Purpose of review
This article argues that the time is approaching for data-driven disease modelling to take centre stage in the study and management of neurodegenerative disease. The snowstorm of data now available to the clinician defies qualitative evaluation; the heterogeneity of data types complicates integration through traditional statistical methods; and the large datasets becoming available remain far from the big-data sizes necessary for fully data-driven machine-learning approaches. The recent emergence of data-driven disease progression models provides a balance between imposed knowledge of disease features and patterns learned from data. The resulting models are both predictive of disease progression in individual patients and informative in terms of revealing underlying biological patterns.

Recent findings
Largely inspired by observational models, data-driven disease progression models have emerged in the last few years as a feasible means for understanding the development of neurodegenerative diseases. These models have revealed insights into frontotemporal dementia, Huntington’s disease, multiple sclerosis, Parkinson’s disease and other conditions. For example, event-based models have revealed finer graded understanding of progression patterns; self-modelling regression and differential equation models have provided data-driven biomarker trajectories; spatiotemporal models have shown that brain shape changes, for example of the hippocampus, can occur before detectable neurodegeneration; and network models have provided some support for prion-like mechanistic hypotheses of disease propagation. The most mature results are in sporadic Alzheimer’s disease, in large part because of the availability of the Alzheimer’s disease neuroimaging initiative dataset. Results generally support the prevailing amyloid-led hypothetical model of Alzheimer’s disease, while revealing finer detail and insight into disease progression.

Summary
The emerging field of disease progression modelling provides a natural mechanism to integrate different kinds of information, for example from imaging, serum and cerebrospinal fluid markers and cognitive tests, to obtain new insights into progressive diseases. Such insights include fine-grained longitudinal patterns of neurodegeneration, from early stages, and the heterogeneity of these trajectories over the population. More pragmatically, such models enable finer precision in patient staging and stratification, prediction of progression rates and earlier and better identification of at-risk individuals. We argue that this will make disease progression modelling invaluable for recruitment and end-points in future clinical trials, potentially ameliorating the high failure rate in trials of, e.g., Alzheimer’s disease therapies. We review the state of the art in these techniques and discuss the future steps required to translate the ideas to front-line application.

Keywords
ageing, Alzheimer, computational, disease progression, generative, modelling, multiple sclerosis, neurodegeneration, neurological, prion, statistical
MRI provides various markers of neurodegenerative disease including regional abnormalities in structure, microstructure, function and metabolism. Changes in these features can be subtle and occur at different disease stages.

Structural MRI reveals changes in brain structure. Abnormalities in the volume of a region of the brain typically reflect neurodegeneration in later stages of disease, although shape changes may occur earlier [3]. Higher specificity to volume loss is attained from short-term longitudinal data (atrophy) than from cross-sectional volumetric comparisons across individuals. Brain volume loss has been used for subject selection and as an endpoint in clinical trials for Alzheimer’s disease [4].

Various quantitative or microstructural MRI techniques, for example using diffusion MRI [5,6], magnetization transfer [7], relaxometry [8], susceptibility imaging [9,10] or combinations thereof [11,12,13], can reveal changes at the cellular level that may precede macrostructural changes.

Diffusion MRI and functional MRI also provide insight into brain connectivity. Diffusion MRI, via tractography, enables the study of structural connectivity, which can be affected in neurodegenerative diseases [14**], and may act as a network supporting pathogen propagation [15,16]. Functional MRI based on blood-oxygen-level-dependent contrast imaging [17] is an indirect measure of neuronal activity based on blood flow. It can reveal regions active in cognitive tasks, which can be disrupted in disease [18]. Moreover, temporal correlations of this activity highlight functional connectivity and its disruption in disease [19,20]. Similar functional information can come from electroencephalography or magnetoencephalography [21].

PET uses a radioactive tracer bound to a metabolite to map out various processes across the brain. Notable pathogenic processes include glucose hypometabolism [22] and concentrations of β-amyloid [23] or tau [24**]. Abnormalities of each are common markers of neurodegenerative diseases.

**Cognitive and other clinical presentations**

Cognitive and other clinical presentations play a key role in differentiating between neurodegenerative diseases.
diseases. Tests that measure cognitive impairment are susceptible to learning effects [25] and are confounded by depression/anxiety [26]. In order to improve the sensitivity of cognitive assessments for early detection [27], cognitive composites have been developed [28,29].

