Life.course digital T.wins – I.ntelligent M.onitoring for E.arly and continuous intervention and prevention (LifeTIME): Proposal for a proof-of-concept study

Milne-Ives, Madison; Fraser, Lorna K.; Khan, Asiya; Walker, David; van, Velthoven MH; May, Jon; Wolfe, Ingrid; Harding, Tracey; Meinert, Edward

Published in:
JMIR Research Protocols

DOI:
10.2196/35738

Publication date:
2022

Life.course digital T.wins – I.ntelligent M.onitoring for E.arly and continuous intervention and prevention (LifeTIME): Proposal for a proof-of-concept study

Abstract

Introduction: Multimorbidity, which is associated with significant negative outcomes for individuals and healthcare systems, is increasing in the UK. However, there is a lack of knowledge about the risk factors (including health, behaviour, and environment) for multimorbidity over time. An interdisciplinary approach is essential, as data science, artificial intelligence, and concepts from engineering (digital twins), have the potential to enable personalised simulation of life-course risk for the development of multimorbidity by identifying key risk factors throughout the life course. Predicting the risk of developing clusters of health conditions before they occur would add clinical value by enabling targeted early preventive interventions, advancing personalised care to improve outcomes, and reducing the burden on the UK's healthcare systems.

This study aims to identify key risk factors that predict multimorbidity throughout the lifetime through the development of an intelligent agent using digital twins so that early interventions can be delivered to improve health outcomes. The objectives of this study are to identify key predictors of lifetime risk of multimorbidity, create a series of simulated computational digital twins that predict levels of risk for specific clusters of factors, and test the feasibility of the system.

Methods: This study will use machine learning to identify key risk factors throughout life that predict the risk of later multimorbidity to develop digital twins. The first stage of the development will be the training of a base predictive model. Data from the National Child Development Study (NCDS), the North West London Integrated Care Record (NWL ICR), the Clinical Practice Research Datalink (CPRD), and Cerner's Real World Data will be split into subsets for training and validation, which will be done following the k-fold cross-validation procedure and assessed with the PROBAST risk of bias tool. Two additional datasets - from the early-LIfe data cross-LInkage in Research (eLIXIR) study and the Children and Young People's Health Partnership (CYPHP) randomised controlled trial - will be used in addition to the model to develop a series of digital twin personas that simulate clusters of factors that predict different levels of risk of developing multimorbidity.

Results: The expected results are a validated model, a series of digital twin personas, and an assessment of proof-of-concept.

Discussion: Digital twins could provide an individualised early warning system that predicts the risk of future health conditions and recommends the intervention that is most likely to be effective at minimising that risk. These insights could have a significant positive impact on an individual's quality of life and healthy life expectancy and reduce population-level health burdens.

Keywords: Artificial Intelligence (MeSH), Machine Learning (MeSH), Mulitmorbidity (MeSH), Mental Health (MeSH)

Introduction

Background

Multimorbidity, the co-occurrence of two or more long-term conditions, has been associated with numerous adverse outcomes and overwhelming financial burdens on healthcare systems [1]. Its prevalence in the UK is projected to almost double by 2035, and at least two thirds of the gain in life expectancy above 65 years will be spent with four or more chronic conditions [2]. Multimorbidity is most common in the elderly, but it is seen in patients of all ages [3] and is more prevalent, and at younger ages, in people who live in socioeconomically deprived areas [4]. Health, behavioural, and environmental factors significantly influence individuals' risk and many processes that lead to multimorbidity begin with chronic illness at much earlier ages. Multimorbidity can occur in clusters with predictable disease pathways [5]; however, the specialisation of healthcare has created a challenge for identifying and treating co-occurring health conditions. An interdisciplinary approach combining clinical knowledge of disease pathways, clusters, and risk factors with artificial intelligence (AI) technology could improve understanding of the bio-social factors that are associated with developing multimorbidity. Understanding key risk factors for multimorbidity could enable doctors to monitor and treat patients more effectively, potentially mitigating or preventing multimorbidity.

Rationale

A holistic, patient-centred system will be essential to address the challenge of preventing and managing multimorbidity and will require a combination of prediction, monitoring, and intervention. Digital twins - a concept developed for engineering that uses real world data and AI to provide a virtual representation of a physical counterpart - are starting to be explored in healthcare. The concept of the digital twin can be interpreted in relation to patients as a means of improving diagnostics and treatments, processing vast amounts of data to develop predictive health trajectory for individuals [6]. Our aim is to adapt this engineering solution to draw from continuously updating individual data on factors and health outcomes to simulate an individual's future health status [7].

