# AI in Nursing Homes

## Utilizing artificial intelligence to enhance data analysis in long-term care for older adults

**Coen W.J.G. Hacking**

# AI in Nursing Homes

Utilizing artificial intelligence to enhance data analysis in long-term care for older adults

**Coen Willem Johannes Gerardus Hacking**

*The research presented in this thesis was conducted at CAPHRI Care and Public Health Research Institute, Department of Health Services Research, Maastricht University. CAPHRI participates in the Netherlands School of Public Health and Care Research CaRe. The research was funded by the Limburg Living Lab in Ageing and Long-Term Care.*

Printing: Grafistar

Cover design: Coen Hacking

# AI in Nursing Homes

## Utilizing artificial intelligence to enhance data analysis in long-term care for older adults

**PROEFSCHRIFT**

*Voor het behalen van de graad van Doctor aan de Universiteit Maastricht, onder gezag van Rector Magnificus, Prof. dr. Pamela Habibović, overeenkomstig met het besluit van het College van Decanen, te verdedigen in het openbaar op donderdag 20 februari 2025, om 16.00 uur.*

door

**Coen Willem Johannes Gerardus Hacking**

**Promotores**
Prof. Dr. J.P.H. Hamers (Universiteit Maastricht)
Prof. Dr. H. Verbeek (Universiteit Maastricht)

**Copromotor**
Dr. S. Aarts (Universiteit Maastricht)

**Beoordelingscommissie**

| | |
|---|---|
| Voorzitter: | Prof. dr. S.M.G. Zwakhalen (Universiteit Maastricht) |
| Leden: | Prof. dr. A.L.A.J. Dekker (Universiteit Maastricht) |
| | Prof. dr. D. Mahr (Universiteit Maastricht) |
| | Dr. O.E. Scharenborg (TU Delft) |
| | Prof. dr. S.U. Zuidema (UMC Groningen) |

*”There’s a lot of ugliness in this world. Disarray.*
*So many of my memories were ugly.*
*But the things I held onto until the end weren’t the ugly ones.*
*I remember the moments that I saw what they were really capable of.*
*Moments of kindness, here and there.*
*They created us, and they knew enough of beauty to teach it to us.*
*Maybe they can find it themselves.*
*There is ugliness in this world. Disarray.*
*But I choose to see the beauty.”*

- Dolores Abernathy, Westworld

# Contents

# CHAPTER 1

## General Introduction

In recent years, there has been a growing demand for resident perspectives in the assessment of the quality of care in nursing homes, as these perspectives are crucial for improving the quality of care [1–3]. Quality of care and quality of life are closely related to the well-being of residents. This demand has led to the collection of large volumes of data, including quantitative data from questionnaires, sensors, and wearables, as well as qualitative data from interviews and observations [2, 4, 5]. Narrative data, such as residents' experiences in nursing homes or during events like the COVID-19 pandemic, have become particularly valuable [2, 5–7]. They provide insight into aspects like engagement, satisfaction, quality of life, and quality of care. Gathering these narratives involves various methods, including written narratives and transcribed interviews. However, analyzing these large volumes of textual data often involves manual analysis, which is labor-intensive and prone to human error and bias, limiting objectivity and reproducibility [8–10].

The volume and variety of data in nursing homes are expanding significantly, driven by Internet of Things (IoT) devices, interviews, questionnaires, and electronic health record (EHR) systems. IoT devices generate continuous data streams regarding residents' health and well-being [4], while EHR systems store vast amounts of clinical and administrative data [11]. This complex data ecosystem poses substantial challenges in data management, analysis, and interpretation. Manual data entry, transcription, and analysis are time-consuming and amplify the downsides of human error and bias [12]. Advanced analytical approaches and robust data management strategies are necessary to integrate these diverse data types effectively to distill their knowledge [13].

This dissertation explores the application of artificial intelligence (AI) methods to streamline data analysis in long-term care for older adults (LTC). LTC encompasses a range of services designed to meet the needs of people with chronic illnesses or disabilities who cannot care for themselves for long periods [14]. Nursing homes are a critical component of LTC, providing 24-hour medical and personal care [15]. In the Netherlands, approximately 115,000 older adults resided in nursing homes as of 2019, according to the Statistics Office of the Netherlands (CBS) [16]. These facilities cater to individuals with complex care needs, such as those suffering from dementia or severe physical impairments. Nursing homes have seen a growing integration of digital technology and data in recent years. Innovations in EHRs, IoT devices, and advanced analytical tools have paved the way for enhanced care delivery and better health outcomes [4, 17]. However, the full potential of these technologies has not been reached. Therefore, the goal of this dissertation is to provide scalable solutions for data collection and analysis, addressing challenges such as manual labor intensity, human error, and bias, to offer deeper insights into the quality of care. In Box 1.1, we provide an overview of key terminologies related to data science and artificial intelligence.

**Box 1.1**: Data Science and Artificial Intelligence.

### Data Science

Data Science is an interdisciplinary field that uses scientific methods, processes, algorithms, and systems to extract knowledge and insights from structured and unstructured data [18]. It combines aspects of statistics, computer science, and information science to analyze and interpret complex data.

### Artificial Intelligence

Artificial Intelligence (AI) is a broad field of computer science focused on creating smart machines capable of performing tasks that typically require human intelligence [19]. AI integrates various subfields, including machine learning and natural language processing, to enable machines to learn from experience and adjust to new unseen inputs.

### Machine Learning

Machine Learning is a subset of AI that enables machines to learn from data such as Bayesian networks and decision trees [20]. Machine learning algorithms use statistical techniques to identify patterns in data and make predictions based on these patterns.

### Deep Learning

Deep Learning is a subset of machine learning that uses neural networks to process large and complex datasets [21]. Neural networks are inspired by the human brain, with interconnected layers of artificial neurons that process information.

### Natural Language Processing

Natural Language Processing (NLP) is a branch of AI that enables machines to understand and interpret human language [22]. NLP algorithms analyze text and speech data, extracting meaning from unstructured data. These algorithms can perform tasks such as sentiment analysis, speech recognition, and text summarization.

## 1 Insufficient Usage of Data in Nursing Homes

Currently, there is a large amount of data in nursing homes that is not (fully) utilized. Efficient, accurate, and automated methods for data collection, entry, and analysis are critical for ensuring the scalability and objectivity of quality-of-care assessments [1, 2]. Significant advancements in data processing technologies and methodologies, along with a reevaluation of current practices, are necessary to achieve a data-informed approach [3, 5].

One of the primary challenges in nursing homes is the lack of resources, particularly in manpower and technological infrastructure. The manual processes involved in data collection, entry, and analysis are time-consuming and labor-intensive [23–25]. AI can automate many of these processes, significantly reducing time and labor [18, 26]. For example, machine learning algorithms can automatically transcribe and analyze interview recordings, alleviating the burden of manual transcription and coding analysis [8, 27]. This automation speeds up the process of analysis, which allows for processing larger data volumes than feasible manually [28, 29].

Traditional methods of data analysis in nursing homes, such as manual coding, are often subjective and prone to human error and bias [9, 10]. They also struggle to handle increasing data complexity and volume. AI provides more objective and sophisticated analytical methods [17, 30], analyzing large datasets more efficiently and accurately, and identifying patterns and insights that might be missed by human analysts [31, 32]. AI technologies, such as machine learning algorithms and natural language processing, can assist in analyzing data, identifying patterns and trends that might be missed in manual analysis, and providing more rapid and accurate insights [19, 20, 22]. The integration of AI in data analysis can help in transforming raw data into meaningful information, guiding decision-making, and improving care strategies in long-term care settings.

## 2 Data Collection Regarding Quality of Care in Nursing Homes

Data collection in nursing homes is crucial for assessing and improving quality of care for residents. Two notable methods from the 'Limburg Living Lab in Aging and Long-Term Care' are the Maastricht Electronic Daily Life Observation tool (MEDLO-tool) and the 'Connecting Conversations' approach [23, 33].

## 2.1 MEDLO-Tool

The MEDLO-Tool is a tablet-based observational instrument designed to assess the daily life of people with dementia in nursing homes [33]. It evaluates activities, physical environment, social interaction, and emotional well-being, capturing both quantitative data (numerical information) and qualitative data (textual narratives). This dual-data approach provides comprehensive insights into residents' engagement, interactions, and mood, offering a holistic view of their daily experiences [33]. Additionally, the MEDLO-Tool employs an ecological momentary assessment (EMA) method, measuring residents' day-to-day life through repeated measurements over time [33]. This tool was developed to overcome the limitations of inter-rater reliability in traditional observational methods, enhancing the validity and reliability of data collection in nursing homes [33]. However, currently the tool is only being used by researchers due to limited emphasis on its user-friendliness and practicality for care professionals. This could be solved by taking the theoretical framework of MEDLO and employing a user-centered design to make it more accessible.

## 2.2 Connecting Conversations

The 'Connecting Conversations' method involves interviews to measure the quality of care from the resident's perspective in nursing homes. This approach includes conversations with a resident, a family member, and a care professional, focusing on positive aspects like satisfaction with care. Questions include "How would you grade your life right now on a ten-point scale?," "What could we do to improve that grade?," and "What does a day for you normally look like?" [2, 23]. This method has proven valuable in gaining diverse perspectives on the quality of care, enhancing understanding, and strengthening learning networks among care organizations [2, 3]. These interviews currently result in large volumes of narrative data, which are manually transcribed and coded for analysis. This process is time-consuming and labor-intensive, limiting the scalability and objectivity of data collection [2, 23]. AI could play a pivotal role in automating these processes. This could make the data collection much more scalable.

## 3 Opportunities for Data in Nursing Homes

Currently, many of the methods used to collect and analyze data in nursing homes are manual and time-consuming. This limits the scalability and objectivity of quality-of-care assessments. AI technologies offer opportunities to automate these processes, enabling more efficient and accurate data collection and analysis [18, 26]. However, to encompass the nuanced and complex nature of quality of care, these technologies must be tailored to the specific needs and challenges in nursing homes [2, 33].

Recent developments in data science are revolutionizing the management and interpretation of large datasets in healthcare, particularly in nursing homes. Techniques such as machine learning, neural networks, and text mining offer promising avenues to improve quality of care by enabling more precise predictions and better understanding of care outcomes. For example, analyzing EHRs can help identify risk factors for adverse outcomes like falls or medication management issues [34, 35]. These techniques can also facilitate personalized care plans by predicting individual risks and needs based on data-driven insights [36, 37]. The implementation of these technologies in nursing homes, although in the early stages, represents a forward-thinking approach to addressing the complex needs of an aging population, promising significant improvements in care delivery and quality [38, 39].

Advancements in NLP have significantly enhanced our ability to interpret and understand text, crucial in healthcare settings where understanding resident narratives and medical reports is vital [2, 5, 23–25]. Improved NLP capabilities aid in extracting accurate and relevant information from medical records and other textual sources, leading to better-informed decisions and quality of care.

Automatic Speech Recognition (ASR) technology has also seen considerable improvements, now capable of transcribing speech with higher accuracy [40–46]. This is particularly significant in healthcare and nursing home settings, where converting spoken words from interviews, doctor's notes, or group discussions into text can be a labor-intensive process [2, 17]. The enhanced ASR technology can accurately transcribe these conversations, even in noisy environments or with speakers having diverse accents. This advancement not only saves time but also ensures that valuable qualitative data is captured more efficiently and accurately, contributing to more comprehensive and person-centered care.

## 4 Aim

This thesis aims to leverage AI advancements to improve the quality of care in long-term care (LTC) for older adults. The specific goals are to:

- Develop tools and methodologies to support data collection and analyses in LTC by automating manual processes.
- Investigate the possibilities of AI in analyzing textual data regarding the quality of care.

## 5 Outline

The structure of this study begins with **Chapter 2**, which describes what data science techniques are currently being employed in long-term care for older adults in a scoping review. After this, the thesis is split into two parts. In Part I, the possibilities for automation in data collection and observation are addressed in two chapters. This exploration is deepened in **Chapter 3** through the development of an automatic speech recognition (ASR) model using interview data from nursing homes. **Chapter 4** moves towards a practical application, with the co-design and development of a software application for real-time observations in a nursing home setting. In Part II, the analysis of quality of care data is discussed in two chapters. **Chapter 5** shows where text mining can used to delve into themes related to residents' experiences regarding quality of care in nursing homes. **Chapter 6** presents a comparison between text mining and manual coding methods. Finally, this thesis concludes with a synthesis of the findings, methodological and theoretical considerations, and recommendations for future research and practice in **Chapter 7**.

## Bibliography


[1] J. Pols, "Enacting appreciations: Beyond the patient perspective," *Health Care Analysis*, vol. 13, no. 3, pp. 203–221, 2005.

[2] K. Sion *et al.*, "The feasibility of connecting conversations: A narrative method to assess experienced quality of care in nursing homes from the resident's perspective," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 5118, 2020.

[3] K. Y. Sion *et al.*, "Experienced quality of post-acute and long-term care from the care recipient's perspective–a conceptual framework," *Journal of the American Medical Directors Association*, vol. 20, no. 11, pp. 1386–1390, 2019.

[4] E. Ahmed *et al.*, "The role of big data analytics in internet of things," *Computer Networks*, vol. 129, pp. 459–471, 2017.

[5] M. Murray *et al.*, "Narrative data," *The SAGE handbook of qualitative data collection*, pp. 264–279, 2018.

[6] K. Y. Sion *et al.*, "How to assess experienced quality of care in nursing homes from the client's perspective: Results of a qualitative study," *BMC geriatrics*, vol. 20, no. 1, pp. 1–12, 2020.

[7] V. Club, *The experiences of seniors during the corona/covid-19 crisis*, 2020. [Online]. Available: `https://www.leydenacademy.nl/the-experiences-of-seniors-during-the-corona-covid-19-crisis/`.

[8] A. Strauss and J. Corbin, *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA, 1998.

[9] N. Norris, "Error, bias and validity in qualitative research," *Educational action research*, vol. 5, no. 1, pp. 172–176, 1997.

[10] P. Mackieson, A. Shlonsky, and M. Connolly, "Increasing rigor and reducing bias in qualitative research: A document analysis of parliamentary debates using applied thematic analysis," *Qualitative Social Work*, vol. 18, no. 6, pp. 965–980, 2019.

[11] M. Ross, W. Wei, and L. Ohno-Machado, ""big data" and the electronic health record," *Yearbook of medical informatics*, vol. 23, no. 01, pp. 97–104, 2014.

[12] A. Anderson, J. Kleinberg, and S. Mullainathan, "Assessing human error against a benchmark of perfection," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 11, no. 4, pp. 1–25, 2017.

[13] K. A. Mills, *Big data for qualitative research*. Taylor & Francis, 2019.

[14] L. Shi and D. A. Singh, *Essentials of the US health care system*. Jones & Bartlett Learning, 2022.

[15] S. Bryan *et al.*, *A systematic review of research evidence on: (a) 24-hour registered nurse availability in long-term care, and (b) the relationship between nurse staffing and quality in long-term care*, Jul. 2010.

[16] Statistics Netherlands (CBS), *Aantal bewoners van verzorgings- en verpleeghuizen 2019*, Accessed: 2024-06-16, 2020.

[17] T. Delespierre, P. Denormandie, A. Bar-Hen, and L. Josseran, "Empirical advances with text mining of electronic health records," *BMC medical informatics and Decision Making*, vol. 17, no. 1, pp. 1–15, 2017.

[18] V. Dhar, "Data science and prediction," *Commun ACM*, vol. 56, no. 12, pp. 64–73, 2013.

[19] E. B. Hunt, *Artificial intelligence*. Academic Press, 2014.

[20] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[21] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. MIT press, 2016.

[22] K. Chowdhary and K. Chowdhary, "Natural language processing," *Fundamentals of artificial intelligence*, pp. 603–649, 2020.

[23] K. Sion *et al.*, "The validity of connecting conversations: A narrative method to assess experienced quality of care in nursing homes from the resident's perspective," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 5100, 2020.

[24] A. Scheffelaar, M. Hendriks, N. Bos, K. Luijkx, and S. Van Dulmen, "Protocol for a participatory study for developing qualitative instruments measuring the quality of long-term care relationships," *BMJ Open*, vol. 8, no. 11, 2018.

[25] A. Scheffelaar, M. Janssen, and K. Luijkx, "The story as a quality instrument: Developing an instrument for quality improvement based on narratives of older adults receiving long-term care," *Int J Env Res Pub He*, vol. 18, no. 5, 2021.

[26] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining.," in *Ldv Forum*, Citeseer, vol. 20, 2005, pp. 19–62.

[27] A. Strauss and J. Corbin, *Basics of qualitative research*. Sage publications, 1990.

[28] F. Popowich, "Using text mining and natural language processing for health care claims processing," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 1, pp. 59–66, 2005.

[29] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, "Text mining in healthcare. applications and opportunities," *J Healthc Inf Manag*, vol. 22, no. 3, pp. 52–6, 2008.

[30] M. Hofmann and A. Chisholm, *Text mining and visualization: case studies using open-source tools*. CRC Press, 2016, vol. 40.

[31] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar, and G. Srivastava, "An accurate deep learning model for clinical entity recognition from clinical notes," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3804–3811, 2021.

[32] A. H. Azeemi and A. Waheed, "Covid-19 tweets analysis through transformer language models," *arXiv preprint arXiv:2103.00199*, 2021.

[33] B. De Boer, H. Beerens, S. M. Zwakhalen, F. E. Tan, J. P. Hamers, and H. Verbeek, "Daily lives of residents with dementia in nursing homes: Development of the maastricht electronic daily life observation tool," *International Psychogeriatrics*, vol. 28, no. 8, pp. 1333–1343, 2016.

[34] B. E. McGarry, A. D. Gandhi, M. Syme, S. D. Berry, E. M. White, and D. C. Grabowski, "Association of state covid-19 vaccine mandates with staff vaccination coverage and staffing shortages in us nursing homes," *JAMA Health Forum*, vol. 3, no. 7, e222363, 2022.

[35] T. Jiang, S. Qian, D. Hailey, J. Ma, and P. Yu, "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*, vol. 245, pp. 892–895, 2017.

[36] C. L. F. Sun *et al.*, "Predicting coronavirus disease 2019 infection risk and related risk drivers in nursing homes: A machine learning approach," *J Am Med Dir Assoc*, vol. 21, no. 11, 1533–1538 e6, 2020.

[37] K. Volrathongchia, P. F. Brennan, and M. C. Ferris, "Predicting the likelihood of falls among the elderly using likelihood basis pursuit technique," in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2005, 2005, pp. 764–768.

[38] A. Withall, G. Karystianis, D. Duncan, Y. I. Hwang, A. H. Kidane, and T. Butler, "Domestic violence in residential care facilities in new south wales, australia: A text mining study," *Gerontologist*, vol. 62, no. 2, pp. 223–231, 2022.

[39] M. Suzuki, R. Yamamoto, Y. Ishiguro, H. Sasaki, and H. Kotaki, "Deep learning prediction of falls among nursing home residents with alzheimer's disease," *Geriatr Gerontol Int*, vol. 20, no. 6, pp. 589–594, 2020.

[40] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, vol. abs/2006.11477, 2020. arXiv: 2006.11477.

[41] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *CoRR*, vol. abs/2106.07447, 2021. arXiv: 2106.07447.

[42] S. Chen *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[43] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "Data2vec: A general framework for self-supervised learning in speech, vision and language," *arXiv preprint arXiv:2202.03555*, 2022.

[44] S. Yang *et al.*, "SUPERB: speech processing universal performance benchmark," *CoRR*, vol. abs/2105.01051, 2021. arXiv: `2105.01051`.

[45] J. Li *et al.*, "Recent advances in end-to-end automatic speech recognition," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[46] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised automatic speech recognition: A review," *Speech Communication*, vol. 139, pp. 76–91, 2022.

# CHAPTER 2

## Data Science Techniques to Gain Novel Insights into Quality of Care: A Scoping Review of Long-Term Care for Older Adults

This chapter was published as:

Hendriks, A., Hacking, C., Verbeek, H., & Aarts, S. (2024). Data science techniques to gain novel insights into quality of care: A scoping review of long-term care for older adults. Explorations in Digital Health Technology, 2(1), 67–85.

https://doi.org/10.37349/edht.2024.00012

## Abstract

**Aim**: 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 organisations providing long-term care for older people (LTC) 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 CINAHL was searched using keywords related to data science techniques and LTC. The screening and selection process were carried out by two authors and 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/NLP and support vector machines 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 include studies concluded that these techniques are valuable for its specific purposes.

**Conclusions**: This review reveals the limited use of data science techniques on data accumulated in or by LTC facilities. 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 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.

# 1 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, questionnaires, 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 analysed individually, but can also be combined; velocity relates to the fast rate at which the data is collected and stored, and veracity 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, 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 behaviour, 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 for older adults (LTC) seems to be lagging behind other healthcare areas such as hospitals [7, 8]. Hence, LTC organisations 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 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 we have, 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 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 analyse (large amounts of) data collected in or by LTC organisations 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.

# 2 Material and methods

A scoping review was conducted. Both the recently updated guidelines for scoping reviews by the Joanna Briggs Institute [17] as well as the PRISMA extension for scoping reviews checklist were followed [18].

## 2.1 Search Strategy

PubMed and CINAHL were deployed for relevant studies. The search was conducted in December 2022. MeSH terms, standardized keywords manually assigned by indexers of the National Library of Medicine, were used. Box 2.1 displays the search string that was used.

**Box 2.1**: Query Parameters.

(”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])

## 2.2 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 in 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. (2015) [9]: ”long-term care occurs in a residential facility or care home 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 behavioural 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.

## 2.3 Selection Process

The screening and selection process was carried out by two authors (AH + SA) (see flow-chart): 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. Afterwards, the abstracts of titles which fitted the pre-specified inclusion criteria were screened. Abstracts which 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 which 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.

## 2.4 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, analysing technique, metric used, conclusion, limitations, and whether ethical approval had been obtained (see Table 2.1). The overall findings were reported by means of narrative synthesis based on the two postulated aims. In order to provide a broad overview on this topic, a methodological quality assessment of the included works was not performed, consistent with the methodology of scoping reviews [17].

# 3 Results

The search strategy yielded 1,488 studies. After the screening of titles and of abstracts, 71 studies were read and assessed for eligibility based on a detailed analysis of their full texts (see Figure 2.1). Finally, 27 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-analysing techniques, or being conducted in a setting other than LTC. The selection process is visualized in the flowchart shown in Figure 2.1.

## 3.1 Characteristics of the Included Studies

A detailed overview about the characteristics of each included study is shown in Table 2.1. 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 United States [20–25], four in Australia [26–31], three in Japan [30–32] and China [33–35], two in Korea [36, 37], France [38, 39], and 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 study population varied greatly between publications. About half of the studies reported that they had obtained ethical approval by a review board.

