Data science in neurodegenerative disease: its capabilities, limitations, and perspectives

Sepehr Golriz Khatami[a,b], Sarah Mubeen[a], and Martin Hofmann-Apitius[a,b]

Purpose of review

With the advancement of computational approaches and abundance of biomedical data, a broad range of neurodegenerative disease models have been developed. In this review, we argue that computational models can be both relevant and useful in neurodegenerative disease research and although the current established models have limitations in clinical practice, artificial intelligence has the potential to overcome deficiencies encountered by these models, which in turn can improve our understanding of disease.

Recent findings

In recent years, diverse computational approaches have been used to shed light on different aspects of neurodegenerative disease models. For example, linear and nonlinear mixed models, self-modeling regression, differential equation models, and event-based models have been applied to provide a better understanding of disease progression patterns and biomarker trajectories. Additionally, the Cox-regression technique, Bayesian network models, and deep-learning-based approaches have been used to predict the probability of future incidence of disease, whereas nonnegative matrix factorization, nonhierarchical cluster analysis, hierarchical agglomerative clustering, and deep-learning-based approaches have been employed to stratify patients based on their disease subtypes. Furthermore, the interpretation of neurodegenerative disease data is possible through knowledge-based models which use prior knowledge to complement data-driven analyses. These knowledge-based models can include pathway-centric approaches to establish pathways perturbed in a given condition, as well as disease-specific knowledge maps, which elucidate the mechanisms involved in a given disease. Collectively, these established models have revealed high granular details and insights into neurodegenerative disease models.

Summary

In conjunction with increasingly advanced computational approaches, a wide spectrum of neurodegenerative disease models, which can be broadly categorized into data-driven and knowledge-driven, have been developed. We review the state of the art data and knowledge-driven models and discuss the necessary steps which are vital to bring them into clinical application.

Keywords

Artificial intelligence, data-driven models, knowledge-driven models, neurodegenerative disease, virtual patient cohort

INTRODUCTION

With the silver tsunami (i.e., an aging population) sweeping across the world, neurodegenerative diseases (NDDs) are becoming endemic, placing a disproportionate level of burden on older adults (those aged 65 years or over) [1]. NDDs affect nearly 50 million people worldwide and roughly 10 million new cases are reported every year (https://www.who.int/news-room/fact-sheets/detail/dementia). To prevent the occurrence of these conditions, slow their progression, and reduce their global socioeconomic impact, a deeper understanding of the pathophysiology underlying these diseases is necessary.

[a]Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, Sankt Augustin and [b]Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

Correspondence to Sepehr Golriz Khatami, Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53754 Sankt Augustin, Germany. E-mail: sepehr.golriz.khatami@scai-extern.fraunhofer.de

Curr Opin Neurol 2020, 33:249–254

DOI:10.1097/WCO.0000000000000795

KEY POINTS

- Diverse knowledge and data-driven NDD models have been developed for various purposes.
- These models are limited in their clinical applications because of deficiencies with clinical studies.
- Knowledge-driven models can enhance the interpretive power of empirical, data-driven models by incorporating relevant background knowledge.
- Artificial intelligence has the potential to overcome deficiencies with clinical studies and the limited applicability of certain models by simulating virtual patients.

Understanding the cause of NDDs is challenging because of the complex nature of these diseases and the existence of dysregulations at different biological scales, ranging from mutations at the genetic level to structural and functional alterations of the brain at the clinical level. For this reason, a broad variety of biomarkers throughout all modalities, including imaging and nonimaging have been studied. However, effectively translating these extensive biomarker modalities into a clinical application remains a challenging task.

In recent years, computational approaches that analyze these biomarker modalities have led to a wide range of models that help to understand NDDs. Existing models can be placed in two primary categories namely, data-driven models and knowledge-driven models. Although data-driven models are informed directly by patient-level data, knowledge-driven models rely on reasoning over findings of previously published studies. Here, we highlight recent advancements of diverse data-driven models in the context of their applications and describe knowledge-driven models and their applications in NDD research. Finally, we propose the use of artificial intelligence in this field to overcome the limitations associated with clinical data upon which such models are built in order to generate new avenues for better disease comprehension.

