Exploration of Digital Health Technologies

Open Exploration

Open Access Systematic Review

Data science techniques to gain novel insights into quality of care: a scoping review of long-term care for older adults

Ard Hendriks[1,2], Coen Hacking[1,2], Hilde Verbeek[1,2], Sil Aarts[1,2*]

[1]Living Lab in Ageing and Long-Term Care, Maastricht University, 6211 LK Maastricht, The Netherlands

[2]Department of Health Services Research, CAPHRI Care and Public Health Research Institute, Faculty of Health Medicine and Life Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands

***Correspondence:** Sil Aarts, Living Lab in Ageing and Long-Term Care, Maastricht University, 6211 LK Maastricht, The Netherlands; Department of Health Services Research, CAPHRI Care and Public Health Research Institute, Faculty of Health Medicine and Life Sciences, Maastricht University, 6200 MD Maastricht, The Netherlands. s.aarts@maastrichtuniversity.nl

Academic Editor: Andy Wai Kan Yeung, The University of Hong Kong, China

Received: July 17, 2023 **Accepted:** November 3, 2023 **Published:** April 12, 2024

Cite this article: Hendriks A, Hacking C, Verbeek H, Aarts S. Data science techniques to gain novel insights into quality of care: a scoping review of long-term care for older adults. Explor Digit Health Technol. 2024;2:67–85. https://doi.org/10.37349/edht.2024.00012

Abstract

Background: The increase in powerful computers and technological devices as well as new forms of data analysis such as machine learning have resulted in the widespread availability of data science in healthcare. However, its role in organizations providing long-term care (LTC) for older people LTC for older adults has yet to be systematically synthesized. This analysis provides a state-of-the-art overview of 1) data science techniques that are used with data accumulated in LTC and for what specific purposes and, 2) the results of these techniques in researching the study objectives at hand.

Methods: A scoping review based on guidelines of the Joanna Briggs Institute. PubMed and Cumulative Index to Nursing and Allied Health Literature (CINAHL) were searched using keywords related to data science techniques and LTC. The screening and selection process was carried out by two authors and was not limited by any research design or publication date. A narrative synthesis was conducted based on the two aims.

Results: The search strategy yielded 1,488 studies: 27 studies were included of which the majority were conducted in the US and in a nursing home setting. Text-mining/natural language processing (NLP) and support vector machines (SVMs) were the most deployed methods; accuracy was the most used metric. These techniques were primarily utilized for researching specific adverse outcomes including the identification of risk factors for falls and the prediction of frailty. All studies concluded that these techniques are valuable for their specific purposes.

Discussion: This review reveals the limited use of data science techniques on data accumulated in or by LTC facilities. The low number of included articles in this review indicate the need for strategies aimed at the effective utilization of data with data science techniques and evidence of their practical benefits. There is a need for a wider adoption of these techniques in order to exploit data to their full potential and, consequently, improve the quality of care in LTC by making data-informed decisions.

Keywords

Data science, long-term care, analyzing methods, big data, big data analytics, older adults, nursing homes

Introduction

Data science is a rapidly evolving field that offers many valuable applications for healthcare and may be defined as a set of fundamental principles that support and guide the extraction of information and knowledge from often vast amounts of data, also known as “big data”. Big data refers to large amounts of data that often originate from different sources [e.g., websites, electronic health records (EHRs), questionnaires, and interviews], are collected quickly, and are often not only numerical in nature. Although no single widely accepted definition of big data appears to be available, the concept is often described using the four V’s [1]: volume, variety, velocity, and veracity. Volume refers to large volumes of data, while variety applies to the different forms and domains of data that can be analyzed individually, but can also be combined, velocity relates to the fast rate at which the data is collected and stored, and veracity refers to the quality.

Examples of data science techniques often used for the analyses of vast amounts of healthcare data include data- and text-mining, machine learning (ML), pattern recognition, and neural networks [2]. Systematic reviews on the effectiveness of big data in healthcare have concluded that it may lead to positive changes in health behavior, as well as improved public health policy-making and overall decision-making [3–5]. In addition, these studies argued that the vast amounts of data have the potential to improve the quality of care while simultaneously reducing the costs, as well as lowering readmission rates and supporting policy-makers and clinicians in developing public policy and service delivery, in addition to assisting hospital management with improving the efficiency of care services and the provision of personalized care to patients [2, 6]. Despite these promising benefits, the use of these vast amounts of data and innovative data science methods in long-term care (LTC) for older adults seems to be lagging behind other healthcare areas such as hospitals [7, 8]. Hence, LTC organizations are not currently using the growing amount of data they collect on a daily basis to gain novel insights and foster improvements.

LTC may be characterized as a “set of services delivered over a sustained period of time to people who lack some degree of functional capacity” and can be provided either at home or in LTC facilities such as nursing homes (NHs) or assisted living facilities [9–11]. In many countries, LTC is being confronted with significant demographic changes and staff shortages while trying to provide high levels of care and remain financially sustainable [12]. Emerging technological advances and the continuous implementation of digitalization have the potential to mitigate these challenges, at least partly. Information is of utmost importance: the more high-quality data there is, the more optimally care can be organized [13]. As volumes of data continue to pile up and data science gradually penetrates all parts of healthcare, the possibilities of data science for providing novel information, and thus knowledge, related to quality of care for clients and quality of work for staff in LTC can be considered endless. However, the role of data and data science (techniques) in LTC remains unclear.

Published reviews conducted regarding LTC focused on specific individual smart technologies such as sensors or robotics, and merely examined the technology itself, rather than the data it accumulated [14, 15]. In addition, a recent review on LTC concentrated solely on the acceptability and effectiveness of artificial intelligence (AI) interventions such as smartphone applications, thereby excluding other types of data gathered for LTC [16]. Hence, the literature on the use of data science techniques on data accumulated in LTC has yet to be systematically synthesized. We therefore systematically reviewed the literature on the application of data science techniques to analyze (large amounts of) data collected in or by LTC organizations to gain novel insights. The aim of this review was twofold: 1) to assess what data science techniques are used on data accumulated in LTC and for what specific purposes and 2) to assess the results of these techniques in researching study objectives.

Methods

A scoping review was conducted. Both the recently updated guidelines for scoping reviews by the Joanna Briggs Institute [17] as well as the preferred reporting items for systematic reviews and meta-analyses (PRISMA) extension for scoping reviews checklist were followed [18].

Search strategy

PubMed and Cumulative Index to Nursing and Allied Health Literature (CINAHL) were deployed for relevant studies. The search was conducted in December 2022. Medical Subject Headings (MeSH) terms, standardized keywords manually assigned by indexers of the National Library of Medicine, were used. The following search string was used: ("Big Data"[MeSH Terms] OR "Big Data analytics"[All Fields] OR "data analytics"[All Fields] OR "Data Science"[MeSH Terms] OR "Medical Informatics"[MeSH Terms] OR "Artificial Intelligence" [MeSH Terms] OR "Machine Learning"[MeSH Terms] OR "Deep Learning" [MeSH Terms] OR "Data Mining"[MeSH Terms] OR "text mining" [All Fields]) AND ("Residential Facilities" [MeSH Terms] OR "residential home*" [All Fields] OR "care home*" [All Fields] OR "Assisted Living Facilities"[MeSH Terms] OR "Homes for the Aged" [MeSH Terms] OR "Nursing Homes"[MeSH Terms]).