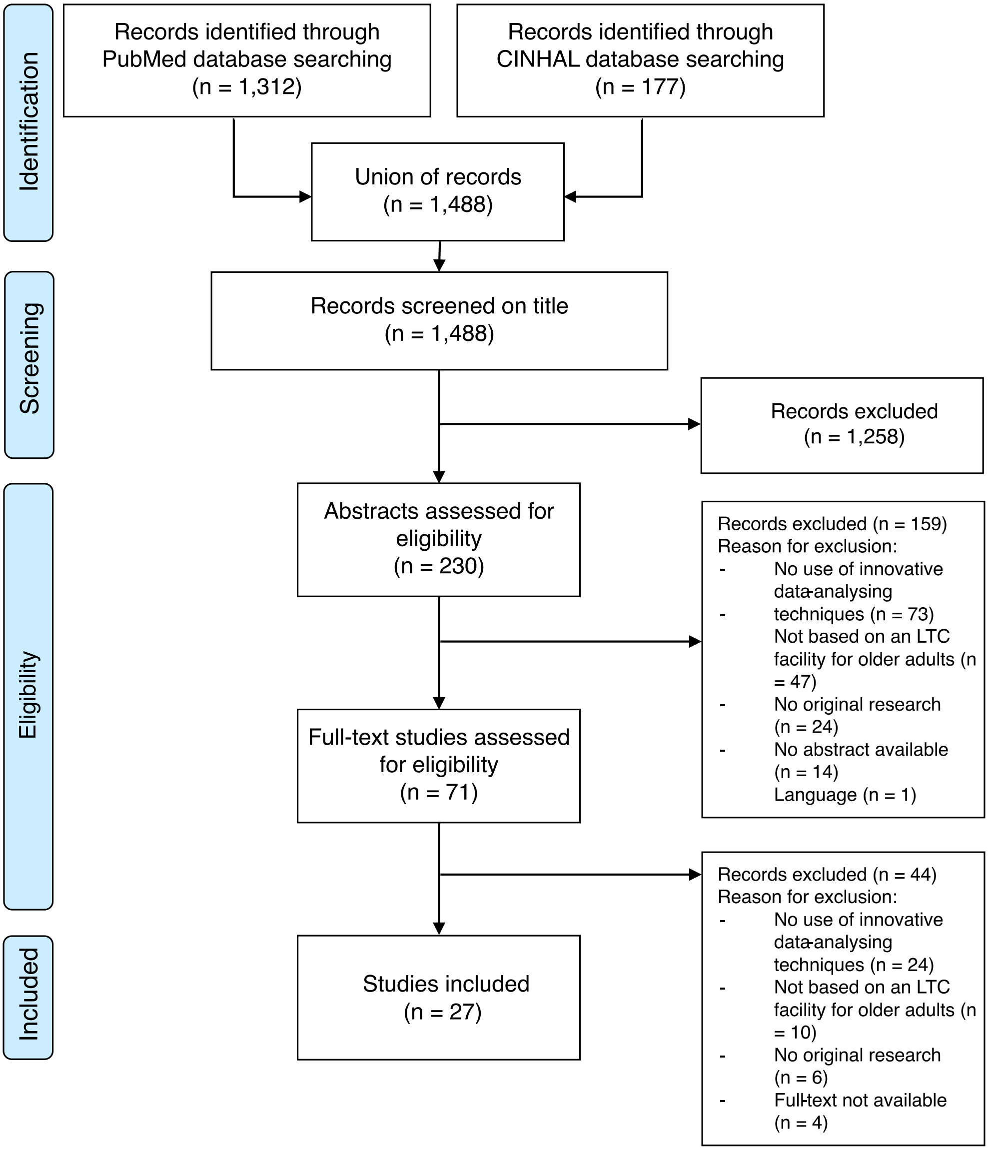



**Figure 2.1:** Flowchart displaying the selection process.

### Data science techniques used and purpose

A diverse set of data-analysing 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 support vector machines (n=4) (i.e. several studies used various methods). In term of metric, accuracy was the most use (n=5); 7 studies did not report a metric. While some studies mentioned to deploy a machine learning technique, other studies refer to the term artificial intelligence or algorithms (i.e. machine learning is a part of artificial intelligence, while algorithms can be considered as part of machine learning, and thus, of artificial intelligence [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 analysing (large) amounts of text. Studies did not report on the use of supervised, unsupervised or semi-supervised methods.

Machine learning 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 electronic medical record data in order to identify risk factors related to medication management. A likelihood basis pursuit data mining technique was employed to predict the likelihood of falls. Artificial intelligence software was used to analyse facial emotion of residents with dementia. Several publications deployed algorithms. One study reported an artificial intelligence algorithm utilized to identify frailty, while another study reported on an algorithm to infer individualized visual models of human behaviour. Modified immune algorithms were used to find the most favourable 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 machine learning techniques (e.g. random forest, logistic regression, naïve bayes etc.) in terms of accuracy and predictive values related to respectively, pressure ulcers, falls and infectious diseases in nursing homes; two of them concluded that a random forest model provided the greatest accuracy and prediction for these outcome measures. Other studies using machine learning techniques were able to quantify and predict the risk of COVID-19 infection in nursing homes, to provide accurate recommendations on potential preferences for a nursing home resident, to map spatial accessibility to high quality nursing home care and to predict falls. A convolutional neural network (CNN) based on deep learning was found to be accurate in fall prediction among nursing home 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 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 analysing large amounts of text, showed that Natural Language Processing (NLP) can be valuable to evaluate agitation in people with dementia, and the identified behaviors can inform improvements in aged care and nursing. A likelihood basis 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 the make-up therapy on cognitive function of female patients. In addition, they reported that AI may be superior to self-reported scales because of its independence om verbal ability and cognition of the patient at hand.

A support vector machine algorithm was found to be capable of accurately identifying frailty among residential care facility 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 residential care facilities in a district of Shanghai. A computerized algorithm provided information on the dynamics of a person-to-person transmitted influenza outbreak in nursing homes, thereby being able to investigate such events. Studies using regression analyses, a more traditional analysing method, showed that COVID-19 outbreaks leaded to adverse outcomes such as reductions in nursing staff levels and that COVID-19 vaccine mandates were associated with increased staff vaccinations.

**Table 2.1:** Results of the data extraction

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 1. Zhu et al. (2022) Australia [26] | To estimate the prevalence of agitated behaviors in people with dementia in nursing homes. | Nursing notes from electronic health records regarding nursing home residents with dementia (n = 3,528). | Rule-based natural language processing (NLP) to detect health terminology, terminology regarding dementia and agitation-related terms. | F-score. | NLP can be valuable to evaluate 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 obtained |
| 2. Wang et al. (2022) China [35] | To develop an early diagnostic tool for Alzheimer's disease using machine learning and non-imaging factors. | Nursing homes in Hangzhou, China (n = 4).Nursing home residents aged 65 or older (n = 654).Community members (n = 1,100). | Logistic regression, support vector machine (SVM), neural network, random forest, extreme gradient boosting (XGBoost), least absolute shrinkage and selection operator (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, quality of care, and reducing associated costs. | Limited study sites. | Ethical approval obtained |
| 3. Huang et al. (2022) China [34] | Using artificial intelligence to improve the time required for nurse–patient interaction. | Nursing home residents (n = 32). | Real-time analysis of streamed video data through a convolutional neural network. | 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 obtained |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 4. Boyce et al. (2022) United States [25] | To develop and validate a novel predictive model that forecasts the risk of falls for nursing home residents 90 days in advance, utilizing data from the Long-Term Care Minimum Dataset and drug therapy records. | Nursing home residents (n = 3985) in 2011, 2012, 2013, 2016-2018 from the University of Pittsburgh Medical Center Senior Communities nursing homes. | A machine learning approach, known as classification and regression tree (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 nursing home staff in identifying individual residents' fall risk within 90 days, potentially leading to targeted interventions, with further research needed to test the model's performance in different health systems and validate its optimal integration into the nursing home clinical workflow. | - 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. (2022) United Kingdom [42] | To determine the prevalence of atrial fibrillation and temporal trends by year of care home entry, and associations between AF and adverse health outcomes including stroke, transient ischaemic attack (TIA), major bleeding, myocardial infarction (MI), cardiovascular hospitalisation and mortality. | Nursing home 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 possible missed due to positive recordings of diagnoses. | Not mentioned |
| 6. Hacking et al. (2022) Netherlands [43] | To explore different text mining methods to analyze the quality of care in a nursing home 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, topic clustering using k-means clustering of word2vec vectors. | Not mentioned. | The study demonstrates the usefulness of text mining to extend our knowledge regarding quality of care in a nursing home setting. With the rise of textual (narrative) data, text mining can lead to valuable new insights for long-term care for older adults. | - 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 obtained |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 7. McGarry et al. (2022) United States [24] | To examine the association of state COVID-19 vaccine mandates with staff vaccination coverage and staffing shortages at nursing homes. | 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 nursing homes. | Not mentioned. | State vaccine mandates for nursing home staff were associated with increased staff vaccine coverage without exacerbating staffing shortages. | - Data self-reported by nursing homes, 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. (2022) United States [23] | 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 nursing homes 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. (2022) Japan [32] | To evaluate the therapeutic effect of makeup therapy. | Female nursing home residents with dementia (n = 34). | Faces were photographed at baseline and after 3 months and was analysed with artificial intelligence software (version of Microsoft Azure Face modified to 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 obtained |
| 10. Reddy, O'Neill & O'Neill (2022) Ireland [44] | 1. 1. To measure and map US county-level spatial accessibility to high quality nursing home care.2. 2. To discover the most relevant socio-demographic variables associated with these levels. | Certified nursing homes 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 southwest and along the Pacific coast. Factors such as 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 were not considered. | NR |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 11. Withall et al. (2022) Australia [29] | 3. 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. | The technique involves using the 'General Architecture for Text Engineering'. It develops rules and uses dictionaries of terms related to mental illnesses, abuse types, and injuries to automatically extract and categorize relevant information. 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 nursing homes, due to potential underreporting. | Not mentioned |
| 12. Tadokoro et al. (2021) Japan [31] | To evaluate the immediate effect of makeup therapy on dementia patients. | Female nursing home residents(n = 36). | Faces were photographed before and after treatment and were analysed with artificial intelligence software (version of Microsoft Azure Face modified to Japanese patients). | P-values, correlation coefficients. | Makeup therapy is a promising non-pharmacological approach for the immediate elevation of behavioural 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 obtained |
| 13. Lee et al. (2021) Canada [45] | To determine predictors associated with 30-day mortality after a positive SARS-CoV-2 test. | Residents in long-term care homes (n = 84.142). | Random survival forest model. | AUC (Area under the ROC curve). | 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 severity of comorbidity was included. | This study did not require approval by a Research Ethics Board and did not require individual consent' |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 14. Garcés-Jiménez (2021) Spain [41] | It was hypothesized that anticipating an infectious disease diagnosis by a few days could significantly improve a patient's well-being an reduce the burden oon emergency health systems. | Residents (n=60) in nursing homes (n=2). | Data was analysed using three ML algorithms: 1) naïve Bayes, 2) filter classifier, 3) 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. (2021) Korea [37] | 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. | Nursing homes (n = 60). Nursing home 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 is a 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 obtained |
| 16. Lee et al. (2020) Korea [36] | To compare different ML methods for predicting falls. | Nursing homes (n = 60). Nursing home 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 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 obtained |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 17. Ambagtsheer et al. (2020) Australia [28] | 1. 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.2. To identify best-performing candidate algorithms. | Residential care facilities(n = 10). Residential care facility residents (n = 592). | A frailty prediction system was designed based on the electronic Frailty Index identification of frailty. Classification algorithms used are K-Nearest Neighbours, Decision Tree and SVM. Each classification algorithm was applied to six cases to find out which best predicted frailty compared to the electronic Frailty Index. | Accuracy. | AI techniques show potential in accurately identifying frailty in RCFs based on data held in administrative databases. A 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 obtained |
| 18. Buisseret et al. (2020) Belgium [46] | 1. To design a method combining clinical tests and motion capture sensors in order to optimize risk of fall prediction.2. To assess the ability of AI to predict risk of fall from solely sensor raw data. | Nursing homes (n = 4). Nursing home residents (n = 73). | A Timed Up and Go test and a six-minute walking test were performed and combined with residents equipped with a homemade wearable inertial 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 on 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 in 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 obtained |
| 19. Cheng & Cui (2020) China [33] | To optimize the configuration of residential care facilities, while considering the demand of three stakeholders (government, elderly, investor), by development of a multi-objective spatial optimization model. | Residential care facilities in the Jing'an District of Shanghai. | A multi-objective spatial optimization model was developed with the goals of maximizing efficiency and equity of RCF configuration, minimizing travel costs of elderly and maximizing profits of investors. | Not mentioned. | The existing configuration of RCFs was shown to be irrational regarding quantity, location and scale. A significant gap is concluded to be present between the service supply of RCFs and the demand of elderly. Overall, the optimization model improved efficiency and equity, reduced the travel costs of 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 |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 20. Sun et al. (2020) United States [20] | To inform about preventive measures for COVID-19 infection by identifying and assessing risk and possible vectors of infection, using a ML approach. | Nursing homes (n = 1146). Nursing home residents(n = NR). | A self-constructed dataset including information on the NHs' facility and community characteristics were used to create predictive features. A tree-based gradient boosting algorithm was used. | ROC (Area under the curve), sensitivity, specificity. | A ML gradient boosting model is useful to quantify and predict the risk of infection in NHs. It provides data-driven support for infection control policies 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. These factors support early identification and management of COVID-19 infections in NHs. | - 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. (2020) Japan [30] | 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). | Nursing home (n = 1) Nursing home 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 was assessed using suitable measures. A CNN was created which focused on multiple complicating factor patterns. 1000 bootstrap datasets were generated for each group using actual sample datasets and used to propose a CNN algorithm. | 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. These predictions may facilitate the development of individualized regimes based on the predicted risk of falling, and thus aid in the prevention of falls. | - 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.<br>- 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 obtained |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 22. Gannod et al. (2019) United States [21] | To explore the application and utility of a recommender system to preference assessment, based on data mining and ML techniques. | Nursing homes(n = 28). Nursing home 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 a ML recommender system, using Apriori algorithm and logistic regression. The algorithm identifies association rules that are based upon data from section F of the MDS 3.0, i.e. the algorithm suggests additional PELI-NH preference items to ask tailored to the individual. | 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 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 cognitive 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 |
| 23. Delespierre et al. (2017) France [38] | 1. To illustrate how text mining of clinical narratives can enhance EMR data.2. To demonstrate the convergence of information between clinical narrative extracted data and EMR data. | Nursing homes(n = 127). Nursing home residents(n = 1015). | 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 were used.- Matching residents with their associated clinical narratives relied on physiotherapy care observations that varied between NHs. | Ethical approval obtained |
| 24. Jiang et al. (2017) Australia [27] | To identify risk factors related to medication management using text data 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 labelled with representative keywords. Records of care homes failing to meet medication management accreditation standards were considered for analysis. Apache OpenNLP was used to extract a word cloud indicating 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 relatively low sample size.- The reports used possessed inadequate details about why failure happened. | NR |

| Authors | Objective | Setting & study population | Analysing technique | Metric used | Conclusion | Limitations | Ethics |
|---|---|---|---|---|---|---|---|
| 25. Fernández-Llatas et al. (2013) Spain [40] | To present a set of algorithms based on process mining that may help professionals to infer and compare individualized visual models of human behaviour. | Nursing home(n = 1).Nursing home residents (n = 9). | The eMotiva process mining framework combining algorithms and visualisation interfaces were 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. The visualisation tool compared and highlighted behaviour patterns. | Not mentioned. | The process mining technology was useful to infer and present individual models to experts, representing human behaviour in a visual and understandable manner. | Limited number of cases used for observation. | NR |
| 26. Lapidus & Carrat (2010) France [39] | To develop a computerised algorithm able to identify the likeliest transmission paths during a person-to-person transmitted illness outbreak. | Nursing home residents(n = NR). | A computerised algorithm was built using information about 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. The algorithm was also applied to real data about an influenza outbreak in NHs. | 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. Volrathongchia et al. (2005) United States [22] | To evaluate the application of a KDD process using a Likelihood Basis 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 Basis 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 Basis 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 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 Basis Pursuit technique required that there is little correlation among the predictor variables.- Only five or six variables were included within the Likelihood Basis Pursuit technique. | Ethical approval obtained |

# 4 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 organisations providing LTC to analyse 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 organisations [48]. In addition, LTC organisations include a variety of data sources that all collect information in various forms: e.g. medical data in electronic health records, 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 on 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 scientist and computing experts (e.g. data scientists) may help to address the challenges within this specific care echelon.

Several different data-analysing 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/natural language processing, a process aimed and analysing 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 electronic health records (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 quality of life. These data would be perfectly suitable for data exploration using text-mining/NLP. For example, text fields in EHRs can be analysed: 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 machine learning, artificial intelligence and algorithms seem to be used interchangeable and for the same purpose: to describe the method that was deployed (e.g. some studies report to deploy an ′artificial intelligence method′, others report to use machine learning or ′a powerful algorithm to predict′). While the terms artificial intelligence, machine learning 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 used 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 are 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 organisations [12, 52, 53]. Hence, these studies underscore that novel data-analysing 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 long-term care, 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 analysing 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 able 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**

LTC: Long-term care for older adults
EHR: Electronic health records
NLP: Natural language processing
CNN: Convolutional Neural Networks

# Bibliography


[1] Nictiz, "Big data in healthcare (rep.)," Tech. Rep., 2017.

[2] R. Raja, I. Mukherjee, and B. K. Sarkar, "A systematic review of healthcare big data," *Hindawi Scientific Programming*, pp. 1–15, 2020.

[3] S. Dash, S. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: Management, analysis and future prospects," *Journal of Big Data*, vol. 6, no. 54, 2019.

[4] S. Khanra, A. Dhir, A. K. M. N. Islam, and M. Mäntymäki, "Big data analytics in healthcare: A systematic literature review," *Enterprise Information Systems*, vol. 14, no. 7, pp. 878–912, 2020.

[5] C. S. Kruse, R. Goswamy, Y. Raval, and S. Marawi, "Challenges and opportunities of big data in health care: A systematic review," *JMIR Med Inform*, vol. 4, no. 4, e38, 2016.

[6] N. Mehta and A. Pandit, "Concurrence of big data analytics and healthcare: A systematic review," *Int J Med Inform*, vol. 114, pp. 57–65, 2018.

[7] S. Aarts, "Ethical usage of data in long-term care: How do we proceed?" In *Moral Design and Technology*. Wageningen: Wageningen Academic, 2022, pp. 267–282.

[8] S. Aarts, R. Daniels, H. Verbeek, and J. Hamers, "Data in de langdurige zorg," *Tijdschrift voor Ouderengeneeskunde*, 2020.

[9] A. M. Sanford *et al.*, "An international definition for "nursing home"," *J Am Med Dir Assoc*, vol. 16, no. 3, pp. 181–4, 2015.

[10] E. Stallard, "Long term care for aging populations," in *International Encyclopedia of Public Health*. Oxford: Oxford Academic Press, 2017.

[11] D. Grabowski, "Long-term care," in *Encyclopedia of Health Economics*. San Diego: Elsevier, 2014, pp. 146–151.

[12] S. Spasova, R. Baeten, S. Coster, D. Ghailani, R. Peña-Casas, and B. Vanhercke, "Challenges in long-term care in europe. a study of national policies," European Commission, Tech. Rep., 2018.

[13] B. Feldman and E. Martin, "Big data in healthcare - hype and hope," *Computer Science*, pp. 122–125, 2012.

[14] S. H. Tak, L. E. Benefield, and D. F. Mahoney, "Technology for long-term care," *Res Gerontol Nurs*, vol. 3, no. 1, pp. 61–72, 2010.

[15] T. Krick, K. Huter, D. Domhoff, A. Schmidt, H. Rothgang, and K. Wolf-Ostermann, "Digital technology and nursing care: A scoping review on acceptance, effectiveness and efficiency studies of informal and formal care technologies," *BMC Health Serv Res*, vol. 19, no. 1, p. 400, 2019.

[16] K. Loveys *et al.*, “Artificial intelligence for older people receiving long-term care: A systematic review of acceptability and effectiveness studies,” *Lancet Healthy Longev*, vol. 3, no. 4, e286–e297, 2022.

[17] M. D. J. Peters *et al.*, “Updated methodological guidance for the conduct of scoping reviews,” *JBI Evid Synth*, vol. 18, no. 10, pp. 2119–2126, 2020.

[18] A. C. Tricco *et al.*, “Prisma extension for scoping reviews (prisma-scr): Checklist and explanation,” *Ann Intern Med*, vol. 169, no. 7, pp. 467–473, 2018.

[19] A. Y. Zadeh and M. Shahbazy, “A review into data science and its approaches in mechanical engineering,” *arXiv preprint arXiv:2012.15358*, no. 5th International conference, 2020.

[20] C. L. F. Sun *et al.*, “Predicting coronavirus disease 2019 infection risk and related risk drivers in nursing homes: A machine learning approach,” *J Am Med Dir Assoc*, vol. 21, no. 11, 1533–1538 e6, 2020.

[21] G. C. Gannod, K. M. Abbott, K. Van Haitsma, N. Martindale, and A. Heppner, “A machine learning recommender system to tailor preference assessments to enhance person-centered care among nursing home residents,” *Gerontologist*, vol. 59, no. 1, pp. 167–176, 2019.

[22] K. Volrathongchia, P. F. Brennan, and M. C. Ferris, “Predicting the likelihood of falls among the elderly using likelihood basis pursuit technique,” in *AMIA annual symposium proceedings*, American Medical Informatics Association, vol. 2005, 2005, pp. 764–768.

[23] K. Shen, B. E. McGarry, D. C. Grabowski, J. Gruber, and A. D. Gandhi, “Staffing patterns in us nursing homes during covid-19 outbreaks,” *JAMA Health Forum*, vol. 3, no. 7, e222151, 2022.

[24] B. E. McGarry, A. D. Gandhi, M. Syme, S. D. Berry, E. M. White, and D. C. Grabowski, “Association of state covid-19 vaccine mandates with staff vaccination coverage and staffing shortages in us nursing homes,” *JAMA Health Forum*, vol. 3, no. 7, e222363, 2022.

[25] R. D. Boyce *et al.*, “Falls prediction using the nursing home minimum dataset,” *J Am Med Inform Assoc*, vol. 29, no. 9, pp. 1497–1507, 2022.

[26] Y. Zhu *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*, vol. 48, no. 4, pp. 57–64, 2022.