There is little established evidence on effective means of identifying risk factors or preventing and managing multimorbidity throughout the life course [8,9]. The vast majority of multimorbidity studies aim to identify disease clusters at a single point in time, providing little information about how multimorbidity develops over time within individuals [10]. This study will develop a system where innovative AI approaches are used to analyse complex longitudinal data and to predict levels of risk of multimorbidity. Using simulation, digital twins could identify key risk factors, consider weaknesses in source data through quantified under-reporting, model potential adverse health outcomes and, given adequate information, make recommendations for interventions that are most likely to be effective. Digital twins aim to make significant enhancements within healthcare and advance knowledge of disease [6]. Furthermore, used in conjunction with clinician experience and knowledge, will aid decision making and highlight risk factors early on, so that steps can be taken to try and mitigate them and avoid later health consequences. The proposed digital twin solution has the potential to have a significant positive impact on individuals and healthcare systems; however, because the proposed solution is a new and interdisciplinary innovation that will

need significant development and evaluation, this research will not be possible without funding.

Hypothesis and Aim and Objectives

The aim of this study is to identify key risk factors that predict multimorbidity throughout the lifetime through the development of an intelligent agent using digital twins. The hypothesis is that this intelligent agent can use a wide range of bio-social variables (including physical and mental health, behavioural, socio-economic, relational, and environmental conditions throughout an individual's life) to predict risk of multimorbidity.

The three main objectives of the study are to:

1. Identify key indicators that most accurately predict life-time risk of developing multimorbidity;
2. Implement artificial intelligence via a digital twin to simulate factors impacting people throughout their life to identify when they are at risk of later developing multimorbidity, enabling early and preventive interventions based on the critical indicators collected through patient-generated monitoring and medical records;
3. Evaluate the feasibility of the digital twin system by assessing the validity and performance of the predictive model.

Methods

Study design

Using an implementation science theoretical framework and a retrospective cohort design, this will be a feasibility study aiming to address the following research question: Can historical data and data captured via a dynamic remote monitoring system be used to develop a digital twin that can predict individual risk of developing multimorbidity over a lifetime? The theoretical framework that has best fit with this study is that of implementation research: development of innovations [11]. Here the framework provides guidance regarding design and conduct, facilitating innovative ideas into practice. The framework gives ideas for various intended research pursuits and suggests ways to achieve these [11].

Three longitudinal databases will be used to determine how the presence or absence of particular factors occurring in childhood (health, behavioural, and environmental) is associated with the development of health conditions and multimorbidity over the life course. The data is longitudinal and is needed to inform about childhood factors and health outcomes over the life course for the model to identify potential early predictors of later multimorbidity. The study will consist of two phases: prototype development and validation, during which the model will be trained to identify these associations using subsets of the datasets and tested on the remaining subsets. These datasets will be used to determine how the presence or absence of particular factors is associated with the development of multimorbidity over the life course. Table 1 and Figure 1 provide an overview of the study design and framework.

Table 1. PICO table

Population	Data for system development will be drawn from all age groups
Intervention	An intelligent agent using digital twins that includes a dynamic remote monitoring system and a general predictive model
Comparator	Real-world evidence via longitudinal data collected from databases will provide a means of validating the predictions made by the model
Outcomes	Predicted risk of developing health conditions and multimorbidity over the lifetime

Figure 1. LifeTIME study logic diagram

Data collection

The healthcare data that will be used to train and validate the model will be collected from several pre-established databases. To satisfy the data requirements of model training and validation, a few longitudinal datasets will be used and a variety of factors - including mental and physical health, behavioural, and environmental (including socio-economic) indicators will be examined. The cohort used to train and validate the model will be built using data from previously established databases, split into random subsets. It will build upon our extensive experience of data linkage, the harmonisation of multiple sources of patient related e-records [12] and use of AI for health data analytics [13,14]. Using multiple databases will allow for a broader range of variables to be included in model development. Four longitudinal databases (the National Child Development Study [15], the Clinical Practice Research Datalink [16,17], the North West London Integrated Care Record (Discover-NOW) [18], and Cerner's Real-World Data™ from UK and Irish healthcare services [19]) were selected because they cover a range of mental and physical health outcomes, health behaviours, and other characteristics over decades in a diverse UK population (see Table 2).