Data-driven models

The multifaceted nature of NDDs demands quantification of a wide variety of biomarkers of all modality types, including imaging and nonimaging, such as cerebrospinal fluid samples and *omics* data. To translate these biomarker modalities into clinical application, they can be subjected to computational approaches independently (unimodal) or in combination (multimodal). Unimodal-based models overlook the complexity of a disease as they do not consider the interdependence between different biological modality measurements. Nonetheless, in NDD research, certain modalities, such as genetic [2,3] and neuro-imaging [4–6], are well-suited for unimodal biomarker analysis. Specifically, genetic biomarkers are used in these analyses as NDDs are partially driven by genetics [7] and the imaging modality is used as it can estimate pathological changes occurring in patients [8]. In contrast, multimodal-based models provide a more comprehensive overview of a disease by integrating a variety of biomedical and clinical biomarkers.

In the following, we review current developments in the field of multimodal-based models. We start with models that provide an overview of biomarker dynamics in the time course of disease which can facilitate disease diagnosis and patient staging and then highlight models which assist in patient prognosis. Finally, we conclude with models that are used for patient stratification.

Disease progression monitoring, diagnosing, and patient staging

Disease progression, or biomarker trajectory, and the current disease stage can be estimated by two primary models. The traditional type models the trajectory of biomarkers based on discrete disease stages, however, a finite number of stages fail to capture continuous changes related to disease progression over time [9–11]. Alternatively, in the contemporary type, disease progression is modeled based on measured biomarkers (e.g., mini mental state examination) [12] and thus, the disease time course is regarded as a continuous process. Although the traditional models were developed by reasoning over previously published studies, the more contemporary ones were developed using diverse computational approaches. These include linear and nonlinear mixed models (N/LMMs) [13–15], differential equation models (DEMs) [15], self-modeling regression models (SMORs) [16], and event-based models (EBMs) [17–19].

Although a diverse set of contemporary models exist, there are trade-offs between the techniques that have been used to develop them. For example, an N/LMM model [14] can make an assumption on the shape of biomarker trajectory (e.g., exponential curves), whereas a DEM model [15] and SMOR model [16] can loosen this assumption. In DEMs, each biomarker is treated independently, yet SMORs pool data from all available biomarkers to estimate the dynamics of biomarkers over the course of disease [20]. In contrast to SMORs and DEMs which provide continuous biomarker trajectories, EBMs provide a discrete description of the biomarkers dynamics [20,21]. This type of model does not include any time information between rate of biomarker changes, which limits its capability in disease monitoring [21]. However, in contrast to other models, EBMs can also address individual deviations from a generic disease progression model [22].

Patient prognosis

Risk models can provide prognostic information by predicting the probability and the time of future incidence of disease. Diverse computational approaches such as the Cox-regression technique [23–25], Bayesian network models [26], and deep-learning-based approaches [27,28] have been used to establish such models. There also exists a trade-off between current implementations of these approaches.

Although in the Cox-regression-based models, relationships among features are restricted by a number of assumptions and the causal structure cannot be modeled, Bayesian networks can model the underlying causal relationship between predictive risk variables [29■■]. This enables Bayesian networks to ask ‘what-if’ questions and predict risk at the individual-level as the effects in Bayesian networks are represented by directed arrows and thus, with any manipulation of the independent variable, the model can predict its influence on the dependent variable [30].

Furthermore, the prediction accuracies in current implementations of Bayesian networks and deep-learning-based approaches are notably higher than those for Cox-regression models as the former approaches are well suited for dealing with high dimensional data. In Cox-regression and Bayesian network-based models, feature selection is done manually or semiautomatically and thus relies upon prior knowledge from researchers. However, because deep learning algorithms can automatically infer features that can help to predict future incidence of disease, they perform better compared with the Cox-regression and Bayesian network-based models [31]. Moreover, the current implementation of deep-learning-based models are capable of accepting any irregular length of data as an input without preprocessing, in contrast to Cox-regression and Bayesian network-based models where a preprocessing step is required for handling unequal time-series and missing values.