**Cerebrospinal fluid**

Pathogenically abnormal levels of, for example, misfolded proteins (or neurofibrillary tangles) in the brain can be detected in cerebrospinal fluid (CSF) [30,31]. Changes are often detectable many years prior to the clinical presentation of symptoms. However, CSF markers suffer from considerable variability within and between laboratories and assays [31,32], require a highly invasive spinal tap and lack spatial specificity within the brain.

**Fixed variables**

Fixed-variables such as demographics, genetics, environmental or lifestyle factors are also influential on disease progression. For example, neurodegeneration differences exist between sexes [33], ApoE4 carriage is a genetic risk factor in Alzheimer’s disease [34] and various environmental factors may influence neurodegeneration [35].

**ADVANCED COMPUTATIONAL MODELS OF DISEASE PROGRESSION**

Data-driven disease models have emerged in recent years that provide uniquely rich disease signatures without relying on a-priori classification or staging of individuals. These models contrast with an earlier modelling paradigm (Fig. 1) including hypothetical models, such as those presented in [36–38], which sketch disease signatures like those in Fig. 1a but are

![Fixed-variables](image_url)

**FIGURE 1.** Old paradigm disease progression models. (a) It shows the hypothetical model of [36], which illustrates qualitative sigmoid evolution in AD of scalar biomarkers such as CSF Aβ level, cognitive test scores and hippocampal volume or atrophy. The lack of quantitative information prevents direct diagnostic usage. (b) It shows a traditional longitudinal model of AD atrophy [39] by binning individuals a-priori into ‘mild’, ‘moderate’ and ‘severe’ classes based on cognitive test scores. The model can potentially match new individuals to the same stages using imaging data, but must exclude cognitive scores to avoid circularity. AD, Alzheimer’s disease.
not informed directly by measured data; traditional longitudinal models, for example [39–41] (Fig. 1b), which regress measured data against a prespecified variable that defines the disease stage; and pattern recognition approaches [42–44], which learn to classify cases from a labelled training database. Hypothetical models aid understanding, but lack the quantitative nature required for classification or prediction. Traditional models are inherently limited by, typically crude, a-priori staging and can model only variables that do not contribute to the stage assessment. Pattern recognition approaches are similarly limited by the a-priori labelling and provide no explicit disease signature to enhance understanding.

**Discriminative models**

Supervised machine-learning approaches use labelled data to discriminate between individuals who will/will not progress clinically, or do/do not have a disease [42,43,44]. They have shown the value of including both neuroimaging and non-imaging markers [44]. Straightforward classification approaches do not provide insight on the full disease time course, although many supervised machine-learning techniques also have the capability for regression so might be used to predict variables indicative of disease progression. However, such an approach still suffers from inherent limitations of the a-priori labelling. Unsupervised learning techniques, or clustering, offer potential in disease subtyping and staging [45,46,47]. Direct application of these ideas simply identifies groups of patients with similar appearance, but does not account for the temporal nature of progressive diseases where a single phenotype may appear very different at the beginning to the end of the full disease time course.

The recent rise of deep learning techniques potentially benefits discriminative models of disease progression, but such approaches typically need large amounts of data for both training and validation. Current medical data sets rarely approximate the data sets of billions of images that deep learning has used to show great success in, for example, computer vision tasks such as in [48], but the technology nevertheless may prove beneficial, particularly by exploiting transfer learning [49].

**Generative models**

Advanced generative models of neurodegenerative disease progression avoid the need for a-priori staging to construct a quantitative signature from all available biomarkers. They fall broadly into two categories: models for unstructured data and models for structured data.

**Models for unstructured data**

These models work with a vector of scalar biomarkers with no assumption of a particular structure or relationship among them. They are designed to fuse together information from the constellation of multimodal biomarkers typically available in studies of neurodegenerative diseases.

The event-based model (EBM) [50–52] (Fig. 2) learns an ordering and uncertainty in the ordering, in which a set of biomarkers becomes abnormal directly from a measured data set, possibly entirely cross-sectional, and requires no predefined staging variable. EBM uses the idea that more individuals from a cohort containing a spectrum of disease stages will show abnormality in biomarkers that change early in the progression. The EBM has been used to study familial Alzheimer’s disease [50,52], Huntington’s disease [50] and sporadic Alzheimer’s disease [51,52].