[27] T. Jiang, S. Qian, D. Hailey, J. Ma, and P. Yu, “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*, vol. 245, pp. 892–895, 2017.

[28] R. C. Ambagtsheer, N. Shafiabady, E. Dent, C. Seiboth, and J. Beilby, "The application of artificial intelligence (ai) techniques to identify frailty within a residential aged care administrative data set," *Int J Med Inform*, vol. 136, p. 104 094, 2020.

[29] A. Withall, G. Karystianis, D. Duncan, Y. I. Hwang, A. H. Kidane, and T. Butler, "Domestic violence in residential care facilities in new south wales, australia: A text mining study," *Gerontologist*, vol. 62, no. 2, pp. 223–231, 2022.

[30] M. Suzuki, R. Yamamoto, Y. Ishiguro, H. Sasaki, and H. Kotaki, "Deep learning prediction of falls among nursing home residents with alzheimer's disease," *Geriatr Gerontol Int*, vol. 20, no. 6, pp. 589–594, 2020.

[31] K. Tadokoro *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*, vol. 83, no. 1, pp. 57–63, 2021.

[32] K. Tadokoro *et al.*, "Chronic beneficial effect of makeup therapy on cognitive function of dementia and facial appearance analyzed by artificial intelligence software," *J Alzheimers Dis*, vol. 85, no. 3, pp. 1189–1194, 2022.

[33] M. Cheng and X. Cui, "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*, vol. 17, no. 21, 2020.

[34] K. Huang, Z. Jiao, Y. Cai, and Z. Zhong, "Artificial intelligence-based intelligent surveillance for reducing nurses' working hours in nurse-patient interaction: A two-wave study," *J Nurs Manag*, vol. 30, no. 8, pp. 3817–3826, 2022.

[35] H. Wang *et al.*, "Develop a diagnostic tool for dementia using machine learning and non-imaging features," *Front Aging Neurosci*, vol. 14, p. 945 274, 2022.

[36] S. K. Lee, J. Ahn, J. H. Shin, and J. Y. Lee, "Application of machine learning methods in nursing home research," *Int J Environ Res Public Health*, vol. 17, no. 17, 2020.

[37] S. K. Lee, J. H. Shin, J. Ahn, J. Y. Lee, and D. E. Jang, "Identifying the risk factors associated with nursing home residents' pressure ulcers using machine learning methods," *Int J Environ Res Public Health*, vol. 18, no. 6, 2021.

[38] T. Delespierre, P. Denormandie, A. Bar-Hen, and L. Josseran, "Empirical advances with text mining of electronic health records," *BMC Med Inform Decis Mak*, vol. 17, no. 1, p. 127, 2017.

[39] N. Lapidus and F. Carrat, "Wtw–an algorithm for identifying "who transmits to whom" in outbreaks of interhuman transmitted infectious agents," *J Am Med Inform Assoc*, vol. 17, no. 3, pp. 348–53, 2010.

[40] C. Fernandez-Llatas, J. M. Benedi, J. M. Garcia-Gomez, and V. Traver, "Process mining for individualized behavior modeling using wireless tracking in nursing homes," *Sensors (Basel)*, vol. 13, no. 11, pp. 15 434–51, 2013.

[41] A. Garces-Jimenez *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*, vol. 18, no. 24, 2021.

[42] L. A. Ritchie *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*, vol. 51, no. 12, 2022.

[43] C. Hacking, H. Verbeek, J. P. H. Hamers, K. Sion, and S. Aarts, "Text mining in long-term care: Exploring the usefulness of artificial intelligence in a nursing home setting," *PLoS One*, vol. 17, no. 8, e0268281, 2022.

[44] B. P. Reddy, S. O'Neill, and C. O'Neill, "Explaining spatial accessibility to high-quality nursing home care in the us using machine learning," *Spat Spatiotemporal Epidemiol*, vol. 41, p. 100 503, 2022.

[45] D. S. Lee *et al.*, "Predictors of mortality among long-term care residents with sars-cov-2 infection," *J Am Geriatr Soc*, vol. 69, no. 12, pp. 3377–3388, 2021.

[46] F. Buisseret *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)*, vol. 20, no. 11, 2020.

[47] T. A. T. Institute, "Data and ai glossary," vol. 2023, 2023.

[48] I. Lee, "Big data: Dimensions, evolution, impacts, and challenges," *Business Horizons*, vol. 60, pp. 293–303, 2017.

[49] M. Spruit and R. Vroon, "Towards healthcare business intelligence in long-term care. an explorative case study in the netherlands," *Computers in Human Behavior*, vol. 30, pp. 698–707, 2014.

[50] J. Roski, G. W. Bo-Linn, and T. A. Andrews, "Creating value in health care through big data: Opportunities and policy implications," *Health Aff (Millwood)*, vol. 33, no. 7, pp. 1115–22, 2014.

[51] . S. Sun, C. Luo, and J. Chen, "A review of natural language processing techniques for opinion mining systems," *Computer Science*, pp. 10–25, 2017.

[52] D. Anthony, D. Alosoumi, and R. Safari, "Prevalence of pressure ulcers in long-term care: A global review," *J Wound Care*, vol. 28, no. 11, pp. 702–709, 2019.

[53] E. J. Cameron, S. K. Bowles, E. G. Marshall, and M. K. Andrew, "Falls and long-term care: A report from the care by design observational cohort study," *BMC Fam Pract*, vol. 19, no. 1, p. 73, 2018.

[54] B. Mittelstadt, “The ethics of biomedical ‘big data’ analytics.,” *Philosophy and technology*, vol. 32, no. 1, pp. 17–21, 2019.

# PART I

## Gathering Quality of Care Data

This part discusses the development of tools and methodologies to support data collection and analyses in LTC by automating manual processes.

# CHAPTER 3

## The Development of an Automatic Speech Recognition Model Using Interview Data From Long-Term Care for Older Adults

This chapter was published as:

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2022). The development of an automatic speech recognition model using interview data from long-term care for older adults. Journal of the American Medical Informatics Association, 30(3), 411-417.

https://doi.org/10.1093/jamia/ocac241

## Abstract

**Objective**: In long-term care for older adults, interviews are often used to collect client perspectives. These interviews are often recorded and transcribed verbatim; a time-consuming and tedious task. Automatic speech recognition (ASR) could provide a solution. However, current ASR systems aren't effective for certain demographical groups. This study aims to show how data from specific demographical groups such as older adults and/or people with accents can be used to develop an effective ASR.

**Materials and Methods**: An initial ASR model was developed using Mozilla Common Voice dataset. Audio and transcript data (34h) from interviews with residents, family, and care professionals on quality of care, were used to improve the ASR model. Interview data was continuously processed to reduce the word error rate (WER).

**Results**: Due to background noise and mispronunciations, an initial ASR model had a WER of 48.3% on the interview data. After processing interview data, the WER was reduced to 17.3%. When tested on speech data from residents, family and care professionals a median WER of 22.1% was achieved: residents displaying the highest WER (22.7%). The resulting ASR model was shown to be at least 6x as fast as compared to manual transcription.

**Discussion**: The current method was shown to decrease the WER, verifying its efficacy. Moreover, using local transcription of audio can be beneficial to the privacy of interview participants.

**Conclusion**: The current study has shown that interview data stemming from a long-term care setting can be effectively used to reduce the WER of an ASR model. In daily practice, researchers reported that while ASR model didn't produce perfect results, it saved a lot of time during transcription.

# 1 Introduction

In recent years, client perspectives have become increasingly important in health care research [1]. For example, in long-term care for older adults (LTC), the perspectives of residents, family, and care professionals have become a valuable source of information about the quality of care and quality of life [2, 3]. To assess these perspectives, often narrative data is collected. Narrative data can be defined as stories that describe the experiences and emotions in someone's life [4]. Narrative data can be collected as text (e.g., through open text fields in questionnaires) or as audio, acquired by means of recorded interviews [2, 5]. To be able to objectively analyze the latter type of data, the corresponding audio recordings are transcribed verbatim (i.e., written out literally into text) [2].. Currently, in LTC, the transcription of interview recordings is conducted manually by researchers; however, as the number of interviews increases, manual transcription can become very time-consuming and costly. An alternative approach could rely on automatic speech recognition (ASR). ASR is the process of automatically transcribing speech data into written text [6].

Various studies have shown that English ASR models can achieve a word-error-rate (WER) below 4%, meaning that of every 100 predicted words, less than 4 words are incorrectly transcribed [6–12]; however, error rates may be higher when applied in real-world scenarios. Research has shown that for non-native English speakers, the word error rates can be considerably higher, resulting in error rates above 20% [13, 14]. Moreover, demographic factors such as a person's age, gender, or accent can influence speech features (e.g., pitch, clarity), which increases error rates. For example, older adults are more likely to suffer from speech impairments, possibly leading to audio recordings that are more difficult to recognize by ASR systems [15, 16].

To improve the accuracy, and thus reduce WERs of current ASR approaches, a more diverse dataset is required, such as data describing a higher diversity in types of voices and accents. More diverse data can lead to an ASR model that is better at recognizing speech irrespective of accent, age, or other demographic factors. To increase the diversity of data, a novel data collection method should be employed. As interview data is abundantly available in health care research, data could be collected from interview recordings with corresponding transcripts. For example, such data could include interviews with residents, formal and informal caregivers, and healthcare managers; however, extensive preprocessing is required to be able to use interview data for developing an ASR model with a lower WER. For example, full-length interview recordings are too long for training an ASR model. Thus, audio should be split into segments of an appropriate length before they can be used for improvement of an ASR model [7, 8].

Current ASR models still display high WERs for some languages, such as Dutch, or for speech stemming from people with specific accents. Therefore, this study aims to investigate how data from specific demographic resources could be used to develop an ASR model that achieves lower WERs regardless of age, accent, or other speech features. Additionally, we aim to investigate the time-effectiveness of the method, compared to the current baseline of manual transcription.

# 2 Methods

To develop an ASR model, speech data was collected from existing datasets, such as the Mozilla Common Voice NL (MCV-NL) corpus, as well as interview data from LTC for older adults. Preprocessing of interview data was required. The steps of the iterative ASR model training process are outlined in Figure 3.1: (a) an initial model was trained using existing datasets, (b) transcripts were split into segments of an appropriate length and labeled (i.e., text was predicted from speech data) using the previously trained model, (c) interview transcripts are used to provide automatic text corrections, (d) the model was further fine-tuned using additional data, and finally (e) training of the ASR model was finished if the WER could no longer be decreased; otherwise, steps b through d were repeated until the WER remained the same. After steps a and d, the current state of the model was evaluated by comparing the segments transcribed using the ASR model to the ones that were manually transcribed.

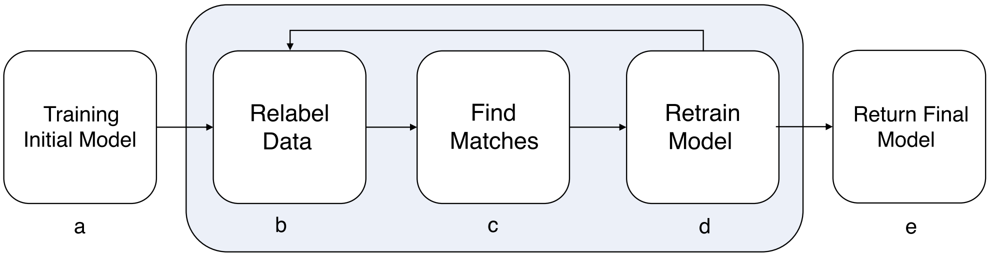



**Figure 3.1:** An overview of the iterative ASR model training process.

## 2.1 Sample and Participants

A total of 232 interviews were conducted at 5 different LTC organizations in the south of the Netherlands over a span of two years. These interviews about the quality of care were conducted with residents, family members, and care professionals. Various types of wards were included, including wards for older people with dementia. In total, 50% of interviews were conducted in the regional dialect (i.e., Limburgish). The medical ethical committee of Zuyderland (the Netherlands) approved the study protocol (17-N-86). Information about the study was provided to all interviewers, residents, family members, and caregivers by an information letter. All participants provided written informed consent; residents with legal representatives gave informed consent themselves (as well as their legal representatives) before and during the conversations. See Sion et al. for more information [2].

## 2.2 Data collection

Data was collected from various sources, including the "Corpus Gesproken Nederlands" (CGN) dataset, the Dutch 'Mozilla Common Voice' dataset (MCV), and the audio recordings and transcripts from interviews with nursing home residents, their family members, and care professionals. The data from interviews was used to construct a dataset for additional fine-tuning of the ASR model. The distribution of all datasets is shown in Table 3.1.

**Table 3.1:** The various data sources used for creating an ASR model. For each dataset, a description is given of its sources, whether preprocessing was required and how many hours of data are included.

| Dataset | Preprocessing | Hours audio | Uttered words |
|---|---|---|---|
| CGN: a dataset of general Dutch data, collected from audiobooks, television, radio, and other sources | No | 960 | 1,855,763 |
| MCV-NL (Mozilla Common Voice Dutch) (Version 6.1) [17]: recorded fragments collected through crowd sourcing with a wide variety of different voices. | No | 61 | 187,559 |
| Connecting Conversations: audio collected from interviews about quality-of-care in nursing home with residents, their family members, and care professionals | Yes | 34 | 99,582 |
| Interviews with informal caregivers, formal caregivers, and health care managers | Yes | 27 | 83,239 |
| Miscellaneous interviews from various individuals about long-term care for older adults | No | 20 | 62,286 |

## 2.3 Model

In the following sections, several components are discussed regarding the preparation of the ASR model. This includes a description of how preprocessing was applied and how the acoustic and language models were constructed. The acoustic and language models are both parts of the deployed ASR model.

### Preprocessing

The MCV dataset didn't require any preprocessing since it entails a dataset for training and includes dedicated train, validation, and test sets [8]. The CGN dataset includes transcripts with time codes for the start and end of each utterance. These time codes describe which part of the audio corresponds with a certain piece of text. Therefore, the time codes were used to split the audio into smaller audio segments with the corresponding text, suitable for deep learning.

For the interview data, only full-length audio files and full-length transcripts exist without timestamps. This would cause issues when trying to train an ASR model as deep learning requires small segments to learn from. Splitting up the speech data is necessary to be able to fit it into memory and avoid exceeding the maximum size (i.e. token limit) the model can process [7, 8]. However, splitting mid-sentence can increase the WER of the ASR model. Therefore, it is important to optimise the time stamps at which the audio is split. To achieve this, several steps were applied, similar to the 'Noisy Student' method [18]. An initial model $M_0$ was trained without the interview data. The interviews were then split into smaller segments $S$ (i.e. smaller than 20 seconds). Then these segments $S$ were transcribed using $M_0$, leading to predictions $P$. These predictions were then aligned with the pre-existing manual transcripts using the Smith-Waterman algorithm. Smith-Waterman is a local alignment algorithm, which can be used to find text similar to predictions in the manual transcripts. If the similarity was bellow p = 90% the text segment discarded. Otherwise, if the text of both was very similar or the same, we assumed that the prediction was correct. Then the manual segments were used as corrections $P'$. Finally, $M_0$ is fine-tuned using $S$ and $P'$. This process is repeated until the word error rate remains the same between iterations (i.e. no additional improvement is possible).

For ASR, it is important that audio segments used for training are short enough to for the model, which may only support speech segments of a limited length [8, 9]. Therefore, some of the audio files required preprocessing to facilitate the development of an ASR model, in the form of splitting and text alignment. Segments that are too short don't contain the information necessary to develop a speech recognition model, while segments that are too long may become more difficult to align with the corresponding part in the audio and may require too many computational resources. It is therefore essential that interview recordings are split into segments; however, splitting the audio at static intervals may result into splitting mid-word or mid-sentence, which could increase the WER [19, 20]. Therefore, the audio of the interviews was split between occurring silences. This process was conducted iteratively, initially splitting on a minimum silence length of 250 ms and repeating this each time with half the minimum silence length until all segments are of the correct length.

### Acoustic model

An acoustic model predicts a token for every n = 10ms of the audio (with a small amount of overlap), the probabilities for every possible token in the vocabulary (e.g. all the characters in the alphabet). A token is an index representation of a character (e.g. 0 = ‘a’, 1 = ‘b’) [8]. The acoustic model returns a probability for each possible token and the token with the highest probability is selected (e.g. given P(0) = 0.9 and P(1) = 0.8, than token 1 is selected). This results in a list of tokens, which are then decoded into a piece of text. For example, a 0 is turned back into the letter ‘a’.

In this study, the HuBERT methodology was used for the acoustic model [8]. This methodology shows that a model can be pre-trained using a convolutional neural network (CNN) on a large audio dataset without the need for corresponding transcripts. This pre-training allows the model to learn what speech features are important to make correct text predictions [8]. An English version of the model was used as a starting point for developing a Dutch model.

### Language model

When transcribing text, it’s not always possible to understand every single word, neither for speech recognition software nor for humans (NOTE: kan worden getoond in de transcripten, er staan vaak notities dat iets onverstaan baar is). It is often necessary to rely on language information, such as knowledge about what sentences are grammatically correct, to understand imperfect speech data. For automatic speech recognition this is achieved by using a language model. Language models are models that can encode the probability of (sequences of) words [21, 22]. For example, a sentence such as ‘my dog likes the park’ is more probable to occur than ‘the park likes my dog’. As another example, given a sentence such as ‘my dog likes the [blank]’, a language model can predict which word best fits the blank. By leveraging the language information, it becomes possible to understand sentences, even when not all words could be understood from the speech data. A language model can thus be used to correct errors using language information (e.g. words that are similar or ambiguous based on their pronunciation) [8]. For example, the words ‘then’ and ‘than’ are pronounced exactly the same, but have different spellings and functions in a written sentence. In this study, RobBERT, a pre-existing language model was used [22].

### 2.4 Analyses

To establish whether the proposed method is effective in fine-tuning an ASR model for the purpose of transcribing speech data in LTC for older adults, the method was evaluated in three different ways. Firstly, the accuracy of the model was evaluated by measuring if there's a significant decrease in WER. The WER was evaluated at the initial stage and at every successive iteration. The WER was obtained by evaluating the MCV-NL validation set. If the measured WER values decrease over time, it is an indication that the method may be effective. To ensure that the change in WER was only due to an improvement in the accuracy of the model, the validation set remained the same during all iterations of the process. To further test the effectiveness, interview transcript predictions of the final ASR model were compared to manual transcripts in terms of overlapping words. Secondly, the time-effectiveness of the resulting model was analyzed to establish if using the model, in practice, could reduce time required to transcribe. Lastly, an error analysis related to insertions (i.e. any character that was wrongfully added to the transcript), deletions (i.e. a character is missing from the transcript) and substitutions (i.e. a character was wrongfully exchanged for another) [23]. The ratio of individual error types compared to the total amount of errors were calculated.

## 3 Results

### 3.1 Accuracy

At first, a model was trained using the CGN dataset. This model achieved a WER of 4.6% on its own validation set. A validation set is used to validate the accuracy of a model. However, when validated on speech data from LTC (i.e., the interviews), this model only achieved a WER of 48.3%. A second model, trained only on the MCV-NL dataset, achieved a WER of 23.6% on its own validation set. When applied to speech data from long-term care, this model achieved a WER of 37.6%. This indicates that the first model might have been trained too strongly on Dutch with a typical accent. The second model went through several iterations of the proposed method (see Table 3.3). Every iteration improves the WER of the model and increases the number of correctly recognized segments.

For all segments in the interview data, the WER was calculated. The distribution of these scores has been visualized as boxplots for each of the different groups (i.e., residents, family, and care professionals) in Figure 3.2. These results show that audio from residents show the highest WERs, which means that the speech of this group is most difficult to recognize, with a median of 23.8% correctly aligned text. Family members show a better WER with a median of 22.6%, and care professionals have the best ratios with a median of 21.9%. All groups are shown to have some outliers (i.e., individual cases that are far away from other

**Table 3.3:** The number of correctly recognized interview samples included in each iteration and the resulting word-error-rate (WER) of the validation set.

| Iteration | # Samples | WER (%) |
|---|---|---|
| 0 (Initial) | 0 | 22.9 |
| 1 | 1137 | 21.6 |
| 2 | 2253 | 20.3 |
| 3 | 3429 | 19.4 |
| 4 (Final) | 4567 | 17.3 |

data points) on the high and low end; none of the interviews surpass a WER of 10.0%. The WER on data from the interviews went from 48.3% after training on CGN, to 37.6% after iteration 0 (i.e. where only data from MCV-NL was used) and then to 23.7% after iteration 4 (i.e. where data from interviews has been used four times).

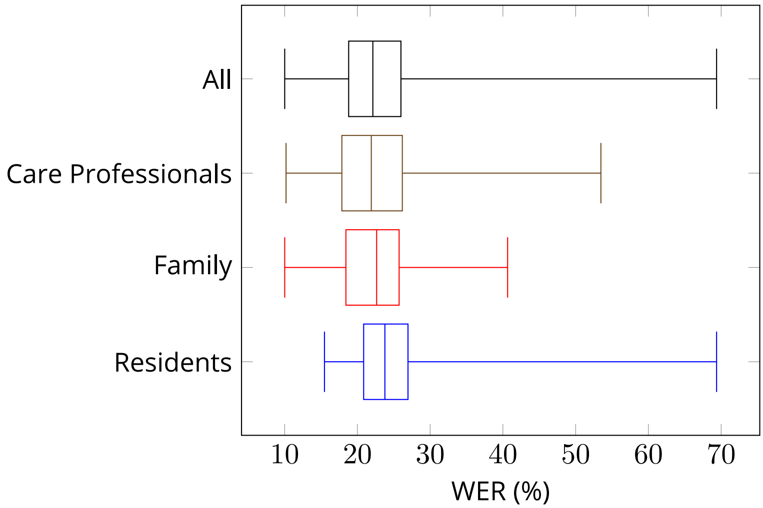



**Figure 3.2:** Results of the ASR model when applied to interview data. For each group (residents, family and care professionals) the word error rates are plotted (x axis).

## 3.2 Time effectiveness

When manually transcribing speech data (i.e., interviews), the speed of transcribing is limited to various factors, such as the playback speed of the audio. At 1x speed (i.e., not faster than recorded), it would be possible to transcribe at most 60 minutes (of audio) per hour (of transcribing). It could be argued that this is an upper limit, as manual transcription may require listening to certain parts of the audio multiple times. With this same logic in mind, when playing the audio back at 10x speed, at most 600 minutes audio per hour could be transcribed manually. No person could type this fast, and this is still significantly slower compared to transcription conducted through ASR. In our experiments, transcription speeds were assessed using an Nvidia RTX 2060. Using this hardware configuration, an average transcription speed of 3,618 minutes audio per hour was reached; however, as the validated output of this transcription had a WER of 22.6%, it should be considered that manual transcription is still required to achieve a 100% correct transcript. For example, in cases where there is a lot of noise in an audio segment or where people are talking at the same time, manual transcription would still be needed.