Patient stratification

NDDs are highly heterogeneous diseases in terms of clinical and biological appearance and progression patterns. As such, stratification of patients based on disease subtypes may lead to improved disease management and the design of better treatments, which in turn brings us closer to the goal of precision medicine. To this end, diverse clustering approaches have been used, such as nonnegative matrix factorization [32], nonhierarchical cluster analysis (e.g., *k*-means clustering) [33,34], and hierarchical agglomerative clustering [35,36]. Although these methods can differentiate subtypes of patients, they are generally not suitable for longitudinal clinical data that often suffer from missing data or unequal time-series measurements because of patient dropout. This is because state-of-the-art distance measure methods (e.g., Euclidean) are unable to compute dis/similarity between samples with different longitudinal measurement lengths [37]. Moreover, Euclidean distance measures often ignore existing temporal correlations between the measurements. Recently, de Jong *et al.* [38■■] proposed an autoencoder-based method to cluster multivariate time-series with many missing values. Although the autoencoder-based model currently only works with equal length time series, it outperformed the clustering approaches that used state-of-the-art distance measures as well as those models which used distance measures specifically designed for unequal time series, such as Dynamic Time Warping.

Knowledge-based models

Knowledge-based models have been developed in parallel to data-driven approaches to facilitate the interpretation of empirical data with background knowledge. Such models have effectively garnered novel insights into several disease areas and have also led to new disease taxonomies. By classifying diseases through data and knowledge-based models, it is possible to establish an alternative approach to the current paradigm of disease classification by clinical appearance. This can facilitate the identification of disease subtypes and associated molecular processes and thereby, help to establish potential disease biomarkers and novel therapeutic targets [39].

NDDs are a particularly complex set of diseases, where short and direct causal links can be challenging to discern. However, NDDs have considerable genetic components and with an abundance of biomedical *omics* data generated from high-throughput technologies, several data-driven approaches can be used to gain insights on these multifaceted diseases. Nonetheless, these approaches tend to lack contextual information; for instance, the cumulative effect of several dysregulated genes with slight alterations can be greater than the effect of a single, highly altered gene. Conversely, knowledge-based models, such as pathway-centric approaches, can incorporate prior knowledge to point at relevant pathways or biological processes and possess greater explanatory power [40]. Thus, by taking into account pathway effects, pathway-centric approaches consider a condition in its broader biological context rather than elucidating specific, individual genes or molecular processes involved in that condition [41]. In the field of NDD research, various pathway analyses have reported significantly enriched pathways in specific NDDs, such as Alzheimer's disease [42–44], and Huntington's disease [45,46], as well as across two or more NDDs [47–49].

Though various pathway-centric approaches markedly improve the interpretive power of *omics* data, these approaches rely upon canonical pathways and representing disease context can be challenging [50]. Moreover, as NDDs tend to be complex and multifaceted, they may only be partially attributable to the involvement of a given number of pathways. As such, elucidating disease-specific mechanisms may be more appropriate for NDDs as compared with applying pathway-centric approaches [51▪▪]. Accordingly, disease-specific knowledge maps, resources which contain mechanisms that are specific to a particular disease, have been developed and can be used to build computational models of disease. Notably, the Disease Maps Project has collected disease maps for several diseases, including AlzPathway for Alzheimer's disease [52] and Parkinson's disease map for Parkinson's disease [53], whereas the NeuroMMSig knowledge graph [51▪▪] has collated candidate mechanisms for Alzheimer's disease, Parkinson's disease, and epilepsy.

Outlook and perspectives of using artificial intelligence in neurodegeneration research: virtual cohorts for data sharing and trial simulation

Although a wide spectrum of computational models has been established, current applications of these models in clinical practice are limited because of deficiencies that come with clinical studies, such as biases toward specific ethnicities, small sample sizes, data missingness, data heterogeneity, and data privacy. In the following, we first elaborate on the inherent challenges in using clinical data, then outline a promising solution to address these challenges and conclude with its potential applications in neurodegeneration research.