Inclusion and exclusion criteria

Publications were included if: 1) they reported on a data science technique for obtaining information from data, which might include "rather novel" techniques such as deep learning and text-mining, but also more "traditional techniques" such as regression analyses. Since there is considerable overlap between math, statistics, data science, and computer science [19] and this review is the first one of his kind, a broad scope was chosen, 2) they were based on data accumulated in or by an LTC facility for older adults, with a facility being considered an LTC facility if it accorded with the following description by Sanford et al. [9] (2015): "LTC occurs in a residential facility or NHs and is primarily intended for those who require assistance with activities of daily living and instrumental activities of daily living, and/or for those who have behavioral problems due to dementia", and 3) they reported original research (e.g., letters to the editor or comments were excluded). Studies were also excluded if they were not published in English and if the full text was not available. The search was not limited by research design or publication date.

Selection process

The screening and selection process was carried out by two authors (AH and SA) (see Figure 1): the data were extracted in duplicate into separate Excel forms (available upon request). The studies yielded from the search strategy were first screened for eligibility based on their titles. Titles that did not comply with the pre-specified inclusion criteria were removed, while ambiguous ones were kept separate and further discussed among all co-authors. Afterward, the abstracts of titles that fit the pre-specified inclusion criteria were screened. Abstracts that did not meet the inclusion criteria were removed and the reasons for removal were noted. The remaining publications were assessed for eligibility based on their full texts. Those that did not meet the inclusion criteria based on their full text were assessed as ineligible and excluded from use in the current review. Again, the reasons for exclusion were noted.

Data extraction and analyses

The standardized form for data extraction in the Joanna Briggs Institute guidance was used as a basis and adapted to meet the needs of the current scoping review [17]. The study characteristics were described in tabulated form: author(s), year of publication, country of origin, objective, setting and study population, analyzing technique, metric used, conclusion, limitations, and whether ethical approval had been obtained (see Table 1). The overall findings were reported by means of narrative synthesis based on the two postulated aims. In order to provide a broad overview of this topic, a methodological quality assessment of the included works was not performed, consistent with the methodology of scoping reviews [17].

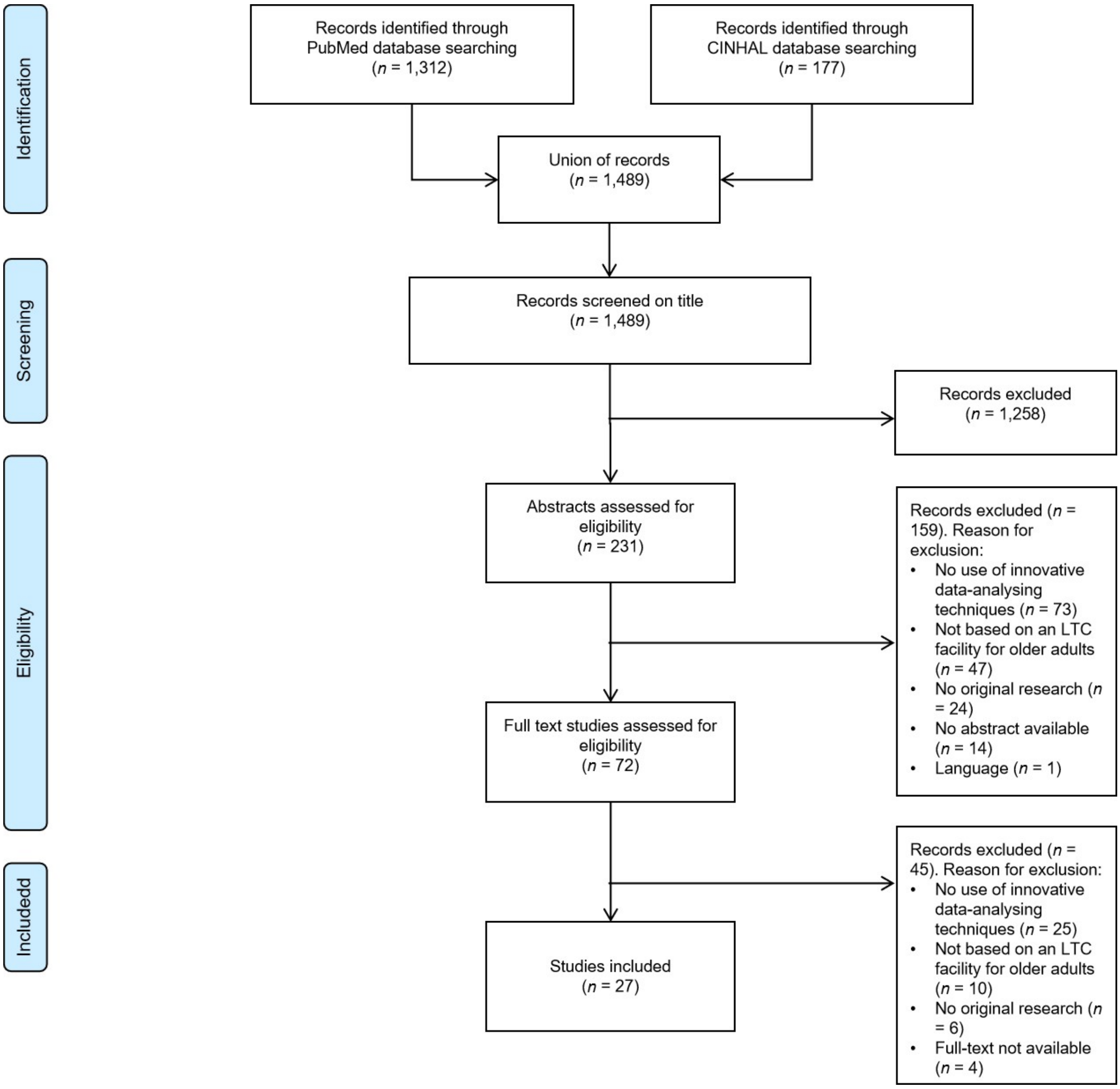

Figure 1. Flowchart displaying the selection process

Table 1. Characteristics and conclusions of included studies

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
1. Zhu et al. [26] (2022) Australia	To estimate the prevalence of agitated behaviors in people with dementia in NHs	Nursing notes from EHRs regarding NH residents with dementia (*n* = 3,528)	Rule-based NLP to detect health terminology, terminology regarding dementia, and agitation-related terms	F-score.	NLP can be valuable in evaluating agitation in people with dementia, and the identified behaviors can inform improvements in aged care and nursing	Relies on the accuracy and completeness of EHRs. The NLP methodology could not capture the entire diversity of writing styles	Ethical approval was obtained