Various researchers from the Limburg Living Lab in Ageing and Long-Term Care in Limburg, The Netherlands were able to use the method by means of a web application. The researchers reported that the application was of added value; the tediousness of verbatim transcription is reduced to making small corrections to the automatically generated transcripts. Some researchers noticed that certain accents could still result in transcription errors, which required more manual correction. "not everything is transcribing correctly, but in most cases $I$ can correct the misspellings without listing to the audio". Interestingly, researchers sometimes reported that even they could not always transcribe the audio, the primary reason being background noise and a lack of clarity in the voice of an interview participant.

## 3.3 Error analysis

The error analysis (show in Table 3.4) indicates that insertions are responsible for the majority of the errors; in the majority of cases, the ASR model inserted an additional character compared to the manual transcript. In contrast, a relatively low ratio of substitutions is observed (i.e., a character is wrongfully replaced by another character). The ratio of deletions may indicate that words are not always recognized.

**Table 3.4:** The distribution of error types of the ASR model from manual to predication *.

| | Total | Residents | Family | Care Professionals |
|---|---|---|---|---|
| Insertions (%) | 70.97 | 73.87 | 73.11 | 76.19 |
| Substitutions (%) | 8.91 | 7.29 | 8.25 | 6.66 |
| Deletions (%) | 20.12 | 18.84 | 18.64 | 17.16 |

* The ratio of insertions indicates where the ASR model wrongfully added a character. The number of substitutions indicates where a character was wrongfully recognized. The ratio of deletions indicates where the ASR model wrongfully did not add a character to the output.

# 4 Discussion

The present study investigated how data from specific demographic groups could be used to develop an ASR model applicable in daily care practice and academia. This method decreases WERs regardless of specific demographic characteristics, such as accents, age, or speech problems. Results show that interview data stemming from a LTC setting can be transformed into a useful resource for improving the recognition performance of an existing ASR model.

The current results indicate that by iteratively processing and adding more interview data, the WER of the ASR model was reduced; every successive iteration of the method was able to recognize more samples (i.e., pairs of audio and text segments) from the interview data, eventually leading to a WER of 17.3%. These results imply that the process of iterative refinement is indeed effective for deploying interview data to decrease the WER of existing ASR models. Besides being tested on the MCV-NL dataset, the ASR model was tested on the interview data stemming from residents, family members, and care professionals. This resulted in slightly higher WERs. This may be due to various factors such as background noise, which is already known to increase WERs [6–8].

In addition, although differences appear to be small between the three groups (i.e., residents, family members, and care professionals), the highest WERs were displayed by residents. This could be related to the age of the persons in this group; since residents are the group with the highest average age, they are more likely to suffer from speech problems [24].

While part of the error of a model could be explained by the limitations of the ASR model itself, errors in manual interview transcripts could also have resulted in a higher WER (comparing an ASR model output to an incorrect transcript will still be perceived as an error of the ASR model). While transcripts ideally are a completely accurate presentation of the verbatim utterances of interview recordings, transcripts aren't perfect text replications of the audio at hand. Manually transcribed interviews may exclude mispronounced words, thinking noises (e.g., uh, uhm), privacy-sensitive information (e.g., names of the interview participant), and words and utterances that could possibly not be understood by the person transcribing the interview.

While there are commercially available systems that appear to achieve lower WERs than the WER conducted in the current study, these services often process data in the cloud (i.e., off-site). Since interviews conducted in a nursing home setting involve some of the most vulnerable people in our society, care and research organizations may require its processing to be conducted within the organization, rather than to hand it to a commercial party (e.g., Google or Microsoft). Hence, developing ASR models for use within an organization may be necessary to guard the privacy of interview participants.

An analysis of the time effectiveness showed that the ASR model transcribed at least 6 times faster than manual transcription. Consequently, a large part of the manual work that is required for transcription can be reduced using the method introduced in the current article. Feedback from researchers at the Limburg Living Lab in Ageing and Long-Term Care mentioned that while transcription results were perceived useful, factors like background noise being present resulted in a higher WER, and thus more errors in transcription which had to be corrected manually. However, many errors were small (i.e. related to the insertion of a letter) and transcripts could be understood 'as is'.

The error analysis shows that few substations occur, which may indicate that words that are recognized are generally spelled correctly. The number of insertions may correspond with non-lexical words, but also words that were omitted from the manual transcript and false positives from background noise. Cases have been observed where the ASR model recognizes words that were missing from the manual transcript. For example, the ASR model transcribed the following text "fthe last question is there anything you want to tell ...", whereas the word question was omitted in the transcript "the last ——is there something you want to tell ...". In this case the only error of the ASR model is the first 'f' in the sentence, while other characters are wrongfully regarded as errors. The number of deletions corresponds with the observed behavior of the ASR model where the beginnings and ends of segments are not correctly recognized. Although in all cases of errors, noise is observed to be a large factor. By using more advanced recording hardware: by using microphones that are less sensitive to background noise, the WER of the ASR model on those audio recordings may be reduced. Additionally, research has shown that many languages share similar speech features [25]. Consequently, the current method of sampling may not only be applicable to the current ASR model, but may also be applied to finetune ASR models aimed at other languages. When comparing the approach in the current study to results in of state-of-the-art methods for English ASR, the WER above 20% on interview data may seem high, however the difficulty of the speech in this dataset may contribute to a higher WER. The robustness (i.e. how well an ASR model performs across different datasets) may still be low even if the WER for one dataset is also low. For example, the English HuBERT model has a reported 1.9% WER on the LibreSpeech-clean dataset, however it has a 58.5% WER on the CHiME6 dataset [25].

Although the method of iterative improvement of the ASR model was shown to be effective and the resulting ASR model was viewed as a useful tool when transcribing interview recordings, the method also has its limitations. Firstly, the Smith-Waterman alignment is used to reduce the WER of text segments from the ASR model. While the Smith-Waterman approach of alignment does ensure a certain quality of data, text segments may still contain errors. This could potentially be improved by giving more context (i.e., words surrounding the text segment) to the ASR model to learn from [19, 26]. A second limitation regards the size of the model. Research implies that to be able to learn more nuanced speech features and reduce the WER of an ASR model, the size of that model and dataset should be increased [27, 28]; however, training and running a larger ASR model than the one discussed here would also require more computational resources, leading to higher costs for the use of the ASR model in daily practice [29].

In the Dutch language, the pronunciation of words may change based on used accent marks, such as *"é"* and *"ö"*. In the present study, this was not taken into account, as the ASR model was finetuned from an existing English model. The model used the characters A–Z, as well as dashes and apostrophes. Future research efforts should focus on using methods that employ characters with accent marks. This could further reduce the WER of the ASR model, as in the current study accent marks are removed, which also reduces the information that the ASR model receives.

In addition, research could focus on verifying if the text produced by the ASR is interpretable for qualitative analyses, to further validate the practical usage of the ASR model presented in this study. Currently, when interviews are conducted in LTC for older adults, these are transcribed and analyzed by hand. In these analyses, themes relevant to a research question are identified within the interviews [30]. When applying the WER metric, sentences which may have a high WER may still be understandable to human readers, and the qualitative analyses (i.e., the identification of themes) can still be conducted without producing erroneous results. For example, if each word in a sentence is slightly misspelled, it could still be interpreted without problems. Thus, the WER doesn't have to be zero for a successful qualitative analysis.

The current study developed and assessed an ASR model for usage on conversational data such as interviews regarding quality of care and quality of life. Research should also focus on finetuning an ASR model for clinical usage. Such a model would be used for speech data that contains clinical or medical terms (e.g. speech data in electronic health records).

## 5 Conclusion

This study shows that full-length interview data (consisting of audio and corresponding transcripts) can be effectively used to improve ASR. This was validated using data stemming from interviews with clients, family members and care professionals from long-term care for older adults. Relatively small real-world data from interviews can contribute to reduce the WER for demographic groups with less common speech features such as older adults. While ASR can certainly reduce the amount of time and (inherently) cost that is required in the manual transcription of audio, the results of the developed ASR system can't yet compete with manual transcription. However, the tediousness of verbatim transcription is reduced to only making (often small) corrections to automatically-generated transcripts.

## 6 Funding statement

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

## 7 Competing interests statement

The authors have no competing interests to declare.

## 8 Contributership statement

C.H. has written the manuscript and the corresponding software and contributed to the concept and design of the study. $J$.P.H. and $H$.V. contributed to data interpretation and carefully revised the manuscript. $S$.A. contributed to the concept, design and analyses of the study and carefully revised the manuscript.

## 9 Data availability

The code, and models discussed in the article will be made available at `https://github.com/coen22/Speech-Recognition`. Our interview data will not be publicly available due to the privacy of our participants. Upon request, our interview data may be provided with restrictions. Data are available from the Limburg Living Lab in Ageing and Long-Term Care (contact via Sil Aarts `s.aarts@maastrichtuniversity.nl`) for researchers who meet the criteria for access to confidential data.

## 10 Acknowledgements

The authors would like to thank the Data Science Research Infrastructure (DSRI) at Maastricht University and Samenwerkende Universitaire RekenFaciliteiten (SURF). Without the usage of their computational resources, the current study could not have been conducted. Moreover, thanks to the contributors of the "huggingface transformers" library, as the library provided many of the components for developing the ASR model in the current study.

## Bibliography


[1] J. Pols, “Enacting appreciations: Beyond the patient perspective,” *Health Care Analysis*, vol. 13, no. 3, pp. 203–221, 2005.

[2] K. Sion *et al.*, “The feasibility of connecting conversations: A narrative method to assess experienced quality of care in nursing homes from the resident’s perspective,” *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 5118, 2020.

[3] K. Y. Sion *et al.*, “Experienced quality of post-acute and long-term care from the care recipient’s perspective–a conceptual framework,” *Journal of the American Medical Directors Association*, vol. 20, no. 11, pp. 1386–1390, 2019.

[4] M. Murray *et al.*, “Narrative data,” *The SAGE handbook of qualitative data collection*, pp. 264–279, 2018.

[5] T. Delespierre, P. Denormandie, A. Bar-Hen, and L. Josseran, “Empirical advances with text mining of electronic health records,” *BMC medical informatics and Decision Making*, vol. 17, no. 1, pp. 1–15, 2017.

[6] S. Yang *et al.*, “SUPERB: speech processing universal performance benchmark,” *CoRR*, vol. abs/2105.01051, 2021. arXiv: `2105.01051`.

[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, “Wav2vec 2.0: A framework for self-supervised learning of speech representations,” *CoRR*, vol. abs/2006.11477, 2020. arXiv: `2006.11477`.

[8] W. Hsu, B. Bolte, Y. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, “Hubert: Self-supervised speech representation learning by masked prediction of hidden units,” *CoRR*, vol. abs/2106.07447, 2021. arXiv: `2106.07447`.

[9] S. Chen *et al.*, “Wavlm: Large-scale self-supervised pre-training for full stack speech processing,” *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[10] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, “Data2vec: A general framework for self-supervised learning in speech, vision and language,” *arXiv preprint arXiv:2202.03555*, 2022.

[11] J. Li *et al.*, “Recent advances in end-to-end automatic speech recognition,” *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[12] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, “Unsupervised automatic speech recognition: A review,” *Speech Communication*, vol. 139, pp. 76–91, 2022.

[13] N. T. Vu, Y. Wang, M. Klose, Z. Mihaylova, and T. Schultz, “Improving asr performance on non-native speech using multilingual and crosslingual information,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[14] M. Kothyari, A. R. Mekala, R. Iyer, G. Ramakrishnan, and P. Jyothi, "Personalizing asr with limited data using targeted subset selection," *arXiv preprint arXiv:2110.04908*, 2021.

[15] M. Meikle and S. Holley, "Communication with patients in residence," in *Care of the long-stay elderly patient*, Springer, 1991, pp. 149–160.

[16] L. Moro-Velazquez *et al.*, "Study of the performance of automatic speech recognition systems in speakers with parkinson's disease.," in *Interspeech*, 2019, pp. 3875–3879.

[17] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, pp. 4211–4215.

[18] Q. Xie, E. H. Hovy, M. Luong, and Q. V. Le, "Self-training with noisy student improves imagenet classification," *CoRR*, vol. abs/1911.04252, 2019. arXiv: 1911.04252.

[19] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," *arXiv preprint arXiv:2004.05150*, 2020.

[20] C.-C. Chiu *et al.*, "A comparison of end-to-end models for long-form speech recognition," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, IEEE, 2019, pp. 889–896.

[21] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[22] P. Delobelle, T. Winters, and B. Berendt, "RobBERT: A Dutch RoBERTa-based Language Model," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 3255–3265.

[23] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology*, vol. 48, no. 3, pp. 443–453, 1970.

[24] L. Werner, G. Huang, and B. J. Pitts, "Automated speech recognition systems and older adults: A literature review and synthesis," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications Sage CA: Los Angeles, CA, vol. 63, 2019, pp. 42–46.

[25] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*, PMLR, 2023, pp. 28 492–28 518.

[26] Z. Liu and N. F. Chen, "Dynamic sliding window for meeting summarization," *arXiv preprint arXiv:2108.13629*, 2021.

[27] J. Hoffmann *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[28] T. Brown *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[29] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[30] A. Strauss and J. Corbin, *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA, 1998.

# CHAPTER 4

## User-Centered Co-Design in Long-Term Care: Developing a MEDLO-Based App for Real-Time Daily Life Observations in a Nursing Home Setting

This chapter was accepted for publication by JMIR Human Factors.

## Abstract

**Background**: Assessing the daily lives of older adults, including their activities, social interactions and well-being is essential, particularly in nursing homes, as it gains insights into their quality of life. Methods such as the Excel-based Maastricht Electronic Daily Life Observation tool (MEDLO) are time-consuming and require extensive manual input, making them difficult to use.

**Objective**: The aim of this study was to develop an app-based version of the MEDLO using a User-Centered Design (UCD) and co-design approach to enhance efficiency and usability. We looked to actively involve researchers and care professionals who have used the MEDLO before, throughout the development process.

**Methods**: Participants included a diverse group of researchers and care professionals experienced in using the MEDLO. The UCD approach involved multiple iterative phases including semi-structured interviews, user research sessions, and application development. Data were analyzed using a qualitative (thematic) approach of UCD and user research sessions. The app, which was preferred to the traditional Excel-based MEDLO, underwent multiple iterations. This method primed the continuous iterative development of the app, aimed for a Minimum Viable Product (MVP).

**Results**: This study included 14 participants, primarily female, from diverse professional backgrounds. Their feedback highlighted the need for efficiency improvements in tool preparation and data management. Key improvements included automated data handling, an intuitive tablet interface, and functionalities such as randomization and offline data syncing.

**Conclusions**: The iterative development process led to an app that aligns with end-user needs, indicating potential for improved usability. Early and continuous user involvement was key in enhancing the application's usability, demonstrating the importance of user feedback in the development process.

# 1 Introduction

Assessing the daily lives of older adults, particularly in the context of nursing homes, is essential for gaining insights into quality of life [1, 2]. It allows health-care providers to detect and address issues related to their physical health (e.g., mobility limitations, pain), mental well-being (e.g., mood disturbances, cognitive decline), and social interactions (e.g., isolation, engagement in activities) in a timely manner [1, 3, 4]. This information can then be used to personalize care plans, thereby potentially improving health outcomes and enhancing the overall quality of life of older adults [5, 6].

Ecological Momentary Assessments (EMA) have been developed to facilitate this process by providing real-time, on-site evaluations of an individual's well-being [5, 7, 8]. EMA involves collecting data on individuals' behaviors and experiences in their natural environments, which can then be used to identify patterns and inform interventions [7]. However, existing EMA tools are mainly used in research settings, and their implementation in clinical practice has been challenging due to the time-consuming nature of data collection and the complexity of the tools [9, 10]. The Maastricht Electronic Daily Life Observation tool (MEDLO-tool) is designed to assess the daily lives of nursing home residents using EMA methodologies [1, 4]. The tool captures real-time information across several key dimensions of daily life, including activity levels (e.g., participation in communal or solitary activities), agitation, mood states, and interactions with staff, fellow residents, and visitors [1, 4, 11]. By systematically observing and recording these aspects, the MEDLO-tool provides a nuanced view of residents' experiences, which can inform personalized care strategies [4]. However, the MEDLO-tool relies on Excel templates that require significant manual input and are not user-friendly for care professionals [1, 4]. The complexity of the Excel-based system poses significant challenges, including data entry errors, time inefficiency, and barriers to widespread adoption in clinical settings [12, 13]. This complexity can be particularly problematic in nursing homes, where staff may have limited time and technological proficiency [13].

Developing a mobile application for the MEDLO-tool could address these challenges by automating data collection processes, reducing manual input, and providing an intuitive interface for users [14, 15]. An app could streamline observations, enable real-time data analysis, and facilitate immediate feedback to care providers, thereby enhancing the tool's utility in clinical practice. Moreover, considering the unique needs of residents with dementia, it is important that such an app is designed to account for cognitive impairments and communication difficulties [16, 17].The development of such an app requires a thorough understanding of the needs and preferences of the end-users. A User-Centered Design (UCD) approach is vital for developing software that truly meets users' needs [14, 18]. UCD emphasizes involving users throughout the design process to ensure that the product aligns with their requirements and preferences [19]. This involves iterative cycles of user need assessment, design, prototyping, and testing [20]. Co-design, a component of UCD, involves direct collaboration between users and designers, requiring the active participation of users in generating ideas, making decisions, and solving problems [21, 22]. This not only ensures the software's functionality but also its usability [21].

Potential barriers to implementing such technology include technological literacy among staff, user engagement, and ensuring data privacy and security [13, 23]. By actively involving care professionals and stakeholders in the design process, we aim to create a tool that is both functional and user-friendly, facilitating its adoption in clinical practice and ultimately improving the quality of care for residents with dementia. Therefore, this study is aimed at developing an app for measuring the daily life of residents with dementia living in nursing homes, using a user-centered and co-design approach.

## 2 Methods

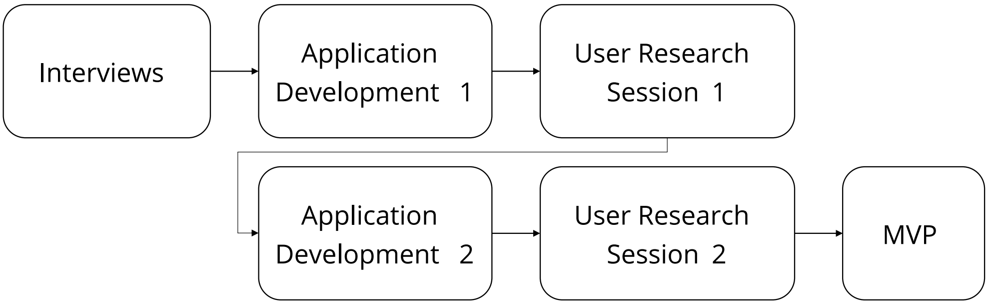



**Figure 4.1:** The process of app development.

This methodological study aimed to develop a mobile application version of the MEDLO-tool for use in long-term care facilities, employing a User-Centered Design (UCD) approach [24–26]. The study unfolded in multiple iterative phases, including semi-structured interviews, user research sessions, application development, and comprehensive data analysis, as illustrated in Figure 4.1. The project culminated in the achievement of a Minimum Viable Product (MVP).

## 2.1 Participants

Participants were selected through purposive sampling to include individuals with experience using the MEDLO-tool. The sample consisted of researchers and healthcare professionals affiliated with Maastricht University and local nursing homes. Inclusion criteria were prior use of the MEDLO-tool and willingness to participate in the study. Fourteen participants were recruited, comprising research assistants, PhD candidates, post-doctoral researchers, senior researchers, occupational therapists, research coordinators, and associate professors.

## 2.2 Data Collection

Data collection involved semi-structured interviews and user research sessions conducted between January and June 2023. Ethical approval was obtained from the Maastricht University Institutional Review Board (approval number XYZ123), and informed consent was secured from all participants.

### Interviews

Semi-structured interviews were guided by a comprehensive list of topics derived from available literature and consultation with the researchers who developed the original MEDLO-tool [1, 4]. The interviews aimed to gather insights into participants' experiences with the Excel-based MEDLO-tool, usability issues, functionality requirements, data analysis techniques, and general impressions [27]. Each interview lasted approximately 45 to 60 minutes and was audio-recorded with participants' consent. The audio was transcribed verbatim for analysis. The interview guide is provided in the supplementary materials.

The interview transcriptions were subjected to a six-phase thematic analysis process, aligning with the methodologies supported by prior research [28–30]. The inductive analysis was conducted by one researcher. A second researcher reviewed the coding to enhance trustworthiness. This comprehensive process involved familiarization with the data, generating initial codes, identifying themes, reviewing themes, defining and naming themes, and ultimately, reporting the results. The resulting report summarized the key findings, interpretations, and implications, to provide valuable insights for the next stages of the development of the app.

### User Research Sessions

Two user research sessions were conducted monthly to involve participants in the iterative development process. Each session lasted approximately two hours and was held at Maastricht University. Participants interacted with app prototypes using tablets and smartphones provided by the research team. The sessions included guided tasks (e.g., completing an observation using the app), open-ended discussions, and interactive feedback activities facilitated by researchers using prototypes and interactive demos. The structure of these sessions was informed by previous co-design methodologies [1, 3, 31].

## 2.3 Data Analysis

### Qualitative Analysis of Interviews

Interview recordings were transcribed verbatim and analyzed using thematic analysis following Braun and Clarke's six-phase approach [28]. Two researchers independently coded the transcripts to enhance reliability. Initial codes were generated and organized into potential themes. Discrepancies in coding were discussed and resolved through consensus meetings. Themes were reviewed and refined to ensure they accurately represented the data. The final themes captured key insights into the usability and functionality of the MEDLO-tool, informing the app development. The overall process of thematic analysis was guided by earlier works in this field [29, 30]. The session guide is provided in the supplementary materials.

### User Research Sessions

Each user research session included guided tasks (e.g., completing an observation using the app), open-ended discussions, and interactive feedback activities facilitated by researchers using mock-ups and interactive demos. Feedback from user research sessions was documented through field notes and audio recordings. An inductive content analysis was conducted by one researcher, who coded feedback and categorized issues into four themes: layout, functionality, errors, and appearance. A second researcher reviewed the coding to enhance trustworthiness. This categorization allowed the development team to prioritize user concerns based on frequency and severity, ensuring a user-centric approach to app refinement.