Ideally, clinical data should be collected in regular intervals for all patients. However, only a limited number of clinical studies have collected longitudinal measurements. Additionally, because of inclusion–exclusion criteria that cannot be avoided or the disproportionate representation of particular ethnicities due to geographic constraints, these studies tend to have biases [54]. Furthermore, most clinical datasets have a relatively small number of samples (fewer than a thousand patients) and a large number of missing observations [55–57]. As such, dealing with these deficiencies generally demands extensive preprocessing such as imputation or discarding of variables. Finally, different clinical studies in equivalent disease contexts usually do not measure the same clinical outcomes and/or molecular data. Nonetheless, even if measurements of identical outcomes and/or data are collected across studies, as their study protocols vary, the data coming from one study is often not interoperable (mappable) to data coming from another. Therefore, clinical data are highly heterogeneous [58]. Additionally, sharing patient data beyond an organization's firewalls is restricted because of legal and ethical constraints. Consequently, there exist data 'silos' which impede the required analyses and comparisons of multiple studies which are so vital to gain comprehensive overviews of a specific disease.

Although a broad range of solutions has been established to address these deficiencies, each of these has its own challenges. For example, imputing missing values can lead to errors and discarding variables that contain a high proportion of missing values can result in information loss and biased conclusions [59]. Similarly, although individual agreements between data users (e.g., research institutes) and data owners (e.g., Alzheimer's disease neuroimaging initiative) can provide access to the data, its usage is restricted to certain activities. For instance, use of the data may not be permitted for teaching and training purposes.

Such shortcomings with clinical data can be overcome with artificial intelligence, and machine learning approaches in particular. These approaches facilitate simulating a synthetic cohort (i.e., virtual cohort) which is informed by actual cohort data and can thus represent the fundamental characteristics of the real cohort [60]. Although this solution has a long history in physiological studies [61,62], and clinical trial simulation [63,64], their application in clinical studies where the focus is on simulating virtual patients across biological scales and modalities (e.g., nonimaging, imaging) is a more recent development. Recently, Gootjes-Dreesbach *et al.* [65▪▪] have developed a variational autoencoder modular Bayesian network which simulates heterogeneous clinical study data as a virtual patient cohort.

Not only do virtual cohorts provide new avenues toward sharing patient-level data without endangering the data privacy of real patients, they can also enable the generation of 'meta-cohorts' by combining the virtual patients obtained from different available clinical data in disease-specific contexts [63]. This 'meta-cohort' not only addresses the deficiency of small sample sizes in clinical studies, but also eliminates their inherent biases (e.g., underrepresentation of certain ethnicities). Moreover, virtual cohorts provide opportunities to conduct counterfactual or 'what-if' scenarios. For example, researchers can add a feature which has not been observed in a specific study (e.g., comorbidity) and investigate how it influences the disease of interest or what the distribution of a particular biomarker would be if a patient's age shifts a number of years [65■■]. Ultimately, virtual cohorts can improve the design of clinical trials and have the potential to bring us closer to the goal of precision medicine.

CONCLUSION

We have reviewed a wide range of established NDD models, from unimodal-based models to multimodal-based ones. We have shown that in contrast to data-driven approaches, knowledge-driven approaches can provide meaningful contextualization and insights into the pathophysiology of disease. We described the deficiencies and limitations of currently available clinical studies in the scope of NDDs and argued the potential of artificial intelligence to overcome these shortcomings so that it is possible to generate new avenues toward better comprehending neurodegenerative disease.

Acknowledgements

The authors thank Daniel Domingo-Fernández for critical reading of the manuscript.

Financial support and sponsorship

The authors receive funding from the IMI project AETIONOMY (https://www.aetionomy.eu/en/vision.html) within the Seventh Framework Program of the European Union under grant agreement No. 115568.