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
2. Wang et al. [35] (2022) China	To develop an early diagnostic tool for Alzheimer's disease using ML and non-imaging factors	NHs in Hangzhou, China (n = 4). NH residents aged 65 or older (n = 654). Community members (n = 1,100)	Logistic regression, SVM, neural network, random forest, XGBoost, LASSO, and best subset models	Sensitivity, specificity, accuracy, AUROC	The developed non-imaging-based diagnostic tool effectively predicts dementia outcomes and can be easily integrated into clinical practice. Its online implementation eliminates barriers to usage, thereby improving dementia diagnosis, care quality, and reducing associated costs	Limited study sites	Ethical approval was obtained
3. Huang et al. [34] (2022) China	Using AI to improve the time required for nurse-patient interaction	NH residents (n = 32)	Real-time analysis of streamed video data through CNN	Accuracy	Automatic monitoring effectively improved the efficiency of nurse-patient interaction. The system achieved an abnormal status recognition accuracy of up to 96.53%	Video data could raise privacy concerns	Ethical approval was obtained
4. Boyce et al. [25] (2022) US	To develop and validate a novel predictive model that forecasts the risk of falls for NH residents 90 days in advance, utilizing data from the LTC MDS and drug therapy records	NH residents (n = 3,985) in 2011, 2012, 2013, and 2016–2018 from the University of Pittsburgh Medical Center Senior Communities NHs	An ML approach, known as CART was used	Precision, recall, specificity, balanced F-measure, threshold	The study successfully developed a novel, easily interpretable fall prediction model using MDS and drug dispensing/administration data, capable of guiding clinicians and NH staff in identifying individual residents' fall risk within 90 days	The model, trained and tested within a single health system, may require additional testing and potential retraining for use in other settings, and it does not currently incorporate promising data from wearable sensors for real-time fall prediction	Not mentioned
5. Ritchie et al. [42] (2022) United Kingdom	To determine the prevalence of AF and temporal trends by year of care home entry, and associations between AF and adverse health outcomes including stroke, TIA, major bleeding, MI, cardiovascular hospitalization, and mortality	NH residents in Wales between 2003 and 2018 (n = 86,602)	Unadjusted logistic regression models to investigate associations with oral anticoagulant usage	95% confidence interval, P-values	The study highlights the need for appropriate blood-thinning medications for stroke prevention and effective management of related heart conditions while emphasizing the need for improved data quality	Certain diagnoses were possibly missed due to positive recordings of diagnoses	Not mentioned

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
6. Hacking et al. [43] (2022) Netherlands	To explore different text-mining methods to analyze the quality of care in a NH setting	Interviews with residents (*n* = 39), family members (*n* = 37), and care professionals (*n* = 49)	Word frequency analyses, correlation analyses, deep learning-based sentiment analysis, and topic clustering using k-means clustering of word2vec vectors	Not mentioned	The study demonstrates the usefulness of text-mining to extend our knowledge regarding the quality of care in an NH setting	Deep learning is less explainable compared to more traditional techniques. Unigram and bigram models don't offer many insights as they contain many words with little significance	Ethical approval was obtained
7. McGarry et al. [24] (2022) US	To examine the association of state COVID-19 vaccine mandates with staff vaccination coverage and staffing shortages at NHs	Data on state COVID-19 vaccine mandate policies were collected from a number of sources, including internet searches using Google, state websites, state memos, and news reports	This study used event study models and linear regressions to analyze the association of state mandates with staff vaccination coverage and staffing shortages in NHs	Not mentioned	State vaccine mandates for NH staff were associated with increased staff vaccine coverage without exacerbating staffing shortages	Data self-reported by NHs, potentially leading to biases. Facilities might underreport staffing shortages due to fear of deficiency citations. Measures might not detect staff departures accurately	Not mentioned
8. Shen et al. [23] (2022) US	To investigate the association of severe outbreaks with staffing measures, such as hires, absences, and departures	Daily shifts (*n* = 333 million) for staff members (*n* = 3.6 million) at facilities (*n* = 15,518) each year on average	This study employs an event study framework with multivariable linear regressions, facility and calendar-time fixed effects, and sensitivity analyses to examine staffing pattern changes during and after a severe outbreak	Not mentioned	Severe COVID-19 outbreaks in NHs lead to significant and lasting reductions in nursing staffing levels, with CNAs experiencing the greatest losses, raising concerns about the potential impact on resident quality of life, morbidity, and mortality	Inability to observe reasons for changes in absences, departures, and new hires. Uncertainty about whether lowered staffing levels were intentional or due to turnover and hiring constraints. Missing early outbreaks not captured by the NHSN data	Requested per Harvard institutional review board policy, but wasn't required because this study uses publicly available data
9. Tadokoro et al. [32] (2022) Japan	To evaluate the therapeutic effect of makeup therapy	Female NH residents with dementia (*n* = 34)	Faces were photographed at baseline and after 3 months and were analyzed with AI software (version of Microsoft Azure Face modified for Japanese patients)	*P*-values, correlation coefficients	Makeup therapy had a chronic beneficial effect on the cognitive function of female patients. The AI facial emotion analysis may be superior to self-reported scales because of its independence on verbal ability and cognition	Small sample. Limited study sites	Ethical approval was obtained

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
10. Reddy et al. [44] (2022) Ireland	To measure and map US county-level spatial accessibility to high-quality NH care. To discover the most relevant socio-demographic variables associated with these levels	Certified NHs in the US	Random forest approaches were used to impute data. Lasso approach was used to select variables for the predictive model	Std. error, *t*-value, *P*-value	Spatial accessibility was high in the Midwest and low in the Southwest and along the Pacific coast. Factors such as the size of the county, ethnicity, and patterns in local employment were related to high-quality care. The ML approach can be used to cast a wide net and select the most important variables	Use of county centroids to represent a county's location. Access to public transport was not considered	NR
11. Withall et al. [29] (2022) Australia	To examine the characteristics of victims and persons of interest regarding domestic violence	A total of 492,393 de-identified, police-recorded domestic violence events from the "new south Wales police force" for the period of January 2005 to December 2016	A rule-based text-mining approach was used to extract data	Percentages	This method demonstrated high precision and recall, highlighting the presence of mental illnesses, types of abuse, and sustained injuries in these narratives	The study is based on police-recorded domestic violence data and may not fully represent the prevalence of elder abuse, especially in NHs, due to potential underreporting	Not mentioned
12. Tadokoro et al. [31] (2021) Japan	To evaluate the immediate effect of makeup therapy on dementia patients	Female NH residents (*n* = 36)	Faces were photographed before and after treatment and were analyzed with AI software (version of Microsoft Azure Face modified for Japanese patients)	*P*-values, correlation coefficients	Makeup therapy is a promising non-pharmacological approach for the immediate elevation of behavioral and psychological symptoms of dementia. The AI software quickly and quantitively evaluated the beneficial effects of makeup therapy	Number of participants was small. Pathological background of dementia was not investigated. Age in the makeup group was higher than in the control group. Total treatment duration was different between the makeup group and the control group	Ethical approval was obtained
13. Lee et al. [45] (2021) Canada	To determine predictors associated with 30 days mortality after a positive SARS-CoV-2 test	Residents in LTC homes (*n* = 84.142)	Random survival forest model	AUC (ROC)	Residents' characteristics related to functional status, comorbidities, and routine laboratory measures were major factors associated with mortality	Asymptomatic transmission of SARS-CoV-2 was not considered. No information on public *vs.* for-profit homes was included. No data on the severity of comorbidity was included	This study did not require approval by a research ethics board and did not require individual consent