Table 2. Characteristics of databases that will be used for model training and validation

Database	Start	# patients	Location	Included data
National Child Development Study (NCDS) [15]	1958	17,415	England, Scotland and Wales	Physical/educational development, economic circumstances, employment, family life, health behaviour, wellbeing, social participation and attitudes
Clinical Practice Research Datalink (CPRD GOLD and Aurum) [16,17]	1987 (GOLD) 1995 (Aurum)	>11 million (GOLD) >19 million (Aurum)	UK (GOLD) England (and N. Ireland starting 2019; Aurum)	Demographics, diagnoses, symptoms, signs, prescriptions, referrals, immunisations, behavioural/lifestyle factors, tests
North West London Integrated Care Record (DiscoverNOW) [18]	2015	>2.3 million	North West London	Data from all care settings (primary care, acute, mental health, community and social care), for all disease areas

Cerner Real-World DataTM [19]		~20 million	UK (30 trusts) Ireland (7 hospitals)	Data recorded in electronic patient records

Model and digital twin development

Vast amounts of data will input into the development of the datasets that will be used to develop a computational model to establish the relationship between bio-social factors and later health outcomes. Machine learning will be used to train a model using a variety of factors. The model will need to be trained on longitudinal data that includes both bio-social characteristics and later health conditions so that links between factors and outcomes can be established and to account for historical differences in risk factor and multimorbidity prevalence.

Once the predictive model has been developed based on the four longitudinal databases detailed in Table 2, two additional datasets will be incorporated into the development of the digital twins. Data from the early-LIfe data cross-LInkage in Research (eLIXIR) study [20] and the Children and Young People's Health Partnership (CYPHP) randomised control trial [21] will be included in the development of the digital twins because they track early life variables and use of healthcare services. Using the preliminary model developed from the longitudinal datasets, and machine learning to identify clusters of factors that best predict the later risk of developing multimorbidity, a series of digital twin personas will be created to provide simulations of different types of people (see Figure 2).

Figure 2. LifeTIME overview

Although the two additional datasets do not contain longitudinal health outcomes, incorporating them at this stage will enable the digital twin personas to be developed based on the clustering of a more comprehensive set of variables. For instance, if a particular variable was identified in the model as a predictor of multimorbidity, these additional datasets can provide more detail about variables that commonly co-occur with that predictor. This is beneficial because these datasets will provide more current data about early life factors than the longitudinal databases. The number of digital twin personas developed will depend on the clustering of the predictive factors. For this study, these personas will be developed to represent children, but in the longer-term, they will be developed to simulate different stages throughout life.

Evaluation

The validity of the developed model and the digital twin prototypes will be assessed using the k-fold cross-validation procedure [22], with k being determined when total cohort size is known. The machine learning algorithms will be tested on the validation dataset to determine which model has the best fit. Model validity will be examined using a variety of indicators, including receiver operating characteristic (ROC) and precision-recall (PR) curves. The PROBAST tool will be used to assess the model's risk of bias [23].

Table 3. Primary and secondary objectives and outcomes

Objective	Primary Outcome	Secondary Outcomes

Identify key indicators that most accurately predict lifetime risk of multimorbidity	Risk factors for multimorbidity	
Assess the validity of a model which identifies variables and predicts life-time risk of developing multimorbidity	Validity	Risk of bias

Ethics

Ethical approval for this study will be sought from the University of Plymouth's Health Faculty Research Ethics and Integrity Committee. All the data being used in this proof-of-concept study will be procured from pre-existing databases. Therefore, all data access will be dependent on the approval of the study protocol by the data controllers for those databases; for instance, the CPRD database requires that study protocols be approved by an Independent Scientific Advisory Committee (ISAC) before data is shared. All storage and use of the data provided will comply with database-specific and GDPR requirements. It will be confirmed before access that the data has been sufficiently anonymized to comply with GDPR requirements and that the databases' patient consent processes extend to cover the sharing of anonymized data for other research purposes (i.e. that their personal health data is being used in a way that could be reasonably expected by the participants). Processes for the removal of patient data - at their request to the specific database - will be established before data use with each of the database managers. Potential ethical issues relating to the development and implementation of digital twins in healthcare will be identified from the literature and discussions with the patient and public steering group.

Results

The study is expected to produce a validated model of the factors that are associated with later development of mental health conditions and multimorbidity, and based on that model, a series of digital personas. The results collected will inform assessments of the feasibility of pursuing further development and evaluation of the digital twin system; if unfeasible in current state, findings will inform understanding of problems and iterative development. The results will provide data to inform a base model that in later studies, will be trained and developed using remote monitoring and data collection from that particular individual in combination with population health data to personalise the digital twin to more closely match the real patient. This study is a starting point.

Discussion

Overview

By associating a patient with a digital twin persona and then personalising it with their evolving data, healthcare providers and researchers could receive a personalised risk score that represents the probability that the patient will develop further health complications. By having the capability to start at a young age, a digital twin system has the potential to provide significant clinical value by identifying risk factors early, so that preventative interventions can be made to reduce the risk of the individual developing long-term health conditions. This

would save healthcare providers' time, hospitals' and healthcare systems' resources, and people from the avoidable negative health, well-being, and economic costs of health conditions.