Conflicts of interest

There are 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

1. Heemels MT. Neurodegenerative diseases. Nature 2016; 539:179–180.
2. Zhang B, Gaiteri C, Bodea LG, *et al.* Integrated systems approach identifies genetic nodes and networks in late-onset Alzheimer's disease. Cell 2013; 153:707–720.
3. Huang X, Liu H, Li X, *et al.* Revealing Alzheimer's disease genes spectrum in the whole-genome by machine learning. BMC Neurol 2018; 18:5.
4. Gupta Y, Lee KH, Choi KY, *et al.* Alzheimer's disease diagnosis based on cortical and subcortical features. J Healthc Eng 2019; 2019:2492719.
5. Lama RK, Gwak J, Park JS, *et al.* Diagnosis of Alzheimer's disease based on structural MRI images using a regularized extreme learning machine and PCA features. J Healthc Eng 2017; 2017:1–11.
6. Manjón JV, Coupé P, Raniga P, *et al.* MRI white matter lesion segmentation using an ensemble of neural networks and overcomplete patch-based voting. Comput Med Imaging Graph 2018; 69:43–51.
7. Wang YA, Kammenga JE, Harvey SC. Genetic variation in neurodegenerative diseases and its accessibility in the model organism Caenorhabditis elegans. Hum Genomics 2017; 11:12.
8. Whitwell JL, Höglinger GU, Antonini A, *et al.* Radiological biomarkers for diagnosis in PSP: where are we and where do we need to be? Mov Disord 2017; 32:955–971.
9. Frisoni GB, Fox NC, Jack CR Jr, *et al.* The clinical use of structural MRI in Alzheimer disease. Nat Rev Neurol 2010; 6:67.
10. Jack CR Jr, Knopman DS, Jagust WJ, *et al.* Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. Lancet Neurol 2010; 9:119–128.
11. Jack CR Jr, Knopman DS, Jagust WJ, *et al.* Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. Lancet Neurol 2013; 12:207–216.
12. Jack CR, Vemuri P, Wiste HJ, *et al.* Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. Arch Neurol 2012; 69:856–867.
13. Yang E, Farnum M, Lobanov V, *et al.* Quantifying the pathophysiological timeline of Alzheimer's disease. J Alzheimer's Dis 2011; 26:745–753.
14. Delor I, Charoin JE, Gieschke R, *et al.* Modeling Alzheimer's disease progression using disease onset time and disease trajectory concepts applied to CDR-SoB scores from ADNI. CPT Pharmacometr Syst Pharmacol 2013; 2:1–10.
15. Villemagne VL, Burnham S, Bourgeat P, *et al.* Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. Lancet Neurol 2013; 12:357–367.
16. Donohue MC, Jacqmin-Gadda H, Le Goff M, *et al.* Estimating long-term multivariate progression from short-term data. Alzheimer's Dement 2014; 10:S400–S410.
17. Oxtoby NP, Young AL, Cash DM, *et al.* Data-driven models of dominantly-inherited Alzheimer's disease progression. Brain 2018; 141:1529–1544.
18. Young AL, Oxtoby NP, Daga P, *et al.* A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain 2014; 137:2564–2577.
19. Fonteijn HM, Modat M, Clarkson MJ, *et al.* An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage 2012; 60:1880–1889.
20. Oxtoby NP, Alexander DC. Imaging plus X: multimodal models of neurodegenerative disease. Curr Opin Neurol 2017; 30:371.
21. Young AL, Oxtoby NP, Schott JM, Alexander DC. Data-driven models of neurodegenerative disease. ACNR 2014; 14:6–9.
22. Young AL, Oxtoby NP, Huang J, *et al.* Multiple orderings of events in disease progression. Inf Process Med Imaging 2015; 24:711–722.
23. Li S, Okonkwo O, Albert M, *et al.* Variation in variables that predict progression from MCI to AD dementia over duration of follow-up. Am J Alzheimer's Dis 2013; 2:12.
24. Barnes DE, Cenzer IS, Yaffe K, *et al.* A point-based tool to predict conversion from mild cognitive impairment to probable Alzheimer's disease. Alzheimer's Dement 2014; 10:646–655.
25. Li K, Chan W, Doody RS, *et al.* Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. J Alzheimer's Dis 2017; 58:361–371.
26. Alexiou A, Mantzavinos VD, Greig NH, *et al.* A Bayesian model for the prediction and early diagnosis of Alzheimer's disease. Front Aging Neurosci 2017; 9:77.
27. Lee G, Nho K, Kang B, *et al.* Predicting Alzheimer's disease progression using multi-modal deep learning approach. Sci Rep 2019; 9:1952.
28. Fisher CK, Smith AM, Walsh JR, *et al.* Machine learning for comprehensive forecasting of Alzheimer's disease progression. Sci Rep 2019; 9:1–14.
29. Khanna S, Domingo-Fernández D, Iyappan A, *et al.* Using multi-scale genetic,
■■ NeuRoImaging and clinical data for predicting Alzheimer's disease and reconstruction of relevant biological mechanisms. Sci Rep 2018; 8:11173.
One of the first examples of a Bayesian network which provides causality inference.
30. Arora P, Boyne D, Slater JJ, *et al.* Bayesian networks for risk prediction using real-world data: a tool for precision medicine. Value Health 2019; 22:439–445.
31. Liang H, Sun X, Sun Y. Text feature extraction based on deep learning: a review. EURASIP J Wirel Commun Netw 2017; 2017:1–12.
32. Scheltens NM, Tijms BM, Koene T, *et al.* Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts. Alzheimer's Dement 2017; 13:1226–1236.
33. Erro R, Vitale C, Amboni M, *et al.* The heterogeneity of early Parkinson's disease: a cluster analysis on newly diagnosed untreated patients. PLoS One 2013; 8:e70244.
34. Mu J, Chaudhuri KR, Bielza C, *et al.* Parkinson's disease subtypes identified from cluster analysis of motor and non-motor symptoms. Front Aging Neurosci 2017; 9:301.
35. Nettiksimmons J, DeCarli C, Landau S, *et al.* Biological heterogeneity in ADNI amnestic mild cognitive impairment. Alzheimer's Dement 2014; 10:511–521.