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
14. Garcés-Jiménez et al. [41] (2021) Spain	It was hypothesized that anticipating an infectious disease diagnosis by a few days could significantly improve a patient's well-being and reduce the burden on emergency health systems	Residents (*n* = 60) in NHs (*n* = 2)	Data was analyzed using three ML algorithms: Naive Bayes. Filter classifier. Random forest	*P*-values	Infectious diseases can be predicted based on the vital signs collected. Its cost-effective implementation allows disadvantaged areas and less accessible populations to be reached	Need to extend the period of sampling	"Ethical consideration for setting clear limits for the research and protecting people's privacy was implemented"
15. Lee et al. [37] (2021) Korea	To compare a variety of ML methods in terms of their accuracy, sensitivity, specificity, positive predictive values, and negative predictor values by validating real datasets in order to predict factors for pressure ulcers	NHs (*n* = 60). NH residents (*n* = NR)	Representative ML algorithms (random forest, logistics regression, linear SVM, polynomial SVM, radial SVM, and sigmoid SVM) were used to develop a prediction model	Accuracy, sensitivity, specificity, negative predictor values, and positive predictive values	The random forest model had the greatest accuracy and was powerful. ML methods were able to identify many factors that predict pressure ulcers in NHs, including both NH characteristics (e.g., hours per resident day of director and number of current residents) and resident characteristics	NR	Ethical approval was obtained
16. Lee et al. [36] (2020) Korea	To compare different ML methods for predicting falls	NHs (*n* = 60). NH residents (*n* = NR)	Representative ML algorithms (random forest, logistics regression, linear SVM, polynomial SVM, radial SVM, and sigmoid SVM) were applied to a pre-processed NH dataset to develop a prediction model	Accuracy, sensitivity, specificity, negative predictor values, and positive predictive values	The random forest model was the most accurate and is therefore a powerful algorithm to discern predictors of falls in NHs. Organizational characteristics (e.g., current number of residents) as well as personal factors should be considered for effective fall management	The number of falls may have been overestimated or underestimated as self-collected data from NHs was used. No differentiations were made in the type of falls, slips, and/or fall-related injuries. Relatively small sample size to train a stable ML model. Parameter tuning was not included	Ethical approval was obtained

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
17. Ambagtsheer et al. [28] (2020) Australia	To assess the effectiveness of AI algorithms compared to the electronic Frailty Index in accurately identifying frailty, based on a routinely-collected residential aged care administrative dataset. To identify best-performing candidate algorithms	RCFs (*n* = 10). RCF residents (*n* = 592)	A frailty prediction system was designed based on the electronic Frailty Index identification of frailty. Classification algorithms used are k-nearest neighbors, decision tree, and SVM	Accuracy	AI techniques show potential in accurately identifying frailty in RCFs based on data held in administrative databases. An SVM algorithm was found to be the best-performing. Frailty identification may enable service providers to anticipate and avoid potentially harmful impacts on residents	Most data extractions were performed manually using formulas in MS Excel. An NLP technique would be more efficient and accurate. Data came from a single aged care service provider. The dataset was relatively small	Ethical approval was obtained
18. Buisseret et al. [46] (2020) Belgium	To design a method combining clinical tests and motion capture sensors in order to optimize the risk of fall prediction. To assess the ability of AI to predict the risk of falls from solely sensor raw data	NHs (*n* = 4). NH residents (*n* = 73)	A Timed Up and Go test and a six-min walking test were performed and combined with residents equipped with a homemade wearable sensor gathering kinematic data. An AI algorithm based on deep learning was created. Models based on CNN were trained and tested in order to find the optimal accuracy of the risk of fall prediction	Accuracy, confusion matrices, *P*-values	The Timed Up and Go test was able to predict falls and the homemade wearable sensor was able to measure differences between fallers and non-fallers. It is shown that the combination improves the accuracy of risk of fall prediction at six months and that the AI algorithm trained by raw sensor data has an accuracy of 75% in fall prediction	Small size of the dataset	Ethical approval was obtained
19. Cheng and Cui [33] (2020) China	To optimize the configuration of RCFs, while considering the demand of three stakeholders (government, elderly, investor), by development of a multi-objective spatial optimization model	RCFs in the Jing'an district of Shanghai	A multi-objective spatial optimization model was developed with the goals of maximizing the efficiency and equity of RCF configuration, minimizing travel costs of the elderly, and maximizing the profits of investors	Not mentioned	A significant gap is concluded to be present between the service supply of RCFs and the demand of the elderly. Overall, the optimization model improved efficiency and equity, reduced the travel costs of the elderly, and increased the profits of investors	Policy and resource constraints were not considered. Predictions of the elderly population in the future were not considered	NR

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
20. Sun et al. [20] (2020) US	To inform about preventive measures for COVID-19 infection by identifying and assessing risk and possible vectors of infection, using a ML approach	NHs (*n* = 1,146). NH residents (*n* = NR)	A self-constructed dataset including information on the NHs' facility and community characteristics was used to create predictive features. A tree-based gradient boosting algorithm was used	ROC (AUC), sensitivity, specificity	An ML gradient boosting model is useful to quantify and predict the risk of infection in NHs. Several risk factors of infection were identified (e.g., NH county infection rate, NH size, and the number of separate units). The historical percentage of non-Hispanic white residents was found to be a protective factor	COVID-19 outcomes were inconsistently reported across states. Model performance can be inconsistent in diverse geographic areas. Data was gathered from historical reports, therefore it may not reflect real-time NH characteristics	NR
21. Suzuki et al. [30] (2020) Japan	To assess whether a CNN is able to predict the time of falling based on multiple complicating factors (such as age, severity of dementia, lower extremity strength, and physical function)	NH (*n* = 1). NH residents with Alzheimer's disease (*n* = 42)	Residents were classified into three groups: those who fell within 150 days, within 300 days, and those who did not fall. Lower extremity strength, severity of dementia, and physical dysfunction were assessed using suitable measures. A CNN was created which focused on multiple complicating factor patterns	Accuracy	An accuracy of 65% was found. A deep learning CNN method based on multiple complicating factors is able to predict the time of falling among NH residents with Alzheimer's disease	Some information may be lacking, e.g., about the various types of dementia, medication use, depressive symptoms, or the fall history of residents. These factors have been associated with an increased risk of falling. A larger number of participants and an addition of important covariates, such as the ones previously listed, could have led to a more accurate prediction	Ethical approval was obtained
22. Gannod et al. [21] (2019) US	To explore the application and utility of a recommender system to preference assessment, based on data mining and ML techniques	NHs (*n* = 28). NH residents (*n* = 255)	NH residents' preferences were gathered using the PELI-NH interview and section F of the MDS 3.0. The information gathered was used to develop an ML recommender system, using an apriori algorithm and logistic regression	Precision, recall, accuracy, F1-score	A reasonable rate of accuracy and precision was found regarding the provision of recommendations on potential preferences for a resident. The ML recommender system has the potential to reduce the time needed to complete the PELI-NH interview, while simultaneously still incorporating important individualized preferences of residents	Learning approach was evaluated using a relatively small transaction dataset. Only cognitively capable participants were included. The preferences of individuals with some form of cognitive impairment or those who are not able to communicate were not considered	NR