Scalar trajectory models, derived from differential equation models [52–54] or self-modelling regression [55,56], reconstruct typical continuous trajectories of scalar biomarkers from a measured data set. As shown in Fig. 3, this gives a more complete temporal picture than the EBM’s sequence of abnormality. Current implementations of differential equation models [52–54] treat each biomarker independently, whereas self-modelling regression pools information across biomarkers to estimate trajectories on a common time frame. Model estimation requires no staging variable, although the data set must contain at least short-term longitudinal information to support estimates of the temporal scale of the trajectories.

**Models for structured data**

These models work with structured data such as images, geometric meshes or networks in which data values have a well-defined spatial organization.

Spatiotemporal models [57,58,59] construct a four-dimensional picture of how an image or image region typically evolves in shape and/or appearance during a disease or developmental process. Traditional longitudinal models of this type are straightforward to construct by binning individuals according to a predefined variable, such as age or a cognitive test score, and computing statistics within each bin. Most current models [58,59] still rely on a-priori labelling. However, recent advances [57] avoid any predefined stage variable by simultaneously estimating a central image trajectory, a deformation of each time point of each individual to align with the central trajectory and a ‘timewarp’ for
each individual and time point that defines the position along the central trajectory.

Network propagation models [60–64] attempt to explain disease progression in terms of spatial propagation through anatomical or functional networks in the brain, which thus induce measurement changes sequentially along the nodes of a network. The pattern of propagation depends on, and can thus reveal, mechanisms of propagation. In particular, Zhou et al. [62] translate various popular

FIGURE 2. Event-based models of disease progression. (a) An EBM of familial AD from [49]. The set of biomarkers (vertical axis) includes atrophy rates in each cortical region. The positional variance diagram is a visualization of the uncertainty in the ordering: each row is a grayscale histogram of probabilities that the event occurs in each position of the ordered sequence. Groups of interchangeable events appear as large blocks (e.g. later in the progression, lower right), whereas strong confidence in the ordering of events appears as a thin diagonal (e.g. earlier in the progression, upper left). (b) An EBM of sporadic AD learned from the ADNI data set in [50] with more diverse biomarkers: CSF tau and Ab; cognitive test scores; regional atrophy and volumes from MRI. The model defines a staging system based on the set of biomarkers showing abnormality. (c) Stage assignments of cognitively normal, mild cognitive impaired (MCI) and AD individuals showing classification accuracy similar to state-of-the-art pattern recognition approaches [44], but with an explicit generative model. AD, Alzheimer’s disease; ADNI, Alzheimer’s disease neuroimaging initiative; CSF, cerebrospinal fluid; EBM, event-based model.

FIGURE 3. The temporally continuous self-modelling regression approach of [55]. The model shows the characteristic trajectories of a diverse set of biomarkers against a common continuous disease stage variable learned from the ADNI and PAQUID (Personnes Agees Quid) data sets. The model can potentially estimate the disease stage of a new patient by identifying the position along the trajectory set that best matches their data.
models for neurodegenerative disease propagation: ‘nodal stress’ – the most active regions or ‘hubs’ are most vulnerable [65,66]; ‘transneuronal spread’ – a prion-like agent propagates along network connections [15,16,67,68]; ‘trophic failure’ – opposite to ‘nodal stress’, the most isolated nodes are most vulnerable [69,70]; and ‘shared vulnerability’ – evenly distributed vulnerability over networks. Each model predicts different patterns of atrophy severity within networks, enabling comparison of the different propagation hypotheses in various neurological diseases. Dynamical system models [61,63] support the prion-like propagation hypothesis in a similar way. More recently, a multifactorial network model was proposed [64] that considers multiple brain networks and their interactions.

Although spatiotemporal models express the trajectory of much richer objects than scalar trajectory models, currently they typically model just one object (image or image region) derived from a single modality. Current formulations do extend naturally to vector-valued images, or multiple interacting regions, but the methods are memory and processor intensive so such extensions present practical difficulties. Network propagation models are typically used to make a single gross inference of propagation mechanism rather than a detailed disease signature so are somewhat distinct from the other models above and do not directly support diagnosis, staging or prognosis.