36. Hwang J, Kim CM, Jeon S, *et al.* Prediction of Alzheimer's disease pathophysiology based on cortical thickness patterns. Alzheimer's Dement (Amst) 2016; 2:58–67.

37. Lauwers O, De Moor B. A time series distance measure for efficient clustering of input/output signals by their underlying dynamics. IEEE Control Syst Lett 2017; 1:286–291.

38. de Jong J, Emon MA, Wu P, *et al.* Deep learning for clustering of multivariate ■■ clinical patient trajectories with missing values. GigaScience 2019; 8:giz134.

Not many methods have been developed for the clustering of time series data and they often cannot handle missing data. This article developed a method which can not only handle missing data, but can also be useful for short time series data.

39. Saqi M, Lysenko A, Guo YK, *et al.* Navigating the disease landscape: knowledge representations for contextualizing molecular signatures. Brief Bioinform 2018; 20:609–623.

40. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol 2012; 8:e1002375.

41. Reimand J, Isserlin R, Voisin V, *et al.* Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. Nature protocols 2019; 14:482.

42. Hu YS, Xin J, Hu Y, *et al.* Analyzing the genes related to Alzheimer's disease via a network and pathway-based approach. Alzheimer's Res Ther 2017; 9:29.

43. Li X, Wang H, Long J, *et al.* Systematic analysis and biomarker study for Alzheimer's disease. Sci Rep 2018; 8:17394.

44. Patel H, Dobson RJ, Newhouse SJ. A meta-analysis of Alzheimer's disease brain transcriptomic data. J Alzheimer's Dis 2019; 68:1635–1656.

45. Agus F, Crespo D, Myers RH, *et al.* The caudate nucleus undergoes dramatic and unique transcriptional changes in human prodromal Huntington's disease brain. BMC Med Genomics 2019; 12:137.