## 2.4 Application Development

The application was developed using '.NET MAUI' (i.e. a framework designed by Microsoft) for cross-platform compatibility on iOS and Android devices. The server-side (i.e. where the data is sent to) utilized 'ASP.NET Core' (a i.e. a framework designed by Microsoft) for a maintainable and performant application. The development process was iterative, with weekly meetings between developers and researchers to integrate user feedback from the interviews and user research sessions.

Usability testing involved participants completing specific tasks using the app prototypes while researchers observed and noted any issues. Testing sessions were conducted in a controlled environment to simulate actual usage conditions. Feedback from these tests informed further refinements and addressed potential adoption or feasibility issues.

# 3 Results

## 3.1 Sample

The study involved 14 participants with an average age of 36 years (standard deviation [SD] = 10 years; range 24–57 years), and 93% (13/14) of them were female. The participants represented a diverse range of professions, including research assistants (n = 3), PhD candidates (n = 4), post-doctoral researchers (n = 2), a senior researcher (n = 1), an occupational therapist (n = 1), a research coordinator (n = 1), an associate professor (n = 1), and one individual who was currently unemployed. Most participants (n = 9; 64%) were affiliated with Maastricht University, while others worked at healthcare providers (n = 4; 29%) or were unemployed (n = 1; 7%). The average duration of employment at their respective institutions was 6 years (SD = 5 years; range 0.5–16 years).

All participants had prior experience with the MEDLO-tool; twelve participants (86%) had used it directly in their research, while two participants (14%) had used it as inspiration for developing other tools different from MEDLO. All participants had also been in contact with the developers of the MEDLO-tool. Table 4.1 presents the demographic information of the participants in detail.

**Table 4.1:** Demographic information of the participants.

| **Characteristic** | **Details** |
|---|---|
| Age | 24–57 years old (Mean: 36 years; SD: 7.7 years) |
| Gender | Female: 13 (93%), Male: 1 (7%) |
| Time at Institution | 6 months–16 years (Mean: 6 years; SD: 3.3 years) |
| Use of the Tool | Used in research: 12 (86%), Inspiration for another tool: 2 (14%) |
| Contact with Developers | All 14 participants (100%) |
| Affiliations | |
| - University | 9 participants (64%) |
| - Healthcare Provider | 4 participants (29%) |
| - Unemployed | 1 participant (7%) |

## 3.2 Interviews

The participants described the process of using the MEDLO-tool in research projects as comprising seven distinct phases: acquiring informed consent; preparing the Excel sheets before the observation; familiarizing themselves with the faces of the residents who would be observed; conducting the observations; fixing issues with the data (such as missing values or inconsistencies); analyzing the data; and communicating the data back to the care organizations.

Participants emphasized that the preparation of the tool and the subsequent data cleaning were particularly time-consuming and would significantly benefit from automation. One participant stated: *"Every time before we do a new observation, we have to look at the participants, randomize them several times, and input all of this into the Excel sheet."* Another participant highlighted the challenges faced after data collection: *"After the observation is done, I have to go through the whole document to check all the fields that I didn't have time to fill out or that I didn't know how to rate."* These comments reflect the manual and labor-intensive nature of the existing process, which participants found cumbersome.

Participants reported that during observations, they used separate pieces of paper to write down descriptions of each resident's appearance because the observation periods were too short to input this information into the Excel sheet in real time. They suggested that having an additional field to securely write this information directly on the tablet would make the process more accessible. One participant explained: *"We usually arrive at least half an hour earlier, so we can ask the nurses which participant corresponds to which name, and write down what they look like. We do that to find them back more easily during the observation."* Another participant elaborated on the logistical challenges: *"The data entry process on the tablet is intuitive, but we also have to keep track of the patient names and descriptions on paper, we carry the manual, and we use a timer on our phone to ensure that we don't spend too long on any resident. This is a lot to carry around."* Regarding data reporting, participants indicated that the data reported back to the nursing homes primarily consisted of quantitative measures, which could potentially be automated through a dashboard (e.g., PowerBI). Participants expressed that automating this process would reduce the time between data collection and reporting, which currently could be several months. One participant noted: *"When we want to report back the numbers to the care organization, we make a PowerPoint presentation that includes aggregated numbers. This usually takes place several months after the observations."* Another participant suggested: *"I think some of the numbers could be reported back automatically."*

Overall, participants preferred using a tablet compared to using pen and paper, citing the intuitive data entry process. However, they also mentioned that a tool compatible with smartphones would be beneficial, especially for care professionals who might find tablets less convenient. One participant remarked: *"I think the tool works well, but I'm not sure if it would be usable like this for nurses."*

## 3.3 Initial User Research Session

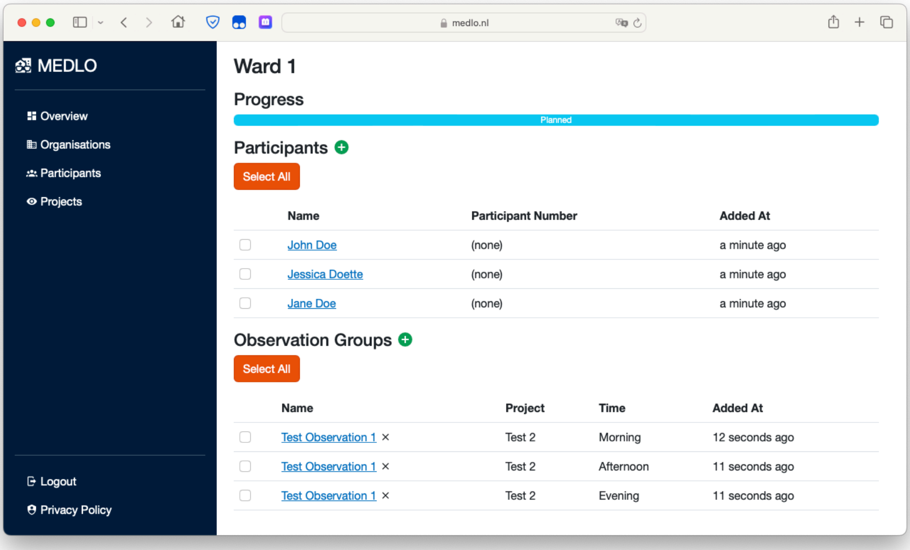



**Figure 4.2:** The dashboard as it was presented to the participants.

Based on the information gathered during the interviews and the suggested points for improvement, an initial version of the application was created and provided to the participants for the first user research session. This first draft of the application consisted of an online dashboard and a mobile app, both of which included the core components but had limited functionality. Figure 4.2 shows the app and dashboard as they were presented to the participants. Table 4.2 provides a detailed overview the feedback received during the initial user research session.

During the initial user research session, participants expressed largely positive sentiments towards the concept of the application and appreciated the efforts made in its development. One participant commented: *"I know we have a lot of complaints about the app, but we really appreciate the work [the development team] has done already."* However, participants also identified several areas for improvement. They expressed concerns regarding certain features of the app, such as the timer functionality, which lacked a reset option. Participants noted that in

**Table 4.2:** Feedback from the initial user research session

| Type | Feedback | Addressed |
|---|---|---|
| Functionality | Implement the functionality to search for participants and view projects. | |
| Functionality | Adjust the timer to enhance its function. Currently, there's no reset option. | Adjusted |
| Layout | Improve the screen layout for better overview and usability. Easy switching between participants via list. | Adjusted |
| Functionality | Allow 'stars' to be given to observations to easily find them later. | Adjusted |
| Functionality | Show an indicator an observation is completely filled in. | Adjusted |
| Functionality | Add the capability to save data and make it available for download. | |
| Functionality | On disruption or sudden app closure, data should remain safe. The app should sync and work offline beforehand. | |
| Functionality | Add a field for linking participant numbers in the app with numbers from other survey forms done in Excel. | Adjusted |
| Error | Update the app according to the latest MEDLO manual. Currently, there are inconsistencies in functions & terms. | Adjusted |
| Layout | Create a summary screen showing filled-in data so far. | |
| Functionality | Implement the ability to view projects in the app. | |
| Functionality | Add an option to upload a floor plan for easy location & participant identification. | |
| Functionality | Ensure the app can be used entirely in Dutch. | Adjusted |
| Layout | Results should be more organized. There should be a filter option. Development over time isn't important. | |
| Appearance | "Morgen" isn't immediately clear. "Ochtend" would be clearer. | Adjusted |
| Appearance | Colors need to be adjusted to match the AWO-L. | Adjusted |

their current practice, they used interval timers on their phones to manage observation times for each resident, and the app's timer did not adequately support this need. One participant suggested: *"What could be nice is some kind of interval timer. That's what we currently use on our phones."* Additionally, participants missed the comprehensive overview provided by the Excel-based MEDLO-tool, as the prototype app had been designed primarily for phone screens and did not offer the same level of data visibility. One participant observed: *"Normally, you could easily scroll back in the overview."* Moreover, there were some unclear aspects of the app, such as certain functionalities being available only through the dashboard and not within the app itself. Participants found labels on buttons to be unclear, which led to confusion during navigation. They recommended improving the screen layout for better usability and enabling easy switching between participants via a list.

Participants raised concerns about data storage and security, even though these aspects had been addressed in the app's design. They questioned how the data was stored and whether it was encrypted. It was explained to them that the data was stored on the device (i.e., phone or tablet) in encrypted form and could be synced to a secure server. Regarding randomization, participants asked how it would be conducted within the app and where they could access this feature. It was clarified that randomization was performed automatically upon creating a new observation group. A participant inquired: *"How do I randomize the participants in the app?"* To which the response was: *"That is done automatically when an observation group is added."*

Participants inquired about features that had not yet been implemented but were planned for future updates. For example, they asked if the app would work offline, which was an important consideration for use in environments with limited internet connectivity. They also asked whether the data could be exported to Excel for further analysis. These features were already part of the application's development roadmap but were not yet available in the user interface at the time of the session. One participant expressed concern: *"I don't know if it's an issue when I lose connection, or whether I lose my data."* Another participant discussed data export capabilities: *"These are the data you see when you download the Excel file. With Excel, it looks a bit different, but what you see here are all the column names."*

## 3.4 Second User Research Session

Following the initial user research session, the app was updated to address the feedback received. Adjustments were made to the timer functionality, screen layout, language translation, and several other features as per the participants' suggestions. A second version of the application was then provided to the participants for the subsequent user research session. In Table 4.3, the feedback received during the second user research session is detailed.

In the second session, participants expressed fewer concerns regarding the application, indicating that many of their initial issues had been resolved. However, they still provided valuable feedback on functionality and appearance. One of the main concerns was that the app was not yet fully translated into Dutch; while significant progress had been made, some parts remained in English. Participants emphasized the importance of complete translation before the app could be ef-

**Table 4.3:** Feedback from the second user research session.

| Type | Feedback |
|---|---|
| Functionality | While a portion of the app is translated, this process should continue to make all elements and features of the app fully available in Dutch. |
| Functionality | Consider adding a feature to remove participants, e.g., when a participant is deceased. |
| Functionality | The criteria for showing and hiding fields, based on what is relevant, during observation should be further extended. |
| Appearance | The criteria for marking an observation as ’complete’ can be confusing. Once an observation has an activity, it should be marked as complete. |
| Appearance | Consider making the stars orange for improved readability and make the star button a different color. |
| Functionality | Offer flexibility in the number of observation rounds that can be added. |
| Functionality | Make the process more efficient by creating a priority list and/or extending the default time block for certain observations. |
| Functionality | Ensure that the manual can be accessed from within the app. It should be clear and comprehensive, aligning with users’ practice. |

fectively used by nurses and other care professionals. One participant remarked: *”Most of the app has been translated, but some of the texts are still only in English.”* Another participant added: *”This all needs to be translated before we can hand this to nurses.”*

Participants also suggested that the criteria for showing and hiding fields during observations should be further refined based on relevance, to streamline the data entry process and reduce cognitive load. They discussed various small changes in the appearance of the app to enhance usability and visual appeal. For instance, they recommended that the ’star’ button used to mark important observations be made a different color to stand out more prominently. One participant suggested: *”Could you make the star button orange to make it stand out a bit more?”* Additionally, participants found the criteria for marking an observation as ’complete’ to be somewhat confusing. They proposed that once an activity has been set for an observation, it should be automatically marked as complete to provide clearer feedback to the user. One participant explained: *”I would prefer it if the observation could be marked as ’complete’ when the activity has been set.”* Participants also recommended offering flexibility in the number of observation rounds that could be added, as different research protocols or care routines might require varying numbers of observations. They suggested making the process more efficient by creating a priority list or extending the default time block for certain observations that typically take longer. Furthermore, participants emphasized the importance of ensuring that the app's manual could be accessed from within the app itself. They stressed that the manual should be clear, comprehensive, and align with users' practical experiences to facilitate ease of use, especially for new users.

By the conclusion of the second user research session, participants expressed optimism about the app's potential to improve their workflow and reduce the time spent on manual data entry and processing. They acknowledged the responsiveness of the development team to their feedback and looked forward to future iterations of the app that would incorporate their latest suggestions.

# 4 Discussion

This study successfully developed an app-based version of the Maastricht Electronic Daily Life Observation tool (MEDLO-tool) aimed at assessing the daily life of residents with dementia in nursing homes. By employing a user-centered design (UCD) approach, we incorporated feedback from participants throughout the development process, resulting in a minimum viable product that aligns with the needs and preferences of end-users.

The iterative development process revealed that larger issues, such as bugs and significant feature requests, were identified and addressed in the initial phases. Subsequent user research sessions showed a decrease in the number and severity of issues, indicating progressive refinement of the app. Participants expressed that the app improved upon the Excel-based MEDLO-tool by streamlining data collection and reducing manual input, thereby potentially enhancing usability.

Balancing end-user feedback with established best practices and user interface guidelines was crucial during development [32–35]. While participants provided valuable suggestions, not all could be implemented due to conflicting requests and constraints related to usability standards. For instance, some users desired a more complex timer feature, whereas others preferred simplicity. Decisions were made to benefit the majority and adhere to platform guidelines to ensure familiarity and ease of use [32, 33].

The effectiveness of the UCD approach in this study aligns with previous research demonstrating its benefits in software development [36, 37]. A key factor in our success was the development team's familiarity with the context of long-term care for older adults. This domain knowledge facilitated insightful conversations between developers and participants, reducing misunderstandings and accelerating the development process. Prior studies have highlighted that cognitive similarity and shared jargon can enhance team performance [38, 39].

### Limitations

Despite these strengths, the study has limitations that warrant consideration. The participant sample consisted primarily of researchers affiliated with the university, which may limit the generalizability of the findings to care professionals who are the intended end-users in clinical settings. While these researchers had extensive experience with the MEDLO-tool, involving a broader range of stakeholders, including nurses and other care staff, could provide additional insights into usability and practical implementation. Furthermore, the app was not tested in a real-life nursing home environment, which could have revealed context-specific challenges and opportunities.

### Future Work

Future research should focus on conducting pilot studies to evaluate the feasibility and effectiveness of the MEDLO app in real-world nursing home settings. Involving care professionals in these studies would help assess the app's usability, accessibility, and impact on daily workflows. Additionally, gathering feedback from residents and their families could provide a more comprehensive understanding of the app's influence on care quality.

Further development could explore integrating advanced technologies, such as artificial intelligence and machine learning, to enhance data analysis and provide predictive insights [40, 41]. For example, natural language processing could be used to automate the interpretation of observational data, aiding care providers in identifying patterns and potential areas for intervention. However, implementing such technologies would require careful consideration of ethical implications, data privacy, and user acceptance.

Future research may also investigate the usability in practice, by letting nurses use the app in their daily work. This could provide valuable insights into the app's integration into existing workflows and its impact on the quality of care provided to residents. Additionally, longitudinal studies could assess the long-term effects of using the app on care outcomes and staff satisfaction. This information could inform further refinements and improvements to the MEDLO app, ensuring its relevance in long-term care.

# 5 Conclusion

This study successfully demonstrates the viability and demand for an app-based Maastricht Electronic Daily Life Observation tool (MEDLO-tool) for assessing residents with dementia in nursing homes. By utilizing a User-Centered Design (UCD) approach, this study addressed the limitations of the existing Excel-based system by offering a more efficient and user-friendly alternative. This study shows that involving users early on in the process and keeping them involved can have a positive effect on the usability of an application. The MEDLO app shows that a UCD approach can provide real benefits in the development of a digital tools in nursing homes.

# Bibliography


[1] B. De Boer, H. Beerens, S. M. Zwakhalen, F. E. Tan, J. P. Hamers, and H. Verbeek, “Daily lives of residents with dementia in nursing homes: Development of the maastricht electronic daily life observation tool,” *International Psychogeriatrics*, vol. 28, no. 8, pp. 1333–1343, 2016.

[2] R. G. Logsdon, L. E. Gibbons, S. M. McCurry, and L. Teri, “Assessing quality of life in older adults with cognitive impairment,” *Psychosomatic medicine*, vol. 64, no. 3, pp. 510–519, 2002.

[3] G. F. Dunton, “Ecological momentary assessment in physical activity research,” *Exercise and sport sciences reviews*, vol. 45, no. 1, p. 48, 2017.

[4] H. C. Beerens *et al.*, “The association between aspects of daily life and quality of life of people with dementia living in long-term care facilities: A momentary assessment study,” *International psychogeriatrics*, vol. 28, no. 8, pp. 1323–1331, 2016.

[5] L. Bajenaru, I. A. Marinescu, C. Dobre, G. I. Prada, and C. S. Constantinou, “Towards the development of a personalized healthcare solution for elderly: From user needs to system specifications,” in *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, IEEE, 2020, pp. 1–6.

[6] A. L. Stewart and A. C. King, “Evaluating the efficacy of physical activity for influencing quality-of-life outcomes in older adults,” *Annals of behavioral medicine*, vol. 13, no. 3, pp. 108–116, 1991.

[7] S. Shiffman, A. A. Stone, and M. R. Hufford, “Ecological momentary assessment,” *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[8] R. C. Moore, C. A. Depp, J. L. Wetherell, and E. J. Lenze, “Ecological momentary assessment versus standard assessment instruments for measuring mindfulness, depressed mood, and anxiety among older adults,” *Journal of psychiatric research*, vol. 75, pp. 116–123, 2016.

[9] C. K. Or and B.-T. Karsh, “A systematic review of patient acceptance of consumer health information technology,” *Journal of the American Medical Informatics Association*, vol. 16, no. 4, pp. 550–560, 2009.

[10] M.-P. Gagnon *et al.*, “Systematic review of factors influencing the adoption of information and communication technologies by healthcare professionals,” *Journal of medical systems*, vol. 36, pp. 241–277, 2012.

[11] H. C. Beerens *et al.*, “The relation between mood, activity, and interaction in long-term dementia care,” *Aging & mental health*, vol. 22, no. 1, pp. 26–32, 2018.

[12] H. Lee *et al.*, "How i do it: A practical database management system to assist clinical research teams with data collection, organization, and reporting," *Academic radiology*, vol. 22, no. 4, pp. 527–533, 2015.

[13] C. S. Kruse, C. Kristof, B. Jones, E. Mitchell, and A. Martinez, "Barriers to electronic health record adoption: A systematic literature review," *Journal of medical systems*, vol. 40, pp. 1–7, 2016.

[14] B. Adipat and D. Zhang, "Interface design for mobile applications," *AMCIS 2005 proceedings*, p. 494, 2005.

[15] B. Martínez-Pérez, I. De La Torre-Díez, and M. López-Coronado, "Mobile health applications for the most prevalent conditions by the world health organization: Review and analysis," *Journal of medical Internet research*, vol. 15, no. 6, e120, 2013.

[16] P. dos Santos *et al.*, "Alzheimer's disease: A review from the pathophysiology to diagnosis, new perspectives for pharmacological treatment," *Current medicinal chemistry*, vol. 25, no. 26, pp. 3141–3159, 2018.

[17] A. Bernstein Sideman, K. L. Harrison, S. B. Garrett, G. Naasan, D. P. C. W. Group, and C. S. Ritchie, "Practices, challenges, and opportunities when addressing the palliative care needs of people living with dementia: Specialty memory care provider perspectives," *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, vol. 7, no. 1, e12144, 2021.

[18] W. Wang, J. Grundy, H. Khalajzadeh, A. Madugalla, and H. O. Obie, "Designing adaptive user interfaces for mhealth applications targeting chronic disease: A user-centric approach," *arXiv preprint arXiv:2405.08302*, 2024.

[19] D. A. Norman, *User centered system design: New perspectives on human-computer interaction*. CRC Press, 1986.

[20] J. D. Gould and C. Lewis, "Designing for usability: Key principles and what designers think," *Communications of the ACM*, vol. 28, no. 3, pp. 300–311, 1985.

[21] L. Tan and D. Szebeko, "Co-designing for dementia: The alzheimer 100 project," *Australasian Medical Journal (Online)*, no. 12, p. 185, 2009.

[22] P. Slattery, A. K. Saeri, and P. Bragge, "Research co-design in health: A rapid overview of reviews," *Health research policy and systems*, vol. 18, pp. 1–13, 2020.

[23] B. Martínez-Pérez, I. De La Torre-Díez, and M. López-Coronado, "Privacy and security in mobile health apps: A review and recommendations," *Journal of medical systems*, vol. 39, pp. 1–8, 2015.

[24] D. McCusker and K. Guzik, "User-centered design for middleware," in Elsevier, 2007.

[25] B. Couture *et al.*, "Applying user-centered design methods to the development of an mhealth application for use in the hospital setting by patients and care partners," *Appl Clin Inform*, 2018.

[26] C. Abras, D. Maloney-Krichmar, J. Preece, *et al.*, "User-centered design," *Bainbridge, W. Encyclopedia of Human-Computer Interaction. Thousand Oaks: Sage Publications*, vol. 37, no. 4, pp. 445–456, 2004.

[27] C. Wilson, *Interview techniques for UX practitioners: A user-centered design method*. Newnes, 2013.

[28] V. Braun and V. Clarke, "Using thematic analysis in psychology," *Qualitative Research in Psychology*, vol. 3, no. 2, pp. 77–101, 2006.

[29] M. Maguire and B. Delahunt, "Doing a thematic analysis: A practical, step-by-step guide for learning and teaching scholars.," *All Ireland Journal of Higher Education*, vol. 9, no. 3, 2017.

[30] A. Strauss and J. Corbin, *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA, 1998.

[31] S. Fox *et al.*, "Co-design of a smartphone app for people living with dementia by applying agile, iterative co-design principles: Development and usability study," *JMIR Mhealth Uhealth*, vol. 10, no. 1, e24483, 2022, ISSN: 2291-5222.

[32] K. Visnapuu, *Consistency, efficiency, and scalability in design systems: A comparative analysis of material design, apple's human interface guidelines, and ibm design systems.* 2023.