Current limitations
One of the most common limitations in disease progression models is the assumption of a common disease trajectory across individuals. Neurodegenerative diseases are highly heterogeneous and clearly violate this assumption. However, it does enable the models to elicit an average or canonical trajectory, which may often be sufficient to capture broad disease characteristics and discriminate different conditions. Various models naturally provide measures and visualizations of the uncertainty in that canonical trajectory, see for example the positional variance diagrams of the EBM in Fig. 2, although these cannot be interpreted directly as pictures of population heterogeneity. Various enhancements aim to estimate features of the population heterogeneity. For example, [71] estimates the distribution of the EBM ranking over Alzheimer’s disease neuroimaging initiative individuals by replacing the single ordering with a Mallows model [72]; [47**] identifies distinct components of Alzheimer’s disease atrophy; and [73**] uses a mixture of EBMs to identify subgroups of the population with distinct orderings. These ideas go some way toward ameliorating this limitation, but further work is required.

Another limitation in disease progression models is the assumption of biomarker independence. This may be valid for early vs. late biomarkers, but it is unlikely to be true for all biomarkers across the full disease time course. Some scalar trajectory models, in particular [55,56], avoid this assumption by explicitly pooling information across biomarkers in order to estimate disease progression. Recent spatially fine-grained models of disease accumulation across the brain use weak assumptions on the spatial correlation structure [74], or explicitly seek clusters with a correlated atrophy pattern [75]. The work in [75] further highlights, particularly for complex models such as spatiotemporal shape models of the brain, the need to balance parsimony and information content. This complexity control is necessary to make inference tractable in large data sets, while retaining the critical information about disease progression.

To date, little effort has gone into incorporating information from fixed variables into disease progression models. Simple approaches, such as constructing separate models for particular genotypes, for example as in [51], highlight that progression does depend on such information. However, lower level integration of such variables into the models may produce interesting new findings.

Other limitations arise from idiosyncrasies of medical data sets. Longitudinal data censoring because of participant dropout can lead to undersampling of later disease stages and, if not accounted for, can bias models towards earlier disease stages. Measurement error can bias even simple models, for example regression dilution, and so explicitly quantifying and expressing uncertainty within disease models is important [54]. Outliers can bias models and frequently arise, for example because of unrelated disease such as a brain tumour causing exaggerated brain volumes, or erroneous data entry. Missing data are also common and, although disease progression models often have formulations that handle missing data in theory [73**], the effect on downstream inference remains to be quantified precisely.

FUTURE AND EMERGING TRENDS
Enhancing existing models
The combination of phenomenological models and mechanistic models offers powerful opportunities to compare hypotheses on disease propagation. Early explorations of mechanistic models [61,62] compare with only end-stage atrophy patterns. However, such models can predict the whole time course of
neurodegeneration. Matching just the overall atrophy pattern is analogous to guessing the story of a movie from only the final scene. The temporal disease signatures uniquely provided by longitudinal data-driven disease progression models, such as the EBM or continuous trajectory models, open up the possibility of matching the whole pattern of appearance of disease — viewing the whole neurodegeneration movie as it unfolds for much stronger inferences on underlying mechanisms.

Recent advances include automatic determination of disease subtypes via data-driven mixture modelling [73]. The use of subtype models to identify the distinct phenotypes among patients provides fundamentally new knowledge that may reveal new subgroups in diseases that are known to be highly heterogeneous, or highlight common comorbidities. This has the potential to aid treatment development through identifying smaller and more homogeneous patient cohorts more likely to have coherent underlying biochemistry and disease, and thus to respond consistently to a single treatment. The subtype models further offer the radical possibility of challenging the traditional labels of dementia subtypes (Alzheimer’s disease, dementia with Lewy bodies, frontotemporal dementia, posterior cortical atrophy and so on.) for the first time. Data-driven subtypes may prove more homogeneous and inherently easier to recognize from clinical data than the traditional labels. They may thus prove simpler targets for drug development thereby supporting precision medicine more readily, by matching individuals to treatments more precisely.

Enhancing these models to include causal inference will be an important advance. A recent multifactorial model [64] incorporated linear systems control theory into a model with causal structure to suggest necessary features of putative treatments that could have disease-modifying effects.