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
23. Delespierre et al. [38] (2017) France	To illustrate how text-mining of clinical narratives can enhance EMR data. To demonstrate the convergence of information between clinical narrative extracted data and EMR data	NHs (*n* = 127). NH residents (*n* = 1,015)	Textual data was extracted from physiotherapy narratives. Data mining techniques were combined. Standard query language and text-mining were used to build physiotherapy variables. Meaningful words were extracted. Principal components and multiple correspondence analyses have been performed	Not mentioned	It is demonstrated that data mining and text-mining techniques can add new, usable, and simple data to EHRs with the goal of improving residents' health and the quality of care	Merely a selected sample of clinical narratives was used. Matching residents with their associated clinical narratives relied on physiotherapy care observations that varied between NHs	Ethical approval was obtained
24. Jiang et al. [27] (2017) Australia	To identify risk factors related to medication management using text-mining	Residential aged care homes (*n* = 3,607)	Data in the form of reports were collected from the website of the Australian Aged Care Quality Agency. The text data was classified and labeled with representative keywords. Apache OpenNLP was used to extract a word cloud indicating the most frequently used words in text reports about medication management	Not mentioned	Using text data mining, 21 risk factors to fail in medication management were identified. "ineffective monitoring process", followed by "noncompliance with professional standards and guidelines" were found to be the biggest risk factors. The gained information may be useful to improve medication management in residential aged care homes	Evidence may be limited due to the relatively low sample size. The reports used possessed inadequate details about why the failure happened	NR
25. Fernández-Llatas et al. [40] (2013) Spain	To present a set of algorithms based on process mining that may help professionals infer and compare individualized visual models of human behavior	NH (*n* = 1). NH residents (*n* = 9)	The eMotiva process mining framework combining algorithms and visualization interfaces was used. Process mining algorithms were used that filter, infer, and visualise workflows. These workflows were inferred from data collected using indoor location systems and bracelets. PALIA was the main algorithm in the framework	Not mentioned	The process mining technology was useful for inferring and presenting individual models to experts, representing human behavior in a visual and understandable manner	Limited number of cases were used for observation	NR

Table 1. Characteristics and conclusions of included studies (*continued*)

References	Objective	Setting and study population	Analysing technique	Metric used	Conclusion	Limitations	Ethics
26. Lapidus and Carrat [39] (2010) France	To develop a computerized algorithm able to identify the likeliest transmission paths during a person-to-person transmitted illness outbreak	NH residents (*n* = NR)	A computerized algorithm was built using information about the natural history of disease and a dataset about the population structure and chronology of observed symptoms. A simulator was used to assess the efficacy and was compared with reference methods	Proportion of infected subjects	The algorithm was able to provide information on the dynamics of an outbreak and may help identify sources of infection in order to take the right preventive actions	Unclear how the algorithm would deal with missing data	NR
27. Volrathongchai et al. [22] (2005) US	To evaluate the application of a KDD process using a likelihood-based pursuit data mining technique able to predict the likelihood of falls	LTC facility residents (*n* = 9,980)	KDD was applied to data from the MDS. A likelihood-based pursuit technique has been used to construct models able to predict the likelihood of falls and the variables contributing to this likelihood. Four variables known to be associated with falls and two variables known to not be associated with falls were included	L1 norm of error, *P*-values	The likelihood-based pursuit technique was able to identify which of the variables were associated with falls and was able to make fall likelihood predictions based on these variables. It has the potential to be useful in assessing fall risk due to its ability to provide probabilities based on the exact combination of variables present in an individual resident	Models constructed using the likelihood-based pursuit technique required that there is little correlation among the predictor variables: Only five or six variables were included in the likelihood-based pursuit technique	Ethical approval was obtained

CNN: convolutional neural network; EMR: electronic medical record; KDD: knowledge discovery in databases; NR: not reported; NLP: natural language processing; PALIA: parallel activity-based log inference algorithm; RCFs: residential care facilities; SVM: support vector machine; XGBoost: extreme gradient boosting; LASSO: least absolute shrinkage and selection operator; CART: classification and regression tree; AF: atrial fibrillation; TIA: transient ischaemic attack; PELI: preferences for everyday living inventory; ROC: receiver operating characteristic; AUC: area under the curve; MI: myocardial infarction; SARS-CoV-2: severe acute respiratory syndrome coronavirus 2; AUROC: area under the receiver operating characteristic curve; NHSN: National Healthcare Safety Network; COVID-19: coronavirus disease 2019; OpenNLP: open natural NLP; MDS: minimum dataset

Results

The search strategy yielded 1,488 studies. After the screening of titles and abstracts, seventy one studies were read and assessed for eligibility based on a detailed analysis of their full texts (see Figure 1). In total, twenty seven studies fulfilled the pre-specified inclusion criteria and were assessed as eligible for use in the current scoping review. The main reasons for exclusion were a lack of data-analyzing techniques, or being conducted in a setting other than LTC. The selection process is visualized in the flowchart shown in Figure 1.

Characteristics of the included studies

A detailed overview of the characteristics of each included study is shown in Table 1. The majority of studies were published between 2020 and the end of 2022. The countries in which the studies were conducted were diverse: six studies were conducted in the US [20–25], four in Australia [26–29], three in

Japan [30–32] and China [33–35], two in Korea [36, 37], France [38, 39], Spain [40, 41], one in the United Kingdom [42], the Netherlands [43], Ireland [44], Canada [45], and Belgium [46]. The number of included LTC facilities and the size of the study population varied greatly between publications. About half of the studies reported that they had obtained ethical approval from a review board.

Data science techniques used and purpose

A diverse set of data-analyzing techniques was used in the studies. The majority of studies reported to have deployed a form of regression (n = 8), text-mining/NLP (n = 8), random forest models (n = 5), and SVMs (n = 4) (i.e. several studies used various methods). In terms of metric, accuracy was the most used (n = 5); 7 studies did not report a metric. While some studies mentioned deploying an ML technique, other studies refer to the term AI or algorithms, i.e. ML is a part of AI, while algorithms can be considered as part of ML, and thus, of AI [47], indicating that different terms are used for interchangeable to report on the data science techniques at hand. In addition, the terms text-mining and NLP are both used to refer to analyzing (large) amounts of text. Studies did not report on the use of supervised, unsupervised, or semi-supervised methods.

ML techniques were used to predict factors for pressure ulcers and falls, to identify and assess the risk of COVID-19 infection, as well as to develop a recommendation system for preference assessment and infectious diseases. A neural network based on deep learning was used to predict the risk and time of falling and to improve nurse-patient interaction. Text-mining was applied to EMR data in order to identify risk factors related to medication management. A likelihood-based pursuit data mining technique was employed to predict the likelihood of falls. AI software was used to analyze the facial emotion of residents with dementia. Several publications deployed algorithms. One study reported an AI algorithm utilized to identify frailty, while another study reported an algorithm to infer individualized visual models of human behavior. Modified immune algorithms were used to find the most favorable solutions for spatial optimization and, lastly, algorithms were also deployed to identify person-to-person transmission paths during an illness outbreak.