[33] T. Gülenman, *Designing better mobile apps: An experimental evaluation of apple's and google's design guidelines: How analysing the human interface guidelines for ios and material design for android better our understanding of the usability challenges app users face and what we can do to overcome key issues.* 2022.

[34] R. E. Eberts, *User interface design*. Prentice-Hall, Inc., 1994.

[35] P. Dey, B. R. Sinha, M. Amin, and H. Badkoobehi, "Best practices for improving user interface design," *Int J Softw Eng Appl*, vol. 10, no. 5, 2019.

[36] J.-Y. Mao, K. Vredenburg, P. W. Smith, and T. Carey, "The state of user-centered design practice," *Communications of the ACM*, vol. 48, no. 3, pp. 105–109, 2005.

[37] C. P. J. M. Van Elzakker and I. Delikostidis, *User-Centered Design of Mobile Geo-Applications*. IGI Global, 2013.

[38] H.-R. Kang, H.-D. Yang, and C. Rowley, "Factors in team effectiveness: Cognitive and demographic similarities of software development team members," *Human Relations*, vol. 59, no. 12, pp. 1681–1710, 2006.

[39] J. He, B. S. Butler, and W. R. King, “Team cognition: Development and evolution in software project teams,” *Journal of Management Information Systems*, vol. 24, no. 2, pp. 261–292, 2007.

[40] A. Ashraf and A. Imam, “Chatgpt’s use case for software engineers,” in *8th International Conference on Computing in Engineering and Technology (ICCET 2023)*, vol. 2023, 2023, pp. 487–492.

[41] B. Roziere *et al.*, “Code llama: Open foundation models for code,” *arXiv preprint arXiv:2308.12950*, 2023.

# PART II

## Analysis of Quality of Care Data

This part investigates the possibilities of AI in analyzing textual data regarding the quality of care.

# CHAPTER 5

## Text Mining in Long-Term Care: Exploring the Usefulness of Computer-Aided Analysis Methods

This chapter was published as:

Hacking, C., Verbeek, H., Hamers, J. P., Sion, K., & Aarts, S. (2022). Text mining in long-term care: Exploring the usefulness of artificial intelligence in a nursing home setting. PLOS ONE, 17(8), e0268281.

https://doi.org/10.1371/journal.pone.0268281

## Abstract

**Objectives**: In nursing homes, narrative data are collected to evaluate quality of care as perceived by residents or their family members. This results in a large amount of textual data. However, as the volume of data increases, it becomes beyond the capability of humans to analyze it. This study aims to explore the usefulness of text-mining approaches regarding narrative data gathered in a nursing home setting.

**Design**: Exploratory study showing a variety of text-mining approaches.

**Setting and Participants**: Data has been collected as part of the project *'Connecting Conversations'*: assessing experienced quality of care by conducting individual interviews with residents of nursing homes (n=39), family members (n=37) and care professionals (n=49).

**Methods**: Several pre-processing steps were applied. A variety of text-mining analyses were conducted: individual word frequencies, bigram frequencies, a correlation analysis and a sentiment analysis. A survey was conducted to establish a sentiment analysis model tailored to text collected in long-term care for older adults.

**Results**: Residents, family members and care professionals uttered respectively 285, 362 and 549 words per interview. Word frequency analysis showed that words that occurred most frequently in the interviews are often positive. Although there are some differences in wording such as the use of *'mother'* and *'breakfast'*, correlation analysis displayed that similar words are used by all three groups to describe quality of care. The majority of interviews displayed a neutral sentiment. Care professionals are more diverse in their sentiment than residents and family members: while some express a more positive sentiment, others express more negativity.

**Conclusions and Implications**: This study demonstrates the usefulness of a text-mining approach to extend our knowledge regarding quality of care in a nursing home setting. With the rise of textual (narrative) data, text-mining can lead to valuable new insights for long-term care for older adults.

# 1 Introduction

Patient perspectives have assumed a central role in various healthcare settings in the assessment of the quality of care [1, 2]. For example, in nursing homes, the perspectives of residents, their family members, and care professionals are seen as an important prerequisite for the improvement of quality of care [3]. To obtain more in-depth information regarding the quality of care so that its essential features can be further explored, narrative data are collected. These data often contain not only the experiences of residents, but also information regarding their engagement, satisfaction, and quality of life [2, 4]. Narrative data can be defined as any data that consist of stories concerning the lives of individuals (e.g., of residents) [5]. Examples of narrative data are interviews regarding the experienced quality of care or stories about the experience of older adults during the COVID-19 pandemic [6, 7]. In addition, certain websites (e.g., Zorgkaart Nederland) allow long-term care receivers to post opinions on the quality of their care and their lives.

To date, several narrative data methods have been developed to evaluate the quality of care and the quality of life of residents [8–10]. In a research context, these often involve audio recordings and verbatim transcriptions thereof. Hence, these methods result in large amounts of unstructured textual data that cannot be analyzed manually. Ordinarily, they are analyzed using coding. This involves summarizing participants' quotes in several words that capture their essence [11, 12]. It is a very time-consuming and tedious task and requires researchers to take a consistently objective approach. Consequently, large amounts of textual data regarding the quality of care gathered daily in nursing homes are only analyzed on a limited scale and with a limited scope. Since the amount of data using narrative methods is only expected to rise because of its expanding importance in science and health care in general, innovative analytical approaches are required.

Computers can process large quantities of data and deliver highly consistent results. Computerized methods, such as those found in data science, could offer a possible solution to overcome the difficulties of manual coding. Data science is a field aimed at extracting knowledge and insights from all kinds of data [13]. Text mining is a sub-category of data science that is directed at the retrieval, extraction, and synthesis of information from text [14]. Text mining includes methods such as frequency distributions, clustering, sentiment analysis, and visualization [14, 15]. In contrast with manual analysis, text mining involves processes and methods that analyze textual data automatically. As a result, it is often used to go beyond the scope of specific projects. It is not intended to replace manual coding analyses, but it does provide a novel way to analyze large amounts of text.

Text mining is already being used to gain insights into the quality of care. For example, it is used in hospitals to process health care claims; to group medical records by patients' symptoms [16, 17] or to predict the number of hospital admissions at an emergency department to avoid overcrowding [18]. A more recent study showed an automated method for extracting clinical entities, such as treatment, tests, drugs and genes, from clinical notes [19]. Another study discussed how text mining can be used to assess the sentiment in tweets towards the COVID 19 pandemic [20]. These types of applications suggest that text mining could also be useful in processing narrative data in long-term care for older adults to acquire novel insights into the quality of care and quality of life. Therefore, the present study aimed to explore the usefulness of text mining approaches to textual data gathered in nursing home care.

# 2 Methods

Several text mining methods were applied to examine narrative data collected within a nursing home setting. These narrative data consisted of interviews with residents of nursing homes, their family members, and their professional caregivers (i.e., triads). To enable automatic processing, all interviews were transcribed verbatim. Several preprocessing steps were applied. A variety of text mining analyses were conducted, including analysis of individual word frequencies and bigram frequencies, correlation analysis, and sentiment analysis. These are methods commonly used to gain insights into textual data [21].

## 2.1 Data collection

The data were collected as part of the project *'Connecting Conversations,'* which assesses the quality of care by conducting separate interviews with residents of nursing homes, family members, and professional caregivers (i.e., triads) [3, 4]. The underlying principles of *'Connecting Conversations'* are *'appreciative inquiry'* and *'relationship-centeredness'* [3]. Appreciative inquiry means that a positive approach is used to focus on what is already good and helpful and to do it more frequently. Relationship-centered care means that the impact of relationships is integral to the process and outcomes of the care experience [22]. It has been shown that appreciative inquiry does not necessarily lead to a more positive conversation [23]. A total of n = 125 interviews were conducted at 5 different care organizations in the south of the Netherlands [8]. "The medical ethical committee of Zuyderland (the Netherland) approved the study protocol (17-N-86)." Information about the study was provided to all interviewers, residents, family members and caregivers by an information letter. All participants provided written informed consent: residents with legal representatives gave informed consent themselves (as well as their legal representatives) before and during the conversations.

A survey was conducted to establish a sentiment analysis model tailored to long-term care. Respondents were presented with a list of sentences that were randomly selected from the transcripts of 'Connecting Conversations.' These sentences had to be assigned according to their sentiment: positive, neutral, or negative. The survey was anonymous, though participants were asked to provide some demographic details (i.e., age, gender, place of residence, and level of education). Participants were allowed to fill in the questionnaire at different times, and each time they were presented with different sentences.

## 2.2 Data preprocessing

Preprocessing is a critical step in text mining. It involves the removal of any noisy and unreliable data from the original dataset [15]. Noise is erroneous information that makes the data more difficult to interpret [24]. An example of noisy data could be stuttering during a phone call or misspelled words in transcribed interviews. A lack of preprocessing could yield erroneous results. The product of data preprocessing is the final data set, which can then be analyzed for meaningful information. Preprocessing and analysis were performed using R and Python. R is a free software package for statistical computing and graphics [25], while Python is a free software package for general-purpose computing [26]. The code for all the analyses can be found at `https://github.com/coen22/Text-Mining-AWO-L`.

To perform text mining analysis such as frequency distributions, correlation analysis, and sentiment analysis, the data were preprocessed in several steps: 1) all transcribed interviews were exported from Word files to Excel files, which were then loaded into R; 2) Words uttered by the interviewer were excluded from all the transcripts so that the results would not include them; 3) since many common words have no special significance (e.g., the, a, and is), removal of these so-called stop words was conducted by using a predefined list consisting of the most common Dutch stop words (n = 100); 4) since words in Dutch containing two letters are often stop words and/or non-informative, the minimum word length (i.e., the minimum number of characters that constitutes a word) was set to three; 5) a stemming approach was conducted to increase statistical significance for the various text mining methods. This approach groups words that refer to the same concept (e.g., nurses -> nurse, tummy -> stomach). It is similar to word embedding. Stemming is applied by computationally identifying the plurals and diminutives and reducing them to their root [27, 28]. This rule-based approach is more conservative than word embedding approaches that group more generally related words (e.g., mother -> family, nurse -> care) [29, 30].

## 2.3 Data analysis

### Word frequencies

To gain an initial understanding of the text, a frequency plot was conducted to visualize the individual words used most frequently in all the interviews (i.e. unigrams). Frequency plots can often be described by Zipf's law, which states that for any piece of text based on natural language, the frequency of any word is inversely proportional to its rank in the frequency distribution [15]. The (n = 50) words used with the highest frequency in the interviews are displayed in the frequency plots.

Bigram plots were also conducted. Bigrams [15] are combinations of two consecutive words such as *'very good.'* The sentence *'I love it here'* could therefore be split up into the bigrams *'I love,' 'love it,'* and *'it here.'* Bigrams can play a crucial role in text classification since they can capture the meanings of words that are not present when analyzing unigrams. For example, the word *'good'* has a different meaning when preceded by the word *'not.'* The 50 most frequent bigrams are displayed in terms of their relative frequency for all interviews with residents, family members of residents, and care professionals, respectively.

### Correlation analysis

It is important to understand which words co-occur in the interviews with residents and family members, residents and care professionals, and care professionals and family members, respectively. Correlation analyses were therefore conducted to assess the correlation (i.e., co-occurrence) of words across the three groups. The words are displayed as a collection of data points, divided over three scatter plots. A log transformation was used for the x and y-axes to account for the skewness of the data (i.e., only some words occurred very frequently) [31]. Three bilateral Pearson correlation coefficients (r) were assessed for residents and family members, residents and care professionals, and care professionals and family members, respectively.

### Sentiment Analysis

Sentiment analysis is the process of computationally identifying sentiment expressed in a piece of text [14, 15]. For example, the sentence *'It's a good day'* could be identified as being positive, while the sentence *'It's a bad day'* could be identified as being negative: the sentence *'Today I went for a walk'* could be neutral, as it does not convey whether the walk is experienced as a positive or negative event.

Previous sentiment analysis models have been based on the general Dutch language [32]. However, as certain words can have a different meaning in a nursing home setting, the sentiment behind these words can also be different. For example, there is a negative sentiment behind the word *'plassen'* (*'peeing'*), whereas in the general Dutch language this world would be considered neutral. Therefore, a general sentiment model may not be suitable for analysis in long-term care.

Using the results of the survey, the sentiment for all occurring unigrams (single words such as *'good'* or *'very'*) and bigrams was calculated as a value between (-1) and (1). A value of (-1) denoted the most negative sentiment, a value of (1) denoted the most positive sentiment, and a value of (0) denoted a neutral sentiment. The calculation was based on how frequently a word was used in a positive or negative context. For example, a word that occurred 9 times as a positive word and once as a negative word was given a sentiment of 0.8 ($(9 * 1 + 1 * -1)/10 = 0.8$). Since any sentiment larger than 0 could be defined as positive, a sentiment of 0.8 was positive. This analysis was carried out for all groups separately: residents (n = 39), family members (n = 37), and care professionals (n = 49), to show the difference between the three groups. The words that were used most often by each group were plotted with the corresponding sentiment of each word.

To show a broad overview of the differences in the overall experienced quality of care between the different groups, sentiment analysis on each interview was conducted [33]. This was illustrated by plotting the proportion of positive, negative, and neutral sentiments for each interview using a ternary plot. A ternary plot is a triangular plot capable of displaying three variables that sum up to the same value [34]. In the case of the present sentiment analysis, the percentages of positive, negative, and neutral sentiments in every interview summed up to 100%.

### Topic Clustering Analysis

Topic clustering is the process of identifying topics (i.e. themes) in text segments and clustering (i.e. grouping) them together based on those topics [35, 36]. Clustering is similar to a qualitative coding process using a grounded approach, as topics are discovered without any prior knowledge [12]. To discover topic clusters within interviews, several steps were conducted. Firstly, for each utterance in the interviews, keywords were extracted using part-of-speech (POS) tagging [33, 37]. POS tagging can identify whether a word is a noun (e.g. nurse, room), a proper noun (e.g. Sarah, Maastricht), a verb (e.g. walking, knitting), or any other type. By extracting nouns as keywords, it is possible to discover overarching semantic topics. Secondly, word2vec was used to calculate similarities in words [29, 30]. Word2vec is an algorithm that creates a vector (i.e. a point in high-dimensional space) for each word in such a way that words that are similar are located together [29]. Lastly, k-means was used to create k clusters of keywords [35]. For example, with k = 2, two clusters could be discovered that correspond to topics 'food' and 'family'. A value for k was calculated using the elbow method [36]. The elbow method balances between having many clusters which are too specific and having few clusters which are too general. For each cluster a topic name was manually assigned, based on the keywords belonging to that specific cluster. For example, if a cluster contained keywords such as 'mother', 'father' and 'daughter', the topic description was formulated as 'family'.

# 3 Results

In total, 125 interviews were analyzed: 39 with residents, 37 with family members, and 49 with staff. A total of n = 202,986 words were uttered. Residents uttered 284.9 words per interview, family members 362.1 words, and care professionals 548.7 words.

## 3.1 Word Frequencies

Figure 5.1 shows the distribution of the most frequently used words in the interviews. A typical Zipf's law pattern is visible, indicating that the most commonly used words accounted for almost half of all the words in the interviews [15]. The frequency of any word was inversely proportional to its rank in the frequency distribution (i.e., the rank-frequency distribution was an inverse relation). The word *'goed'* (*'good'*) was among the most frequently used words, indicating that the interviewees referred to many positive aspects. Moreover, words such as *'eten'* (*'food'*) and *'moeder'* (*'mother'*), provided insights into topics that participants perceived as important aspects of quality care. The word *'food'* was also mentioned frequently, suggesting that eating was another important topic. The frequent use of the word *'mother'* may refer to such residents being mentioned by family members. For example, residents' family members often referred to residents as their *'mother,'* indicating that the individual concerned was often a woman.

Since bigrams often contain more contextual information than unigrams, a bigram plot was conducted (Figure 5.2). The most frequently used bigrams included *'heel erg'* (*'very'*), *'heel goed'* (*'very well'*), *'wel goed'* (*'good'*), *'wel heel'* (*'very well'*), and *'wel leuk'* (*'nice'*). The bigram *'heel erg'* (very) was the most common bigram across all groups. This was used both negatively (e.g. *'very bad'*) and positively (e.g., *'very good'*). A sensitivity analysis of the words that proceeded this particular bigram revealed that the top 3 were *'goed'* (*'good,'* n = 9), *'leuk'* (*'nice,'* n = 7) and *'tevreden'* (*'satisfied,'* n = 7). By contrast, the word *'slechts'* (*'bad'*) only proceeded the bigram *'heel erg'* (*'very'*) once.

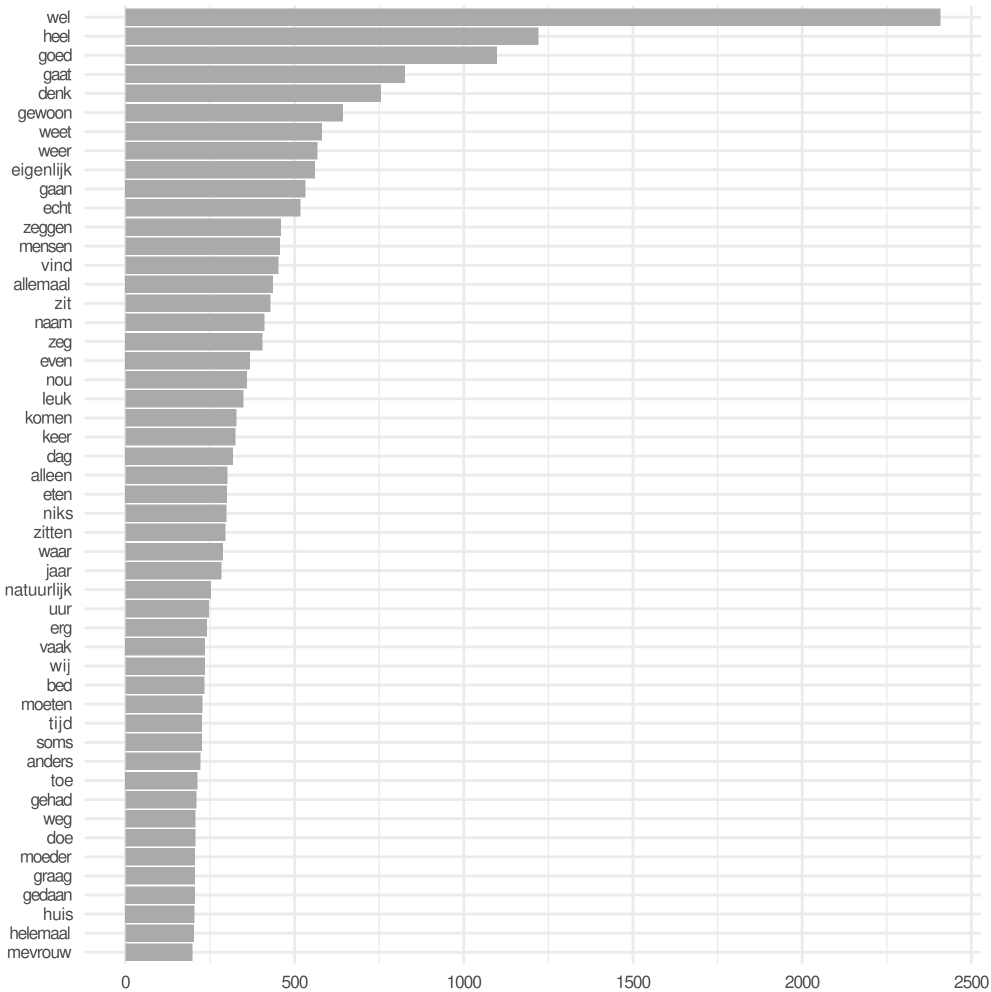



**Figure 5.1:** First (n = 50) most frequently occurring unigrams across all interviews.

## 3.2 Correlation

Figure 5.3 shows the correlation plots between word usage among the three groups, where each point represents a word. For some points, the English translation of the corresponding word is also shown. All three plots are similar. First, they each display an uphill pattern which is indicative of a positive relationship between the two variables. Second, the majority of words form a pattern around the red diagonal line, resulting in a high correlation coefficient r = 0.91, r = 0.83, and r = 0.92, respectively (i.e., word usage was largely similar across all groups). The further away the points are from the red diagonal line, the more different the word usage between groups and perhaps each group's perception of the quality of care. For example, words such as *'hobby'* (*'hobby'*), and *'dochter'*

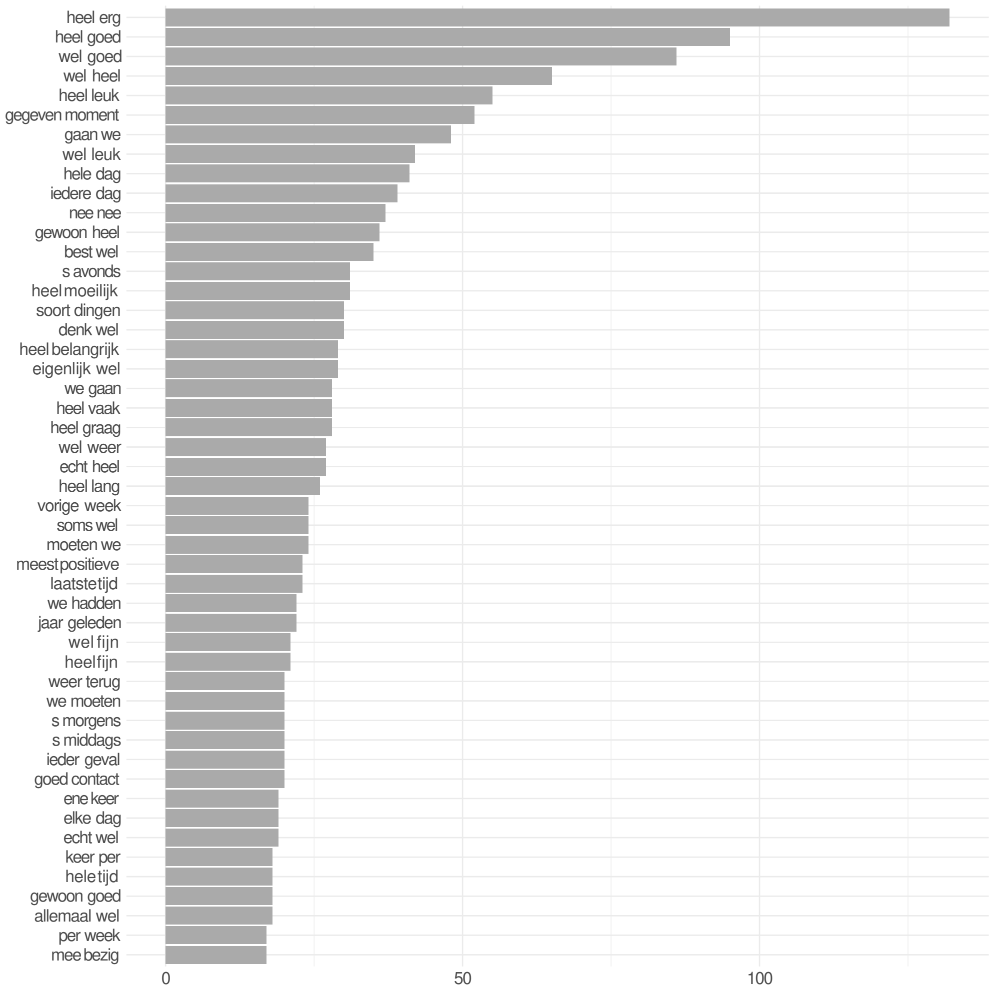



**Figure 5.2:** First (n = 50) most frequently occurring bigrams across all interviews.

(*'daughter'*) occurred more frequently in the interviews with the residents than with family members. This suggests that these subjects were more important to the residents, either negatively or positively. Family members used words such as *'instelling'* (*'institution'*) and *'zaterdag'*(*'Saturday'*) more frequently.