The complementary information provided by each type of disease progression model mentioned above can be combined into uniquely rich, quantitative signatures of disease that avoid the need for a-priori staging. This could be achieved through a single unified modelling framework. For example, there are natural links from the EBM to scalar trajectory models, to spatiotemporal trajectory models. Such a unified modelling framework potentially supports rigorous inference from the collection of disparate information available, balancing evidence from the multiple modalities appropriately.

Translational requirements to realize the potential of disease progression modelling

Currently, disease progression models remain largely in the technical research domain — being developed by computer scientists and statisticians in collaboration with clinical experts. They have great potential for application in the clinic and in clinical trials, but likely require several stages of development and proof of concept to reach the necessary level of maturity.

We must find ways to demonstrate consistent levels of performance. This requires robust and interpretable metrics of predictive performance, which include evaluation against known ground truth in simulations and well-phenotyped data sets (e.g. familial cohorts). Larger, less well-phenotyped data sets support performance evaluation in terms of model stability (e.g. cross-validation and intercentre generalizability), temporal resolution and staging self-consistency.

Clear performance metrics and yardsticks will provide developers of disease models with a clear understanding of how model assumptions bias model predictions and how to interpret and temper those predictions in practice. We must also find ways to transmit that understanding to end users to avoid abuse through misunderstanding. This involves the development of clear user interfaces to make the model output accessible and digestible by nonexpert users, and clear instructions and case studies demonstrating how to use the technology in practical situations.

Thus, key steps that remain for translation are:

1. **External validation of predictive ability of developed models on research data sets.** Several works already show this, but the community still has a need for unbiased evaluation in entirely unseen test sets. Community wide challenges such as [76,77] provide a valuable resource in this way.

2. **Translation between data sets.** Key questions remain on how well within-dataset performance will translate to other data sets, that is if we train models on a data set with a particular acquisition protocol, how well can we expect it to perform on unseen data from different protocols or acquired in different centres? Little exploration of this has been performed.

3. **Demonstration in clinical trials.** Initial steps to demonstrate potential to improve outcome via retrospective analysis of clinical trial data would...
support more ambitious work in demonstrating the potential of multimodal modelling in a prospective clinical trial.

(4) Success in a small number of trials involving model developers can motivate adaptation of the technology for general usage without developer intervention and widespread deployment in treatment development. Such advances will require substantial efforts in interface design and user workflow modelling.

CONCLUSION

We have reviewed data-driven model-based analyses of neurodegenerative disease. We have argued the potential for generative data-driven models to take centre stage in the study and management of neurodegenerative diseases if we are to generate new avenues for disease understanding in the earliest, preclinical stages. This is necessitated by the challenges in monitoring any neurological disease over its full time course, coupled with overlapping phenotypes and lack of a single biomarker that is dynamic across the full disease time course.

The main focus of development and application to date has been in Alzheimer’s disease, but various efforts including the EuroPOND project are expanding the application to other dementias, multiple sclerosis, prion diseases, normal ageing and development, and even non-brain applications. These techniques have the potential for widespread impact in realising precision medicine across many such domains.

Acknowledgements

The authors acknowledge their colleagues in the EuroPOND consortium (http://www.europond.eu), who contributed indirectly to this review and continue to contribute to the field of neurodegenerative disease progression modelling. The EuroPOND coinvestigators are Daniel C. Alexander, Neil P. Octoby, Sebastien Ourselin, Jonathan M. Schott, Olga Ciccarelli, Stanley Durdleman, Olivier Colliot, Meike W. Vernooij, Wiro J. Niessen, Stefan Klein, Frederik Barkhof, Giovanni B. Frisoni, Alberto Bizi, Petra S. Hüppi and Wim Van Hecke.

Financial support and sponsorship

The authors receive funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No. 666992.

Conflicts of interest

The authors declare that they have no conflicts of interest.

REFERENCES AND RECOMMENDED READING

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest


Review of both the background theory of quantitative susceptibility mapping and current applications in neuroimaging.


A new hybrid compartmental model for investigating white matter at the subvoxel scale.


Important review article for this emerging topic in mechanistic disease models.


Practice effects are confounders in monitoring disease progression, particularly for clinical trials. This article focuses on this important issue and suggests ways around the problem.