Future Directions

If the study results are positive, future research could be conducted to develop and incorporate a remote monitoring system that can collect and collate individual data with the population data and assess it in a pilot trial. This will enable the digital twins to be further developed and iteratively improved based on the feedback from the participants. As the system will use sensitive personal data and has ethical implications, safety and efficacy data will be an essential next step on the path to achieving wide-scale patient benefit.

Based on users personal data, healthcare providers and researchers would be able to associate individual patients with the digital twin persona that is the best fit for them. This will then be able to provide a risk score that represents the probability that the child will develop further complications or health conditions. In the future, the digital twin will also be trained with data on interventions so that it can provide recommendations to healthcare providers about what interventions are most likely to be effective for that individual. In the long-term, data collected from this remote monitoring system will be combined with the childhood digital twin personas to increase the variables for prediction and to personalisethe personas based on individual health data.

Limitations

Potential limitations of the study have been identified, with plans for their mitigation, to maximise the potential usefulness of the study.

1. Data linkage: The datasets used in the study are from a variety of databases that were not designed to be integrated and were not collected from the same populations; mitigated by experience of the research team and use of machine learning techniques to aggregate a large amount of data.
2. Data quality: Understanding the errors in data is key to making predictions. This can be more difficult to quantify for survey data (particularly subjective questions) than for technical medical data; mitigated by collecting data from multiple sources with large sample sizes and by prioritising the inclusion of validated questionnaires in the remote monitoring system.
3. Lack of individualisation: Developing a model based on pre-existing datasets that do not all follow specific individuals means that there is a risk that the ‘digital twin’ will only be a predictive model; mitigated by the development of a novel and linked system of collecting remote monitoring and sensing data that can be incorporated into digital twin personas.

Conclusion

Incorporating a remote monitoring system would enable tracking of ongoing lifestyle data, which could be combined with the individual’s personal health data and population health and lifestyle data to potentially increase the potential impact. Personalised patient simulations

could be used to optimise prevention or treatment selection, saving patients unnecessary side effects, reduce the number of treatments attempted, and reduce the need for repeat hospital visits. Prevention and early intervention to manage multimorbidity will support healthy ageing and reduce its negative impacts on quality of life and the workforce.

Author Contributions

MMI and EM conceived of the study topic and designed and drafted the protocol. LF, AK, DW, MvV, JM, IW, and TH contributed to the revision of the protocol. Final revision was conducted by EM.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Conflicts of Interest

EM is the Editor-in-Chief of JMIRx Med.