46. Moss DJ, Flower MD, Lo KK, *et al.* Huntington's disease blood and brain show a common gene expression pattern and share an immune signature with Alzheimer's disease. Sci Rep 2017; 7:44849.

47. Arneson D, Zhang Y, Yang X, *et al.* Shared mechanisms among neurodegenerative diseases: from genetic factors to gene networks. J Genet 2018; 97:795–806.

48. Labadorf A, Choi SH, Myers RH. Evidence for a pan-neurodegenerative disease response in Huntington's and Parkinson's disease expression profiles. Front Mol Neurosci 2018; 10:430.

49. Li MD, Burns TC, Morgan AA, *et al.* Integrated multi-cohort transcriptional meta-analysis of neurodegenerative diseases. Acta Neuropathol Commun 2014; 2:93.

50. Mazein A, Ostaszewski M, Kuperstein I, *et al.* Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. NPJ Syst Biol Appl 2018; 4:21.

51. Domingo-Fernández D, Kodamullil AT, Iyappan A, *et al.* Multimodal mechan- ■■ istic signatures for neurodegenerative diseases (NeuroMMSig): a web server for mechanism enrichment. Bioinformatics 2017; 33:3679–3681.

52. Ogishima S, Mizuno S, Kikuchi M, *et al.* AlzPathway, an updated map of curated signaling pathways: towards deciphering Alzheimer's disease pathogenesis. In: Systems biology of Alzheimer's disease. New York, NY: Humana Press; 2016:; 423–432.

53. Fujita KA, Ostaszewski M, Matsuoka Y, *et al.* Integrating pathways of Parkinson's disease in a molecular interaction map. Mol Neurobiol 2014; 49:88–102.

54. Lawrence E, Vegvari C, Ower A, *et al.* A systematic review of longitudinal studies which measure Alzheimer's disease biomarkers. J Alzheimers Dis 2017; 59:1359–1379.

55. Mueller SG, Weiner MW, Thal LJ, *et al.* The Alzheimer's disease neuroimaging initiative. Neuroimaging Clin N Am 2015; 15:869–877; xi–xi10.

56. Vermunt L, Veal CD, Ter Meulen L, *et al.* European Prevention of Alzheimer's Dementia Registry: recruitment and pre screening approach for a longitudinal cohort and prevention trials. Alzheimers Dement 2018; 14:837–842.

57. Ibrahim JG, Chu H, Chen MH. Missing data in clinical studies: issues and methods. J Clin Oncol 2012; 30:3297.

58. Laake P, Haakon BB. Research in medical and biological sciences: From planning and preparation to grant application and publication. Academic Press; 2015.

59. Kang H. The prevention and handling of the missing data. Korean J Anesthesiol 2013; 64:402.

60. Chase JG, Desaive T, Preiser JC. Virtual patients and virtual cohorts: a new way to think about the design and implementation of personalized ICU treatments. In: Vincent JL, editor. Annual update in intensive care and emergency medicine 2016. Annual update in intensive care and emergency medicine. Cham: Springer; 2016.

61. Carson E, Cobelli C. Modeling methodology for physiology and medicine. Newnes 2013.

62. Keener JP, Sneyd J. Mathematical physiology (Vol. 1). New York: Springer; 1998.

63. Gomez-Cabrero D, Abugessaisa I, Maier D, *et al.* Data integration in the era of omics: current and future challenges. BMC Syst Biol 2014; 8(Suppl 2):I1.

64. Lim SS, Kivitz AJ, McKinnell D, *et al.* Simulating clinical trial visits yields patient insights into study design and recruitment. Patient Prefer Adherence 2017; 11:1295.

65. Gootjes-Dreesbach L, Sood M, Sahay A, *et al.* Variational Autoencoder ■■ Modular Bayesian Networks (VAMBN) for simulation of heterogeneous clinical study data. BioRxiv 2019; 760744.

One of the first examples of an outlook from AI and machine learning that facilitates simulating a synthetic cohort (i.e., virtual cohort) by using heterogeneous clinical study data.