Outcomes of the included studies

All studies concluded that the data science technique used was "effective": each study reported that the data science technique was useful for the study objective at hand. Words and sentences such as "was useful to infer…", "was able to provide information on…", and "can be used to", were stated in the conclusion sections of the included studies.

Three studies compared various ML techniques (e.g., random forest, logistic regression, naive Bayes, etc.) in terms of accuracy and predictive values related to respectively, pressure ulcers, falls, and infectious diseases in NHs; two of them concluded that a random forest model provided the greatest accuracy and prediction for these outcome measures. Other studies using ML techniques were able to quantify and predict the risk of COVID-19 infection in NHs, provide accurate recommendations on potential preferences for an NH resident, map spatial accessibility to high-quality NH care, and predict falls. CNN based on deep learning was found to be accurate in fall prediction among NH residents and to be able to predict the time of falling for those with Alzheimer's disease. In addition, another study deploying CNN, showed that real-time video analyses effectively improved the efficiency of nurse-patient interaction.

Studies using text-mining techniques displayed the ability to identify risk factors related to failed medication management in care homes. In addition, another study analyzing large amounts of text showed that NLP can be valuable in evaluating agitation in people with dementia, and the identified behaviors can inform improvements in aged care and nursing. A likelihood-based pursuit technique was able to identify factors associated with falls and to make fall likelihood predictions based on these factors among LTC facility residents. Two studies using AI for the analyses of facial emotion showed that the AI technique was able to identify the beneficial effect of makeup therapy on the cognitive function of female patients. In addition, they reported that AI may be superior to self-reported scales because of its independence of verbal ability and cognition of the patient at hand.

An SVM algorithm was found to be capable of accurately identifying frailty among RCF residents based on data held in a routinely collected residential care administrative dataset. Moreover, a modified immune algorithm, using data from the geographic information system, was able to evaluate the current configuration of RCFs in a district of Shanghai. A computerized algorithm provided information on the dynamics of a person-to-person transmitted influenza outbreak in NHs, thereby being able to investigate such events. Studies using regression analyses, a more traditional analyzing method, showed that COVID-19 outbreaks led to adverse outcomes such as reductions in nursing staff levels and that COVID-19 vaccine mandates were associated with increased staff vaccinations.

Discussion

The current scoping review is the first to provide an overview regarding the use of data science techniques on data accumulated in LTC. The results show that, even with a very broad scope, only 27 articles were identified in the current review, pinpointing the diminished use of data science techniques deployed in or by organizations providing LTC to analyze the data they accumulate on a day-to-day basis.

Although only a small number of publications were included in this review, and several of these studies included only a small number of participants, all of them concluded that the data science technique at hand was effective and found the data science techniques demonstrated to be useful for the study objective. All the analyses discussed the usefulness of these techniques in qualitative and future potential terms. However, even with the potential benefits (large amounts of) data and data science techniques seem to offer, LTC might struggle with the same problems that other healthcare sectors (e.g., hospitals) were or are still facing: e.g., an absence of knowledge about which data to use and for which purpose, as well as the lack of an appropriate and comprehensive data infrastructure within organizations [48]. In addition, LTC organizations include a variety of data sources that all collect information in various forms: e.g., medical data in EHRs, unstructured textual data based on interviews regarding the experienced quality of care, or real-time data accumulated by sensors or wearables [7, 49]. The integration of these (semi-)structured data, stemming from a large variety of sources, is a challenge in itself. Strategies for mitigating these challenges, including a sufficient data infrastructure and personnel with expertise in data and communication technology are required in order to utilize the full potential of data accumulated in and by LTC [49, 50]. Since the majority of studies included in this review were published in or after 2020 (with 10 articles being published in 2022), the popularity of data science within this care setting may rise in upcoming years. Increasing funding to support research on data accumulation and analyses in LTC, along with integrative collaborations between health scientists and computing experts (e.g., data scientists) may help to address the challenges within this specific care echelon.

Several different data-analyzing techniques were deployed in the included studies, of text-mining/NLP, regression models, and random forest models were the most prevalent. These techniques have already been proven to be useful in other healthcare areas [2, 6], and may therefore be more widely known and used. Interestingly, data science techniques such as text-mining/NLP, a process aimed and analyzing large amounts of natural language data [51], are primarily reported on in 2022. A review conducted in 2018, reported NLP to be among the most used big data techniques in clinical and operational healthcare [6]. In LTC, much quantitative and qualitative information is digitally recorded in EHRs: e.g., client characteristics (e.g., socio-demographic characteristics) and data on various quality indicators (e.g., pressure ulcers) are collected to map the quality of life as well as the quality of life. These data would be perfectly suitable for data exploration using text-mining/NLP. For example, text fields in EHRs can be analyzed: e.g., can certain words (e.g., "imbalance", "restlessness" or even specific types of medication) predict future falls in clients or future agitated behavior? These large amounts of text can thus be utilized to identify and predict critical behavior or symptoms and, in turn, initiate actions in a timelier manner. Interestingly, the terms ML, AI, and algorithms seem to be used interchangeably and for the same purpose: to describe the method that was deployed (e.g., some studies report deploying an "AI method", while others report using ML or "a powerful algorithm to predict"). While the terms AI, ML, and algorithms fall in the same domain as data science and are indeed interconnected, they all do have specific applications and meanings [47]. When reviewing the metric used, accuracy, measuring the number of correct predictions made by a model, was most prevalent. However, various studies do not specify their method(s) or the metric(s). In order to be more transparent about the usefulness of the methods, more information regarding these measurements should be included in upcoming studies. Especially with the rise of emerging methods such as large language models [i.e. ChatGPT], which have the ability to speed up the use of data science techniques in LTC, information about the used methods and metrics is needed in order to indicate their usefulness for daily care practice.

The majority of studies in this review were focused on the prediction of adverse health problems such as falls, pressure ulcers, and infectious diseases. Not surprisingly, these health problems are reported to be among the most prevalent in LTC organizations [12, 52, 53]. Hence, these studies underscore that novel data-analyzing techniques are used to predict the incidence of already well-known daily care problems in LTC.

Surprisingly, not all studies reported that they had obtained approval from an ethical review board or committee. Ethics forms a major concern due to the vulnerability of patients in LTC and due to the inherent sensitivity of health-related data [7, 54]. With the increasing amount of data available in healthcare and, more specifically in LTC, data ethics have become increasingly important in this sector. Ethical mistakes may lead to social rejection or imperfect policies and legislation, perhaps resulting in a diminished acceptance and progress of data science within the field of LTC [54].