## 3.3 Sentiment Analysis

A total of 234 participants assigned a sentiment to 11,519 sentences, which was 56% of the total. The mean age of all participants was 41 (SD: 13.7). Of all participants, 71% were women. Sixty-seven percent of the participants had at least a master'*s degree, while 21% had a bachelor'*s degree; 67% said that they were currently living in the south of the Netherlands.

A scatter plot of the top 40 most commonly occurring words for residents, care professionals, and family members is displayed (Figure 5.4). The x-axis shows the sentiment value for each word between -1 (most negative) and 1 (most positive): the y-axis displays the frequency with which these words occurred. Many of the most common words were similar between residents, their family members, and care professionals. The words *'wel'* (*'well'* or *'quite'*), *'heel'* (*'very'*) and *'goed'* (*'good'*) occurred with very high frequency across all groups, but the sentiment was weakly positive. Words such as *'leuk'* (*'nice'*) and *'fijn'* (*'nice'*) occurred with high frequency with a strong positive sentiment; *'muziek'* (*'music'*) and *'activiteiten'* (*'activities'*) occurred with high frequency with a weakly positive sentiment.

## 3.4 Sentiment Analysis in Triads

To illustrate the expressed sentiments of residents, family members, and care professionals, a ternary plot was created (Figure 5.5). It is 3-dimensional, with a positive, negative, and neutral axis. Each point in the triangle represents either a resident (red), a family member (blue), or a care professional (green). As can be seen, most conversations are closely grouped together and slightly above the absolute middle, meaning that they are mostly neutral and almost equally positive and negative. Residents represent the densest group, implying that they expressed rather similar sentiments, that is, neutral, with equal amounts of positive and negative sentences. Family members have a lower density, implying that they expressed a slightly more diverse range of sentiments; this group was on average more negative regarding the quality of care than the other groups. Care professionals cover the largest area and thereby were the most diverse in terms of sentiment expressed; this group was on average more positive regarding the quality of care than the other groups.

## 3.5 Topic Clustering Analysis

To the topic clustering analysis is displayed in Figure 5.6; each keyword is represented as a dot. The axes have no real-world meaning; they are artificially created to highlight the difference between clusters (i.e. topics) [38]. Only the distance between dots has meaning: dots, and thus keywords, which are closer together are semantically more similar compared to dots that are further apart [29, 38]. Dots with the same color belong to the same cluster. Although most keywords belong to one overarching cluster 'quality of care; (which was expected, as all keywords are related to experienced quality of care in a nursing home setting), utterances still show nuanced differences, leading to the discovery of 12 different, but related topics.

For each different cluster the topic name, the most important keywords and the number of sentences that belong to that topic are displayed (see Table 5.1). Certain clusters are well-defined such as *'health'* and *'food'*, while others display more overlap, such as *'care environment'*. This corresponds with Figure 5.6, as it shows certain clusters overlap very little with other clusters. The clusters *'relations'*, *'time'* and *'life experiences'* are the clusters with the most occurring keywords.

**Table 5.1:** Topics from the cluster analysis including the corresponding number of keywords occurring in the interviews (n = 125).

| **#** | **Topic name** | **Example keywords** | # **Occurrences** |
|---|---|---|---|
| 1 | Relations | Daughter, mother, family, children, man | 1995 |
| 2 | Activities | Flower arranging, television, Christmas, music, movie | 996 |
| 3 | Time | Time, start, hour, day, afternoon, week | 2175 |
| 4 | Care organization | Decision, matter, system, organizations, privacy | 1072 |
| 5 | Daily experiences | Meaning, moment, experience, progression, private | 525 |
| 6 | Physical nursing home environment | Room, neighborhood, door, ambulance, walker | 1106 |
| 7 | Health | Parkinson, hallucinations, miscarriage, eye drops, incident | 423 |
| 8 | Food | Dinner, desert, coffee, sandwich, potatoes | 527 |
| 9 | Life experiences | Life, moment, story, feeling, event | 1827 |
| 10 | Care environment | Nursing home, care, education, help, somatic | 1432 |
| 11 | Physical Appearance | Sewing machine, toe, hands, clothes, nightgown | 860 |
| 12 | Miscellaneous | A little, word, small error, things, remainder | 868 |

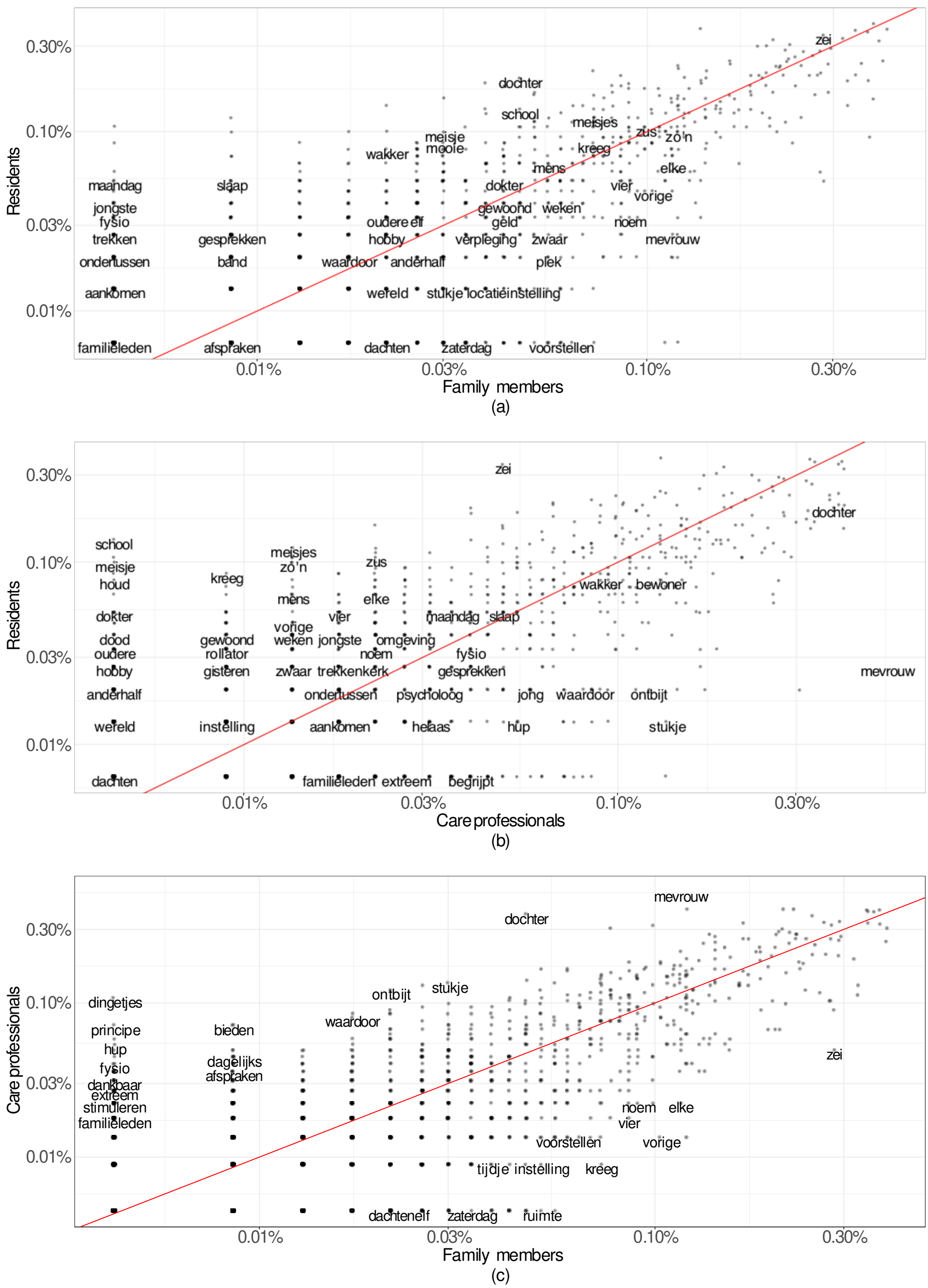
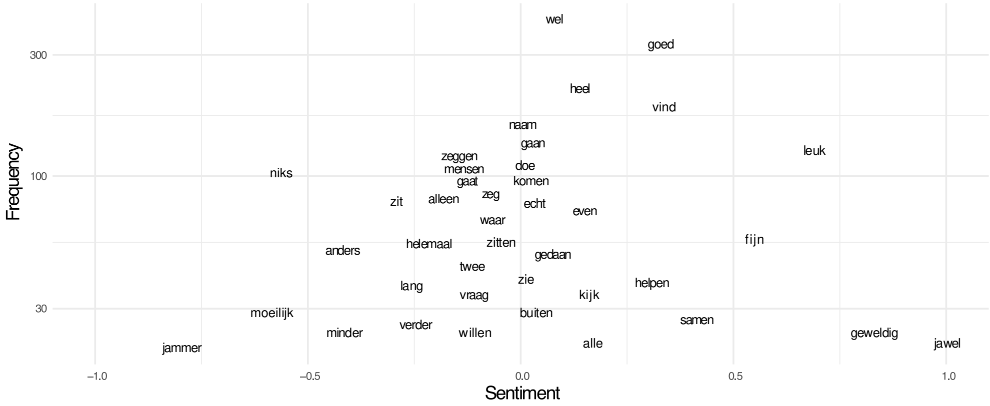
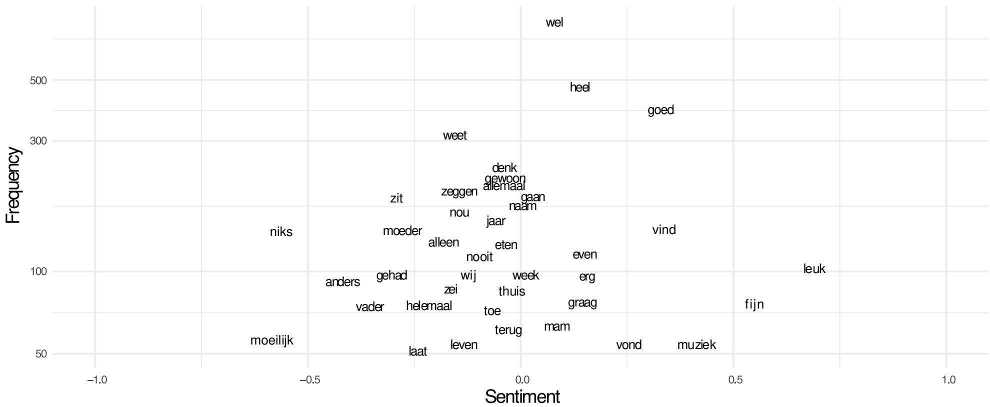


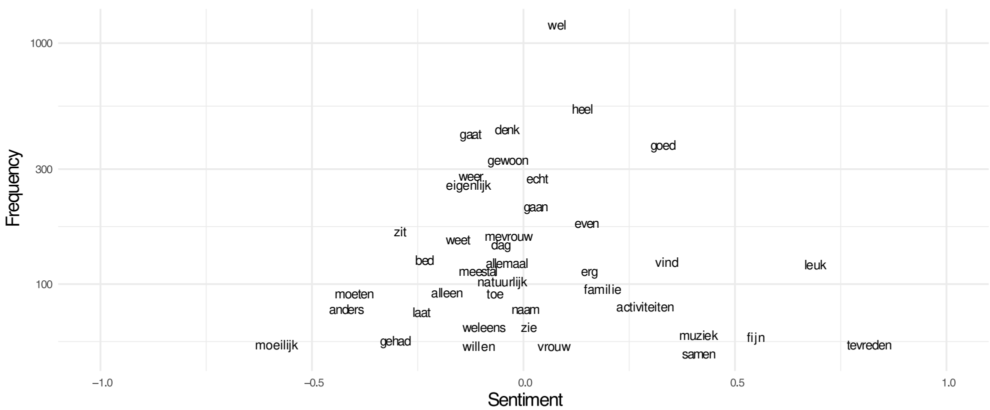

**Figure 5.3:** Bilateral correlations between the words of the 3 groups, respectively: (a) residents and family members; (b) residents and care professionals; and (c) care professionals and family members.




a. Most commonly occurring words among residents (n = 40)




b. Most commonly occurring words among family members (n = 40)




c. Most commonly occurring words among care professionals (n = 40)

**Figure 5.4:** Sentiment analyses: scatter plots of positive and negative words used most often by (a) residents; (b) family members; and (c) care professionals.

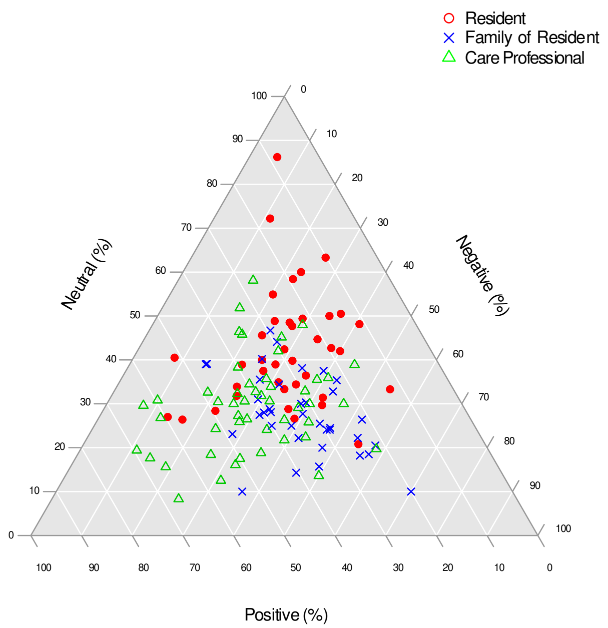



**Figure 5.5:** An overview of the ratios (positive, negative and neutral) for all transcripts.

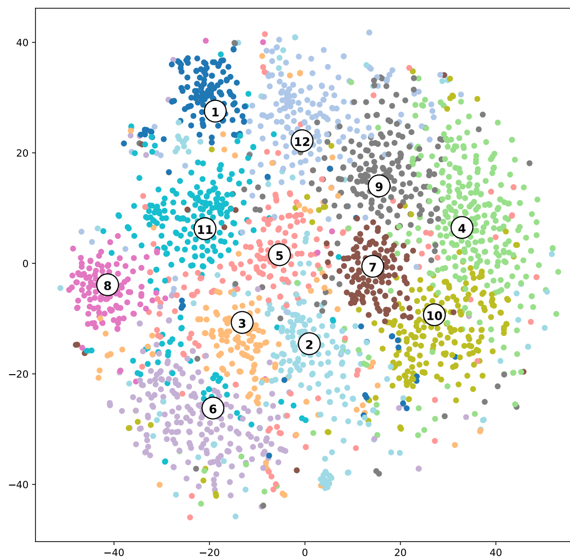


**Figure 5.6:** A visualization of the topic clustering analysis. Clusters are represented by numbers which correspond with the numbers in Table 5.1.

## 4 Discussion

The present study aimed to explore the usefulness of text-mining approaches regarding narrative data gathered in a nursing home setting. The textual information that was automatically gathered from the 125 interviews generated novel insights into the quality of care.

The results showed that the word *'good'* was among the most frequently used words in the interviews, which could indicate that, in general, participants had a positive experience of care. However, it might be argued that individual words have different meanings when preceded by different words (e.g., the word *'good'* can be preceded by the word *'not'* or *'very,'* thereby giving it a different sense). Hence, bigrams (i.e., groups of two consecutive words) were analyzed, and this revealed that the word '*good'* was often preceded by adjectives, indicating magnitude (e.g., *'very good'* or *'very nice'*). These word combinations frequently occurred in the interviews in all three groups, indicating positivity towards the quality of care. Previous research has demonstrated that, when conducting a manual sentiment analysis, words such as *'good'* are indicative of a positive experience regarding quality of care [8]. Correlation analysis showed that the same set of words were used by residents, their family members, and care professionals when discussing the quality of care. These findings imply that the three groups talked about similar topics when discussing the substantive issue. Sentiment analysis highlighted several positive words, including *'muziek'* (*'music'*) and *'activiteiten'* (*'activities'*). Because these words were used frequently in the interviews, it may be inferred that *'music'*and *'activities'* were regarded as important criteria for judging the quality of care [39]. In addition, sentiment analysis showed that the majority of interviews expressed mainly neutral sentiments, though the care professionals were more diverse and positive in their sentiments compared with the other groups [8]. This finding is underscored by previous research that has shown that, in general, care professionals are often more positive than residents or residents' family members concerning the quality of care the residents receive [8, 40]. A topic clustering analysis yielded a variety of topics: while some topics were very clearly defined, including topics such as *'food'* and *'health'*, others were less clearly defined (e.g. *'miscellaneous'*). The large number of occurring keywords related to the topics *'relations'*, *'life experiences'* and *'care environment'* not only highlights the importance of these two topics in relation to experienced quality of care within a nursing home setting [3, 8, 41, 42], but also underscores the validity of the text-mining approach.

The present study is the first to assess quality of care in a long-term care context by analyzing qualitative data through text mining. Making use of the vast amount of text in this way has given a voice to residents, their family members, and care professionals working in nursing homes. However, the study also has several possible limitations. Firstly, the word and bigram frequency analyses only contain absolute numbers. This analyses still contains words and bigrams which have little significance, i.e. words such as 'well' or 'quite', which are less informative. Another limitation is the explainability of the sentiment analysis model which is a 'deep learning' model. Deep learning is the optimization of large models for tasks such as sentiment analysis [33]. While sentiment analysis conducted using a deep learning method often results in more accurate results compared to machine learning models, these latter models can be explained more easily, as unigrams or n-grams (sequences of words) correspond with a certain sentiment prediction. With deep learning models, sentiment can be based on how every word relate to every other word in a text segment. These word relations are calculated from large text datasets and involves many abstract values [33, 43].

### 4.1 Future Work

Future research could focus on combining narrative data with more quantitative measures related to, for example, the prevalence of care problems [44]. This could be achieved by using a text mining approach and various predictive algorithms. It would then be possible to relate narrative data regarding the quality of care to particular care issues (e.g. incontinence and malnutrition), thereby providing a more comprehensive view of quality of care.

Future research could also aim to further explore the text-mining approaches used in the current study. By comparing the text mining approach against the current gold standard of manual coding, the text-mining approach could not only be validated but perhaps also improved. For example, by further improving topic clustering, it may become possible to automate the processes of qualitative coding (i.e. the analyses of qualitative data). As a consequence, analyzing qualitative data may become less time-consuming and more objective.

## 5 Conclusion

To make use of the ever-growing amount of textual data related to the quality of care in long-term older persons' care, innovative and efficient methods are needed. The present study demonstrates the usefulness of a text-mining approach to extend our knowledge regarding the quality of care in a nursing home setting. With the shift to more collections of textual (narrative) data, text-mining in long-term elderly care can lead to valuable new insights that would not have been found using manual analysis.

## 6 Acknowledgments

The authors would like to thank all the contributors of Python, R and the R and Python packages that were used for this study (a full list can be found at `https://github.com/coen22/Text-Mining-AWO-L`). Especially thanks to the contributors of the following packages: *'ggplot package'*, *'tm package'*, *'stopwords package'*, *'Ternary package'*, and the *R-markdown package*.

## Bibliography


[1] J. Pols, "Enacting appreciations: Beyond the patient perspective," *Health Care Analysis*, vol. 13, no. 3, pp. 203–221, 2005.

[2] K. Sion *et al.*, "The feasibility of connecting conversations: A narrative method to assess experienced quality of care in nursing homes from the resident's perspective," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 5118, 2020.

[3] K. Y. Sion *et al.*, "Experienced quality of post-acute and long-term care from the care recipient's perspective–a conceptual framework," *Journal of the American Medical Directors Association*, vol. 20, no. 11, pp. 1386–1390, 2019.

[4] K. Y. Sion *et al.*, "How to assess experienced quality of care in nursing homes from the client's perspective: Results of a qualitative study," *BMC geriatrics*, vol. 20, no. 1, pp. 1–12, 2020.

[5] M. Murray *et al.*, "Narrative data," *The SAGE handbook of qualitative data collection*, pp. 264–279, 2018.

[6] K. Y. Sion, H. Verbeek, S. M. Zwakhalen, G. Odekerken-Schröder, J. M. Schols, and J. P. Hamers, "Themes related to experienced quality of care in nursing homes from the resident's perspective: A systematic literature review and thematic synthesis," *Gerontology and Geriatric Medicine*, vol. 6, p. 2 333 721 420 931 964, 2020.

[7] V. Club, *The experiences of seniors during the corona/covid-19 crisis*, 2020. [Online]. Available: `https://www.leydenacademy.nl/the-experiences-of-seniors-during-the-corona-covid-19-crisis/`.

[8] K. Sion *et al.*, "The validity of connecting conversations: A narrative method to assess experienced quality of care in nursing homes from the resident's perspective," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 5100, 2020.

[9] A. Scheffelaar, M. Hendriks, N. Bos, K. Luijkx, and S. Van Dulmen, "Protocol for a participatory study for developing qualitative instruments measuring the quality of long-term care relationships," *BMJ Open*, vol. 8, no. 11, 2018.

[10] A. Scheffelaar, M. Janssen, and K. Luijkx, "The story as a quality instrument: Developing an instrument for quality improvement based on narratives of older adults receiving long-term care," *Int J Env Res Pub He*, vol. 18, no. 5, 2021.

[11] A. Strauss and J. Corbin, *Basics of qualitative research*. Sage publications, 1990.

