotemporal Dementia. Go to Neurology.org for full disclosures.

Received November 9, 2016. Accepted in final form March 24, 2017.

REFERENCES