The current study is the first to provide information on the use of data science techniques in LTC, potentially raising awareness about the variety of opportunities these techniques may provide to this specific care echelon. This review will provide researchers with a useful base for understanding the overall context of data science techniques deployed in LTC. However, the current results need to be viewed in the light of some possible limitations. Firstly, by focusing on PubMed and CINAHL there is a possibility that work published in journals not covered by this database have been omitted. However, PubMed alone already includes more than 33 million citations and is the most used database in the health domain, especially in LTC. Second, in accordance with the guidelines for scoping reviews, a methodological quality assessment of the included studies was not performed [17]. Hence, no conclusion regarding topics such as incomplete data, the effectiveness of the deployed methods (e.g., in terms of the small number of participants included in some of the studies), or the external validation of the included studies can be formulated.

Since the studies in this review discussed the usefulness of data science techniques in qualitative and potential terms, more quantitative and objective measures are needed. To make these techniques become more widespread and integrated in LTC (as they are in, for example, hospital care), research should provide solid evidence that based on the analyses of data by these types of techniques, health decisions, and outcomes can indeed be improved for individual clients. Hence, in order to implement data-informed LTC, more thorough evidence regarding the usefulness of data science in directly or indirectly changing and improving daily care practice is needed. This could, for example, include the use of metrics such as accuracy and sensitivity/specificity. Future analyses could also focus on investigating the current state of evidence regarding the use of data science techniques with data accumulated in a home-based LTC environment. The application of these techniques in a home-based LTC environment (i.e. community-dwelling older adults receiving care) remains unexplored and the findings of such a review may supplement those of this analysis. As LTC for older adults is also provided at home [10], the combined evidence of both reviews would produce an even more complete overview of the use of these techniques. However, given the small number of included studies in this review, the amount of studies focused on data science techniques used for data accumulated in a home-based LTC environment, might also be quite small.

In conclusion, this review presents a useful starting point for future applications of data science techniques in LTC by creating awareness of the ramifications of data and the corresponding analyzing techniques. Currently, in LTC, data science techniques are used for a variety of purposes and are advantageous for the specific study objective in each of the included studies. Although data science presents promising opportunities to reshape the use of data within this area (especially given the rise of new techniques such as ChatGPT) in order to improve the quality and efficiency of care, the low number of identified articles indicates the need for strategies aimed at the effective utilization of data with data science techniques and evidence of its practical benefits.

Abbreviations

AI: artificial intelligence

CINAHL: Cumulative Index to Nursing and Allied Health Literature

CNN: convolutional neural network

COVID-19: coronavirus disease 2019

EHRs: electronic health records

EMR: electronic medical record

LTC: long-term care

MeSH: Medical Subject Headings

ML: machine learning

NH: nursing home

NLP: natural language processing

RCFs: residential care facilities

SVM: support vector machine

Declarations

Author contributions

AH and SA: Conceptualization, Methodology, Formal analysis, Writing—original draft. CH: Formal analysis. HV: Data curation, Writing—review & editing. All authors carefully revised the manuscript and approved the version to be published.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Ethical approval

Not applicable.

Consent to participate

Not applicable.

Consent to publication

Not applicable.

Availability of data and materials

The raw data supporting the conclusions of this manuscript will be made available by the authors, without undue reservation, to any qualified researcher.

Funding

Not applicable.

Copyright

References

1. Ottenheijm S. Big data in de gezondheidszorg. Definitie, toepassingen en uitdagingen [Internet]. Nictiz; 2015 [cited 2023 Nov 1]. Available from: https://nictiz.nl/app/uploads/2023/06/Big_data_in_de_gezondheidszorg.pdf
2. Raja R, Mukherjee I, Sarkar BK. A systematic review of healthcare big data. Hindawi Sci Program. 2020:1–15.
3. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. J Big Data. 2019;6:54.
4. Khanra S, Dhir A, Islam AKMN, Mäntymäki M. Big data analytics in healthcare: a systematic literature review. Enterp Inf Syst. 2020;14:878–912.
5. Kruse CS, Goswamy R, Raval Y, Marawi S. Challenges and opportunities of big data in health care: a systematic review. JMIR Med Inform. 2016;4:e38.
6. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. Int J Med Inform. 2018;114:57–65.
7. Aarts S. Ethical usage of data in long-term care: How do we proceed? In: Wernaart B, editor. Moral design and technology. Wageningen Academic; 2022. pp. 267–82.
8. Aarts S, Daniels R, Hamers J, Verbeek H. Data in de langdurige ouderenzorg. Tijdschr Ouderengeneeskunde. 2020;2020:92–6.
9. Sanford AM, Orrell M, Tolson D, Abbatecola AM, Arai H, Bauer JM, et al. An international definition for "nursing home". J Am Med Dir Assoc. 2015;16:181–4.
10. Stallard E. Long term care for aging populations. In: Quah SR, editor. International encyclopedia of public health (second edition). Oxford: Academic Press; 2017. pp. 447–58.
11. Grabowski DC. Long-term care. In: Culyer AJ, editor. Encyclopedia of health economics. San Diego: Elsevier; 2014. pp. 146–51.
12. Spasova S, Baeten R, Coster S, Ghailani D, Peña-Casas R, Vanhercke B. Challenges in long-term care in Europe. A study of national policies. Brussels: European Union; 2018 Aug. Report No.: KE-01-18-637-EN-N.
13. Feldman B, Martin EM, Skotnes T. Big data in healthcare hype and hope [Internet]. Scribd Inc.; 2024 [cited 2023 Nov 1]. Available from: https://www.scribd.com/document/530050747/Big-Data-in-Healthcare
14. Tak SH, Benefield LE, Mahoney DF. Technology for long-term care. Res Gerontol Nurs. 2010;3:61–72.
15. Krick T, Huter K, Domhoff D, Schmidt A, Rothgang H, Wolf-Ostermann K. Digital technology and nursing care: a scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies. BMC Health Serv Res. 2019;19:400.
16. Loveys K, Prina M, Axford C, Domènec ÒR, Weng W, Broadbent E, et al. Artificial intelligence for older people receiving long-term care: a systematic review of acceptability and effectiveness studies. Lancet Healthy Longev. 2022;3:e286–97.
17. Peters MDJ, Marnie C, Tricco AC, Pollock D, Munn Z, Alexander L, et al. Updated methodological guidance for the conduct of scoping reviews. JBI Evid Synth. 2020;18:2119–26.
18. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA extension for scoping reviews (PRISMA-ScR): checklist and explanation. Ann Intern Med. 2018;169:467–73.
19. Yousefi Zadeh A, Shahbazy M. A review into data science and its approaches in mechanical engineering. arXiv: 2012.15358v1 [Preprint]. 2020 [cited 2023 Nov 1]. Available from: https://doi.org/10.48550/arXiv.2012.15358
20. Sun CLF, Zuccarelli E, Zerhouni EGA, Lee J, Muller J, Scott KM, et al. Predicting coronavirus disease 2019 infection risk and related risk drivers in nursing homes: a machine learning approach. J Am Med Dir Assoc. 2020;21:1533–8.e6.