References

1. Soley-Bori M, Ashworth M, Bisquera A, Dodhia H, Lynch R, Wang Y, Fox-Rushby J. Impact of multimorbidity on healthcare costs and utilisation: a systematic review of the UK literature. Br J Gen Pract 2021 Jan;71(702):e39–e46. PMID:33257463
2. Kingston A, Robinson L, Booth H, Knapp M, Jagger C, MODEM project. Projections of multimorbidity in the older population in England to 2035: estimates from the Population Ageing and Care Simulation (PACSim) model. Age Ageing 2018 May 1;47(3):374–380. PMID:29370339
3. Nguyen H, Manolova G, Daskalopoulou C, Vitoratou S, Prince M, Prina AM. Prevalence of multimorbidity in community settings: A systematic review and meta-analysis of observational studies. J Comorb 2019 Jan;9:2235042X19870934. PMID:31489279
4. Chapter 5: inequalities in health [Internet]. Public Health England. 2018 [cited 2021 Feb 9]. Available from: https://www.gov.uk/government/publications/health-profile-for-england2018/chapter-5-inequalities-in-health
5. Whitty CJM, MacEwen C, Goddard A, Alderson D, Marshall M, Calderwood C, Atherton F, McBride M, Atherton J, Stokes-Lampard H, Reid W, Powis S, Marx C. Rising to the challenge of multimorbidity. BMJ [Internet] British Medical Journal Publishing Group; 2020 Jan 6 [cited 2021 Apr 26];368. PMID:31907164
6. Björnsson B, Borrebaeck C, Elander N, Gasslander T, Gawel DR, Gustafsson M, Jörnsten R, Lee EJ, Li X, Lilja S, Martínez-Enguita D, Matussek A, Sandström P, Schäfer S, Stenmarker M, Sun XF, Sysoev O, Zhang H, Benson M, Swedish Digital Twin Consortium. Digital twins to personalize medicine. Genome Med 2019 Dec 31;12(1):4. PMID:31892363
7. Wright L, Davidson S. How to tell the difference between a model and a digital twin. Advanced Modeling and Simulation in Engineering Sciences SpringerOpen; 2020 Mar 11;7(1):1–13.
8. Fabbri E, Zoli M, Gonzalez-Freire M, Salive ME, Studenski SA, Ferrucci L. Aging and Multimorbidity: New Tasks, Priorities, and Frontiers for Integrated Gerontological and Clinical Research. J Am Med Dir Assoc NIH Public Access; 2015 Aug 1;16(8):640. PMID:25958334
9. Navickas R, Petric V-K, Feigl AB, Seychell M. Multimorbidity: What do we know? What should we do? Journal of Comorbidity SAGE Publications; 2016;6(1):4. PMID:29090166
10. Cezard G, McHale C, Sullivan F, Bowles J, Keenan K. Studying trajectories of multimorbidity: a systematic scoping review of longitudinal approaches and evidence [Internet]. bioRxiv. medRxiv; 2020. [doi: 10.1101/2020.11.16.20232363]
11. Hamilton A.B and Mittmen, B.S (****) Implementation Science in Health Care Chapter 23, p385 in Dissemination and Implementation Research in Health Oxford University Press DOI: 10.1093/050/9780/90683214.003.0023
12. Zhou S-M, Tsang G, Xie X, Huo L, Brophy S, Lyons RA. Mining electronic health records to identify influential predictors associated with hospital admission of patients with dementia: an artificial intelligence approach. Lancet Elsevier BV; 2018 Nov;392:S9.
13. Zhou S-M, Lyons RA, Bodger OG, John A, Brunt H, Jones K, Gravenor MB, Brophy S. Local modelling techniques for assessing micro-level impacts of risk factors in complex data: understanding health and socioeconomic inequalities in childhood educational attainments. PLoS One 2014 Nov 19;9(11):e113592. PMID:25409038
14. Zhou S-M, Chiclana F, John R, Garibaldi J, Huo L. Type-1 OWA operators in aggregating multiple sources of uncertain information: Properties and real world applications. IEEE Trans Fuzzy Syst Institute of Electrical and Electronics Engineers (IEEE); 2020;1–1.
15. 1958 National Child Development Study [Internet]. UCL Centre for Longitudinal Studies. [cited 2021 Apr 23]. Available from: https://cls.ucl.ac.uk/cls-studies/1958-national-child-developmentstudy-2/

16. Herrett E, Gallagher AM, Bhaskaran K, Forbes H, Mathur R, van Staa T, Smeeth L. Data Resource Profile: Clinical Practice Research Datalink (CPRD). Int J Epidemiol Oxford Academic; 2015 Jun 6;44(3):827–836.
17. Wolf A, Dedman D, Campbell J, Booth H, Lunn D, Chapman J, Myles P. Data resource profile: Clinical Practice Research Datalink (CPRD) Aurum. Int J Epidemiol Oxford Academic; 2019 Mar 11;48(6):1740–1740g.
18. North West London Integrated Care Record (NWL ICR) [Internet]. HDRUK Innovation Gateway. 2020 [cited 2020 Sep 8]. Available from: https://web.www.healthdatagateway.org/dataset/7f23034b-6887-48c0-bbe7-07d099ca9395
19. Real-World Data [Internet]. Cerner. [cited 2021 Apr 23]. Available from: https://www.cerner.com/solutions/real-world-data
20. eLIXIR Study [Internet]. Guy's and St Thomas' Biomedical Research Centre. 2018 [cited 2021 Apr 27]. Available from: https://www.guysandstthomasbrc.nihr.ac.uk/our-work/researchthemes/women-and-childrens-health/elixir-study/
21. Newham JJ, Forman J, Heys M, Cousens S, Lemer C, Elsherbiny M, Satherley RM, Lingam R, Wolfe I. Children and Young People's Health Partnership (CYPHP) Evelina London model of care: protocol for an opportunistic cluster randomised controlled trial (cRCT) to assess child health outcomes, healthcare quality and health service use. BMJ Open [Internet] BMJ Open; 2019 Sep 3 [cited 2021 Feb 22];9(8). PMID:31481366
22. Nguyen D, Long T, Jia X, Lu W, Gu X, Iqbal Z, Jiang S. A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. Sci Rep Nature Publishing Group; 2019 Jan 31;9(1):1–10.
23. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, Reitsma JB, Kleijnen J, Mallett S, PROBAST Group†. PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. Ann Intern Med 2019 Jan 1;170(1):51–58. PMID:30596875