21. Gannod GC, Abbott KM, Van Haitsma K, Martindale N, Heppner A. A machine learning recommender system to tailor preference assessments to enhance person–centered care among nursing home residents. Gerontologist. 2019;59:167–76.
22. Volrathongchai K, Brennan PF, Ferris MC. Predicting the likelihood of falls among the elderly using likelihood basis pursuit technique. AMIA Annu Symp Proc. 2005;2005:764–8.
23. Shen K, McGarry BE, Grabowski DC, Gruber J, Gandhi AD. Staffing patterns in US nursing homes during COVID-19 outbreaks. JAMA Health Forum. 2022;3:e222151.
24. McGarry BE, Gandhi AD, Syme M, Berry SD, White EM, Grabowski DC. Association of state COVID-19 vaccine mandates with staff vaccination coverage and staffing shortages in US nursing homes. JAMA Health Forum. 2022;3:e222363.
25. Boyce RD, Kravchenko OV, Perera S, Karp JF, Kane-Gill SL, Reynolds CF, et al. Falls prediction using the nursing home minimum dataset. J Am Med Inform Assoc. 2022;29:1497–507.
26. Zhu Y, Song T, Zhang Z, Deng C, Alkhalaf M, Li W, et al. Agitation prevalence in people with dementia in australian residential aged care facilities: findings from machine learning of electronic health records. J Gerontol Nurs. 2022;48:57–64.
27. Jiang T, Qian S, Hailey D, Ma J, Yu P. Text data mining of aged care accreditation reports to identify risk factors in medication management in australian residential aged care homes. Stud Health Technol Inform. 2017;245:892–95.
28. Ambagtsheer RC, Shafiabady N, Dent E, Seiboth C, Beilby J. The application of artificial intelligence (AI) techniques to identify frailty within a residential aged care administrative data set. Int J Med Inform. 2020;136:104094.
29. Withall A, Karystianis G, Duncan D, Hwang YI, Kidane AH, Butler T. Domestic violence in residential care facilities in New South Wales, Australia: a text mining study. Gerontologist. 2022;62:223–31.
30. Suzuki M, Yamamoto R, Ishiguro Y, Sasaki H, Kotaki H. Deep learning prediction of falls among nursing home residents with Alzheimer's disease. Geriatr Gerontol Int. 2020;20:589–94.
31. Tadokoro K, Yamashita T, Kawano S, Sato J, Omote Y, Takemoto M, et al. Immediate beneficial effect of makeup therapy on behavioral and psychological symptoms of dementia and facial appearance analyzed by artificial intelligence software. J Alzheimers Dis. 2021;83:57–63.
32. Tadokoro K, Yamashita T, Sato J, Omote Y, Takemoto M, Morihara R, et al. Chronic beneficial effect of makeup therapy on cognitive function of dementia and facial appearance analyzed by artificial intelligence software. J Alzheimers Dis. 2022;85:1189–94.
33. Cheng M, Cui X. Spatial optimization of residential care facility configuration based on the integration of modified immune algorithm and GIS: a case study of Jing'an district in Shanghai, China. Int J Environ Res Public Health. 2020;17:8090.
34. Huang K, Jiao Z, Cai Y, Zhong Z. Artificial intelligence-based intelligent surveillance for reducing nurses' working hours in nurse–patient interaction: a two-wave study. J Nurs Manag. 2022;30: 3817–26.
35. Wang H, Sheng L, Xu S, Jin Y, Jin X, Qiao S, et al. Develop a diagnostic tool for dementia using machine learning and non-imaging features. Front Aging Neurosci. 2022;14:945274.
36. Lee SK, Ahn J, Shin JH, Lee JY. Application of machine learning methods in nursing home research. Int J Environ Res Public Health. 2020;17:6234.
37. Lee SK, Shin JH, Ahn J, Lee JY, Jang DE. Identifying the risk factors associated with nursing home residents' pressure ulcers using machine learning methods. Int J Environ Res Public Health. 2021;18: 2954.
38. Delespierre T, Denormandie P, Bar-Hen A, Josseran L. Empirical advances with text mining of electronic health records. BMC Med Inform Decis Mak. 2017;17:127.
39. Lapidus N, Carrat F. WTW—an algorithm for identifying "who transmits to whom" in outbreaks of interhuman transmitted infectious agents. J Am Med Inform Assoc. 2010;17:348–53.

40. Fernández-Llatas C, Benedi JM, García-Gómez JM, Traver V. Process mining for individualized behavior modeling using wireless tracking in nursing homes. Sensors (Basel). 2013;13:15434–51.
41. Garcés-Jiménez A, Calderón-Gómez H, Gómez-Pulido JM, Gómez-Pulido JA, Vargas-Lombardo M, Castillo-Sequera JL, et al. Medical prognosis of infectious diseases in nursing homes by applying machine learning on clinical data collected in cloud microservices. Int J Environ Res Public Health. 2021;18:13278.
42. Ritchie LA, Harrison SL, Penson PE, Akbari A, Torabi F, Hollinghurst J, et al. Prevalence and outcomes of atrial fibrillation in older people living in care homes in Wales: a routine data linkage study 2003–2018. Age Ageing. 2022;51:afac252.
43. Hacking C, Verbeek H, Hamers JPH, Sion K, Aarts S. Text mining in long-term care: exploring the usefulness of artificial intelligence in a nursing home setting. PLoS One. 2022;17:e0268281.
44. Reddy BP, O'Neill S, O'Neill C. Explaining spatial accessibility to high-quality nursing home care in the US using machine learning. Spat Spatiotemporal Epidemiol. 2022;41:100503.
45. Lee DS, Ma S, Chu A, Wang CX, Wang X, Austin PC, et al.; CORONA Collaboration. Predictors of mortality among long-term care residents with SARS-CoV-2 infection. J Am Geriatr Soc. 2021;69: 3377–88.
46. Buisseret F, Catinus L, Grenard R, Jojczyk L, Fievez D, Barvaux V, et al. Timed up and go and six-minute walking tests with wearable inertial sensor: one step further for the prediction of the risk of fall in elderly nursing home people. Sensors (Basel). 2020;20:3207.
47. Data science and AI glossary [Internet]. London: The Alan Turing Institute; c2024 [cited 2023 Mar 25]. Available from: https://www.turing.ac.uk/news/data-science-and-ai-glossary
48. Lee I. Big data: dimensions, evolution, impacts, and challenges. Bus Horizons. 2017;60:293–303.
49. Spruit M, Vroon R, Batenburg R. Towards healthcare business intelligence in long-term care: an explorative case study in the Netherlands. Comput Hum Behav. 2014;30:698–707.
50. Roski J, Bo-Linn GW, Andrews TA. Creating value in health care through big data: opportunities and policy implications. Health Aff (Millwood). 2014;33:1115–22.
51. Sun S, Luo C, Chen J. A review of natural language processing techniques for opinion mining systems. Inf Fusion. 2017;36:10–25.
52. Anthony D, Alosoumi D, Safari R. Prevalence of pressure ulcers in long-term care: a global review. J Wound Care. 2019;28:702–9.
53. Cameron EJ, Bowles SK, Marshall EG, Andrew MK. Falls and long-term care: a report from the care by design observational cohort study. BMC Fam Pract. 2018;19:73.
54. Mittelstadt B. The ethics of biomedical 'Big Data' analytics. Philos technol. 2019;32:17–21.