[12] A. Strauss and J. Corbin, *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA, 1998.

[13] V. Dhar, "Data science and prediction," *Commun ACM*, vol. 56, no. 12, pp. 64–73, 2013.

[14] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining.," in *Ldv Forum*, Citeseer, vol. 20, 2005, pp. 19–62.

[15] M. Hofmann and A. Chisholm, *Text mining and visualization: case studies using open-source tools*. CRC Press, 2016, vol. 40.

[16] F. Popowich, "Using text mining and natural language processing for health care claims processing," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 1, pp. 59–66, 2005.

[17] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, "Text mining in healthcare. applications and opportunities," *J Healthc Inf Manag*, vol. 22, no. 3, pp. 52–6, 2008.

[18] F. R. Lucini *et al.*, "Text mining approach to predict hospital admissions using early medical records from the emergency department," *International journal of medical informatics*, vol. 100, pp. 1–8, 2017.

[19] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar, and G. Srivastava, "An accurate deep learning model for clinical entity recognition from clinical notes," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3804–3811, 2021.

[20] A. H. Azeemi and A. Waheed, "Covid-19 tweets analysis through transformer language models," *arXiv preprint arXiv:2103.00199*, 2021.

[21] L. Abualigah, H. E. Alfar, M. Shehab, and A. M. A. Hussein, "Sentiment analysis in healthcare: A brief review," in *Recent Advances in NLP: The Case of Arabic Language*, Springer, 2020, pp. 129–141.

[22] M. Nolan, J. Brown, S. Davies, J. Nolan, and J. Keady, "The senses framework: Improving care for older people through a relationship-centred approach. getting research into practice (grip) report no 2.," 2006.

[23] G. Bushe, "Appreciative inquiry is not about the positive," *OD practitioner*, vol. 39, no. 4, pp. 33–38, 2007.

[24] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[25] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2013.

[26] G. Van Rossum and F. L. Drake Jr, *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam, 1995.

[27] J. B. Lovins, "Development of a stemming algorithm," *Mech. Transl. Comput. Linguistics*, vol. 11, no. 1-2, pp. 22–31, 1968.

[28] A. Jonker, “Bag & tag’em, a new dutch stemming algorithm,” 2019.

[29] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” *arXiv preprint arXiv:1301.3781*, 2013.

[30] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[31] R. A. Adams and C. Essex, *Calculus: a complete course*. Addison-Wesley Boston, 1999, vol. 4.

[32] S. Schrauwen, “Machine learning approaches to sentiment analysis using the dutch netlog corpus,” *Computational Linguistics and Psycholinguistics Research Center*, pp. 30–34, 2010.

[33] P. Delobelle, T. Winters, and B. Berendt, “RobBERT: A Dutch RoBERTa-based Language Model,” in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 3255–3265.

[34] M. R. Smith, “Ternary: An r package for creating ternary plots,” 2020.

[35] H.-H. Bock, “Clustering methods: A history of k-means algorithms,” *Selected contributions in data analysis and classification*, pp. 161–172, 2007.

[36] M. Syakur, B. Khotimah, E. Rochman, and B. D. Satoto, “Integration k-means clustering method and elbow method for identification of the best customer profile cluster,” in *IOP conference series: materials science and engineering*, IOP Publishing, vol. 336, 2018, p. 012 017.

[37] W. de Vries, M. Bartelds, M. Nissim, and M. Wieling, “Adapting monolingual models: Data can be scarce when language similarity is high,” in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, Online: Association for Computational Linguistics, Aug. 2021, pp. 4901–4907.

[38] L. Van der Maaten and G. Hinton, “Visualizing data using t-sne.,” *Journal of machine learning research*, vol. 9, no. 11, 2008.

[39] S. H. Tak, S. Kedia, T. M. Tongumpun, and S. H. Hong, “Activity engagement: Perspectives from nursing home residents with dementia,” *Educ Gerontol*, vol. 41, no. 3, pp. 182–192, 2015.

[40] H. Brodaty, B. Draper, and L.-F. Low, “Nursing home staff attitudes towards residents with dementia: Strain and satisfaction with work,” *J Adv Nurs*, vol. 44, no. 6, pp. 583–590, 2003.

[41] M. J. Koren, “Person-centered care for nursing home residents: The culture-change movement,” *Health Affairs*, vol. 29, no. 2, pp. 312–317, 2010.

[42] C. Kingsley and S. Patel, “Patient-reported outcome measures and patient-reported experience measures,” *BJA Education*, vol. 17, no. 4, pp. 137–144, 2017.

[43] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[44] V. A. Huppertz, G.-J. van der Putten, R. J. Halfens, J. M. Schols, and L. C. de Groot, “Association between malnutrition and oral health in dutch nursing home residents: Results of the lpz study,” *JAMDA*, vol. 18, no. 11, pp. 948–954, 2017.

# CHAPTER 6

## Comparing Text Mining and Manual Coding Methods in Analyzing Interview Data on Quality of Care in Long-Term Care for Older Adults

This chapter was published as:

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2023). Comparing text mining and manual coding methods: Analysing interview data on quality of care in long-term care for older adults. PLOS ONE, 17(8), e0268281.

https://doi.org/10.1371/journal.pone.0292578

## Abstract

**Objectives**: In long-term care for older adults, large amounts of text are collected relating to the quality of care, such as transcribed interviews. Researchers currently analyze textual data manually to gain insights, which is a time-consuming process. Text mining could provide a solution, as this methodology can be used to analyze large amounts of text automatically. This study aims to compare text mining to manual coding with regard to sentiment analysis and thematic content analysis.

**Methods**: Data were collected from interviews with residents (n=21), family members (n=20), and care professionals (n=20). Text mining models were developed and compared to the manual approach. The results of the manual and text mining approaches were evaluated based on three criteria: accuracy, consistency, and expert feedback. Accuracy assessed the similarity between the two approaches, while consistency determined whether each individual approach found the same themes in similar text segments. Expert feedback served as a representation of the perceived correctness of the text mining approach.

**Results**: An accuracy analysis revealed that more than 80% of the text segments were assigned the same themes and sentiment using both text mining and manual approaches. Interviews coded with text mining demonstrated higher consistency compared to those coded manually. Expert feedback identified certain limitations in both the text mining and manual approaches.

**Conclusions and Implications**: While these analyses highlighted the current limitations of text mining, they also exposed certain inconsistencies in manual analysis. This information suggests that text mining has the potential to be an effective and efficient tool for analysing large volumes of textual data in the context of long-term care for older adults.

# 1 Introduction

In recent years, client perspectives have become increasingly important in long-term care for older adults (LTC) when assessing the quality of care [1–3]. To gain insight into these perspectives, textual data are often collected, such as electronic health records, policy documents or transcribed interviews with various stakeholders, including residents of nursing homes [2, 4]. When interviews are conducted with stakeholders in nursing homes, textual data may be collected by transcribing audio recordings verbatim from interviews (i.e. literally translating voice into text), and are often referred to as transcripts. This type of data collection often results in large amounts of textual data. To be able to analyse these data, researchers often conduct a so-called coding analysis, which involves manually analysing each transcript (stemming from an interview) to identify text fragments that are relevant to the objective at hand (often a research question) [2, 5]. Each key fragment is summarised using codes (i.e. summaries of several words) that reflect the condensed meaning of that specific fragment [5]. The codes are then clustered based on their similarity, and are grouped into themes [5]. These themes convey a certain topic which is of relevance to the transcript at hand, which often provides a direct or indirect answer to the research question [5]. Although this type of coding is typically performed in a bottom-up manner, it is also possible to apply a top-down approach, in which case a set of themes is constructed in advance [3]. Since text analysis through coding is known to be very time-consuming and prone to bias due to the subjectivity of the researchers, coding is often performed independently by two or more researchers, thereby ensuring a certain level of objectivity. Manual analysis is never completely objective, as researchers are prone to human biases such as generalisations, inferences, and interpretations [6, 7]. which compromise the reproducibility and limit the amount of data that can be analysed.

To overcome the aforementioned drawbacks, text mining could offer a possible solution. Text mining is the process of transforming unstructured text into structured data in order to gain new information and knowledge [8]. and has already been used for knowledge discovery in other domains of health care [4, 9–13]. Knowledge discovery is the process of extracting useful information from a collection of data; for example, a study conducted on electronic health records discussed how text mining could be used to group pathology reports and discharge summaries, based on similar word occurrences [10]. Another study that focused on organising clinical narratives concluded that text mining could be applied to clinical narratives to identify keywords that could help in classifying physiotherapy treatments [4]. These examples highlight the usefulness of text mining in the health care domain.

Recent advancements in the field of text mining have ushered in a variety of new techniques, each with its unique focus and application [14–19]. Some models are particularly good at generating context-aware, human-like text, while others excel at incorporating multi-modal data, such as text and images, for a more comprehensive analysis [14–16]. Moreover, there is a growing emphasis on adapting these models to run efficiently on consumer-grade hardware [17]. Despite these strides in technology, there are still significant challenges in achieving the level of accuracy required for some tasks, and in many cases, human expertise continues to outperform automated methods [17].

To understand the potential usefulness of text mining for qualitative research in long-term care for older adults, it should be compared to the current gold standard of manual coding [20]. This study aims to compare a text mining approach to a manual approach in terms of accuracy, consistency, and expert feedback. Accuracy is a measure of the degree to which the results from the text mining approach are similar to those of the manual approach, whereas consistency is defined as the degree to which an approach (i.e. text mining or manual) finds the same themes for similar pieces of text. Expert feedback is collected to show whether the analyses conducted through text mining are perceived to be correct.

## 2 Materials and Methods

In this study, a comparison was conducted between the use of manual and text mining approaches in a sentiment analysis and a thematic content analysis of qualitative data accumulated in an LTC setting. Two different text mining models were constructed: (i) a sentiment analysis model, and (ii) a thematic content analysis model [21, 22]. Each model was then compared to the respective manual coding approach, based on an accuracy evaluation, a consistency evaluation and expert feedback.

### 2.1 Sample and Participants

Data were collected as part of a project entitled 'Connecting Conversations', which aimed to assess the experienced quality of care in nursing homes from different perspectives [2, 23]. This was achieved by interviewing residents, family members and care professionals at different nursing homes in the South of Limburg [2, 23].

A total of n = 250 interviews were conducted at five different LTC organizations in the southern part of the Netherlands. From those interviews, 234 were transcribed (16 could not be transcribed due to poor audio quality). From the remaining 234 interviews, 61 were analysed manually using thematic content analysis. In addition, 103 interviews were analysed manually using sentiment analysis. All analysis in the manuscript were performed using those 61 and 103 interviews for the thematic content analysis and sentiment analysis respectively.

All interviews were conducted between January 2018 and December 2019. A diverse set of wards were included, including those for older people with dementia [23]. A total of n = 35 interviewers conducted the interviews. These interviewers were part of the project ’Connecting Conversations,’ which aims to assess the experienced quality of care in nursing homes from the resident’s perspective. They primarily come from a long-term care setting and have received specialized training to conduct these interviews. For a more comprehensive understanding of the ’Connecting Conversations’ project, see Sion et. al. 2020a [2]. The medical ethical committee of Zuyderland (the Netherland) approved the study protocol (17-N-86). Information about the study was provided to all interviewers, residents, family members and caregivers by an information letter. All participants provided written informed consent: residents with legal representatives gave informed consent themselves (as well as their legal representatives) before and during the conversations.

## 2.2 Data

The interviews were anonymously collected in the form of audio recordings and were transcribed verbatim (in Dutch) [2]. Personally identifiable information was removed from the transcripts before being coded. The data were coded by three research experts, each working in the *Living Lab on Ageing and Long-Term Care* for over 5 years. All these experts have a minimum of ten years of experience in conducting qualitative research. A total of 103 transcripts were manually coded regarding the sentiment [24]. In this analysis, text segments were manually coded as being either ’positive’ or ’negative’. However, text segments were only coded if the text discussed a topic relevant to the nursing home. A total of 61 transcripts were manually coded using INDEXQUAL, a thematic framework for defining the quality of LTC [3]. The themes provided by INDEXQUAL are ’context’, ’nursing home’, ’person’, ’expectations’, ’personal needs’, ’past experiences’, ’word of mouth’, ’experiences’, ’care environment’, ’relationship-centred care’, ’experienced quality of care’, ’perceived care services’, ’perceived care outcomes’ and ’satisfaction’ [2, 3]. In both cases, transcripts were coded using MAXQDA, and these codes were exported to develop a text mining approach [25].

## 2.3 Text mining models

The models presented in the current study were created using deep learning, a method in which artificial neural networks (ANNs) are used to learn automatically from input data [26]. A Dutch base language model called RobBERT was used [22]. The advantage of using such a model is that language knowledge can be learned from a large dataset of arbitrary (Dutch) text. Two models were developed in the current study: a sentiment analysis model, and thematic content analysis model. The code for the models can be found at: `https://doi.org/10.5281/zenodo.8391747`.

### Sentiment analysis

Sentiment analysis is the process of computationally identifying the sentiment expressed in a piece of text [8, 27]. For example, the sentence 'It's a good day' could be identified as being positive, while the sentence 'It's a bad day' could be identified as being negative. The sentence 'Today $I$ went for a walk,' could be neutral, as it does not convey whether the walk was experienced as a positive or negative event. Coded text segments were passed directly as input to the model, without modification. The sentiment analysis model was trained to classify the sentiment of a given piece of text into one of two categories, i.e. positive or negative. A positive or negative code was only assigned when it was perceived as being relevant to improving the quality of care [24].

### Thematic content analysis

As part of the thematic content analysis, the model was trained to identify the themes present in a given piece of text and to classify them into the relevant themes of the INDEXQUAL coding scheme. Since the number of coded text segments (n = 3867) was insufficient to allow the model to learn all of the themes and sub-themes (n = 16), only the main themes were used: 'Experienced quality of care', 'Experiences', 'Expectations' and 'Context' [3]. Each code containing a subtheme was changed to one of these main themes, and the model was designed to be able to identify multiple themes that may be present in a text segment.

## 2.4 Evaluation

The text mining models were analysed in three ways: an accuracy evaluation, a consistency evaluation, and using expert feedback. The accuracy analysis assessed the ability of each model to correctly classify or predict outcomes based on the input data, while the consistency analysis evaluated their ability to produce consistent results over multiple runs or when applied to different datasets, and expert feedback was used to provide additional insight into the performance and potential biases of the models [28–30].

### Accuracy

The accuracy evaluation aimed to calculate the percentage of text segments that were assigned the same codes in both the text mining approach and the manual approach [8, 27, 31]. For example, if the text mining model for sentiment analysis assigned the same sentiment as the manual approach for all of the sentences, then the model would be considered 100% accurate. To calculate the accuracy, training and validation sets were used: the training set was used to provide feedback to the model to help improve it (i.e. supervised learning), while the validation set was used to evaluate whether what the model had learned so far could be generalised to data that it had not had the chance to learn from [28]. The total amount of data was split, with 90% forming the training set and 10% the validation set. The accuracy score from the validation set was reported, as this is more representative of how a model would perform on unseen data [28]. A confusion matrix was used to display the results of the accuracy evaluation. Such a matrix shows the different cases for each possible choice that either the manual or text mining approach can make. Accuracy was calculated using the formula: (TP + TN) / (TP + TN + FP + FN). In this case, TP is the true positive (i.e. where a code is present in both analyses), TN is the true negative (i.e. where a code is absent in both analyses), and FP is the false positive (i.e. where a code is predicted to be present but is absent in the manual analysis), while FN is the false negative (i.e. where a code is predicted to be absent but is present in the manual analysis). These components help us assess the accuracy of the model's predictions and its performance overall [28].

### Consistency

In the consistency evaluation, both the manual and text mining approach were analysed to determine the consistency of each approach individually. When a coded text is consistent, the expected outcome is that each sentence that is semantically similar will be coded in the same way. A consistency evaluation was conducted by comparing the assigned themes or sentiment between similar sentences; for example, if two sentences were semantically very similar, then it would be expected that these sentences would also be coded with the same themes, and if two sentences were semantically very different, it would be less likely that these would be coded in the same way [30, 32].

#### Expert feedback

To determine whether the output of the models was reliable and comparable to that of manual coding, feedback was collected from the original research experts. This information was collected from three of the research experts who coded the original data, all of whom worked at the Living Lab on Ageing and Long-Term Care for over 5 years. All their feedback was captured in an audio-recorded interview.

The research experts were shown three coded transcripts and were asked to give feedback on them. Without their knowing, the research experts were shown one transcript that had been left unmodified manually coded transcripts (i.e. a transcript that contained the codes as previously analysed by the research experts themselves). After being shown each individual transcript, the research experts were asked to provide feedback on that transcript overall. Their feedback was then analysed to discover potential issues with the text mining approach.

Following this, the research experts were given one large transcript from the validation set in which they were shown both the manual and text mining versions next to each other. This type of comparison allowed them to comment on why the differences between the approaches arose. Their feedback was also used to highlight issues with the accuracy analysis.

# 3 Results

## 3.1 Accuracy

#### Sentiment analysis

The results show that the overall accuracy for the sentiment between the manual approach and the model was 81.8%. Figure 6.1 displays the results of the sentiment analysis in the form of a confusion matrix. It can be seen from the figure that most of the text in the transcripts was not coded with a sentiment, either through the manual process or through text mining. Manually coded text with a negative sentiment was only recognised as positive by text mining in 0.1% of cases, and only 0.3% of the text that was manually coded with a positive sentiment was recognised by text mining as negative. The average accuracy over all transcripts was 88.7% with standard deviation of 8.6%. The minimum was accuracy was 52.1% and the maximum accuracy was 99.6%.

**Text Mining**

| Manual Analysis | Negative | - | Positive |
|---|---|---|---|
| Negative | 1.5 | 3.1 | 0.3 |
| - | 4.3 | 77.1 | 5.2 |
| Positive | 0.1 | 5.1 | 3.2 |

**Figure 6.1:** Confusion matrix comparing sentiment analysis results of the manual and text mining approach. The matrix compares manual coding (rows) against text mining predictions (columns) for sentiment values of the text. Each cell within the matrix represents the percentage occurrence of a particular sentiment alignment (or misalignment) between the manual and text mining approaches. The y-axis of each matrix represents the sentiment as determined through manual analysis, while the x-axis indicates the text mining predictions. The diagonal cells (from top left to bottom right) illustrate the percentage of agreement between the two methods, whereas all off-diagonal cells indicate discrepancies. For instance, the cell at the intersection of the ”Positive” row and the ”Negative” column displays instances where text was manually coded as positive but was predicted as negative by text mining.

### Thematic content analysis

A comparison was conducted between the manually coded INDEXQUAL themes and the codes predicted by the model, and the results indicated that the model achieved an accuracy of 83.7%. Figure 6.2 shows the confusion matrices for the validation set. For all of the themes in general, it was found that most of the text segments that weren't coded by the manual approach, were also not coded by the text mining approach. For the theme 'Context', we found that the text mining approach assigned a code to a text segment much more often compared to the manual approach. The themes of 'Context' and 'Expectations' were absent from most of the manually coded text (in 87.9% and 95.2% of cases, respectively). The themes of 'Experienced Quality of Care' and 'Experiences' were identified correctly by the text mining approach in a higher percentage of text segments compared to 'Context' and 'Expectations'; however, 'Experienced Quality of Care' and 'Experiences' also had higher rates of false positives and false negatives. False positives were cases where text mining incorrectly assigned a particular theme to text segment, and false negatives were cases where text mining incorrectly failed to assign a theme. "The average accuracy over all transcripts was 81.9% with a standard deviation of 8.5%. The minimum accuracy of any transcript was 43.1% and the maximum was 93.4%.

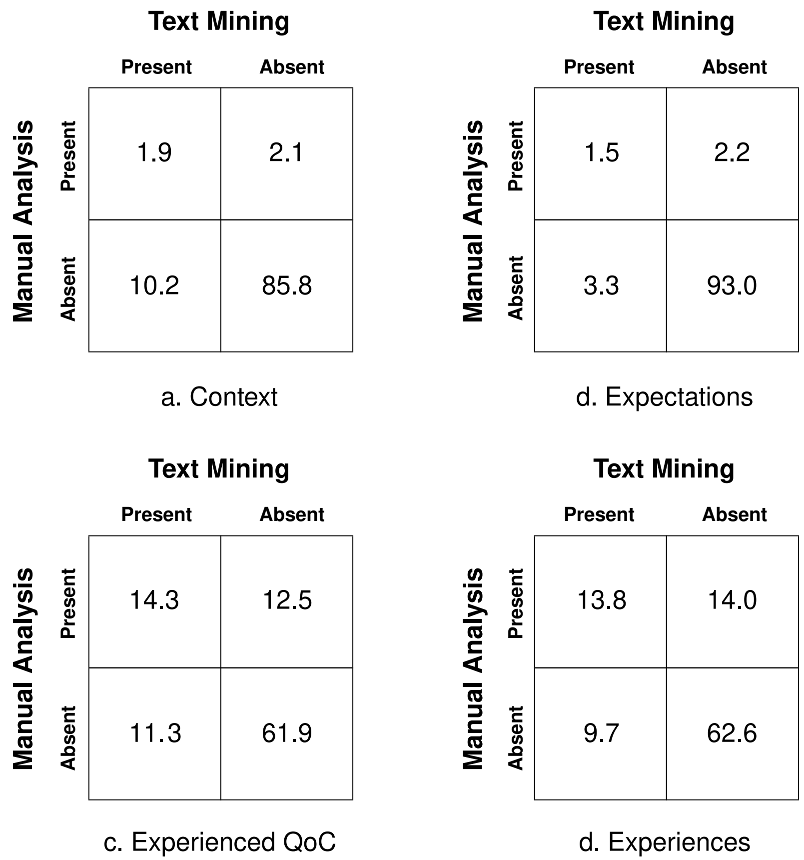



**Figure 6.2:** Comparison of results from the thematic content analysis. A confusion matrix is shown for each of the main INDEXQUAL themes (Experienced quality of care, Experiences, Expectations and Context). The y-axis of each matrix represents the presence or absence of a theme as determined through manual analysis, while the x-axis indicates the text mining predictions. Cells on the diagonals capture instances of agreement between manual coding and text mining for each theme. Off-diagonal cells detail discrepancies, indicating false positives or false negatives. Percentages within cells show the proportion of occurrences for each scenario in relation to the total dataset.

## 3.2 Consistency

### Sentiment analysis

Consistency scores were calculated as part of the sentiment analysis, as shown in Table 6.1. Semantic similarity is a value between 0% and 100%, where a higher percentage indicates that the results were more consistent [33]. On average, the transcripts coded using the sentiment analysis model were more consistent than those coded using the manual approach.

**Table 6.1:** Overview of the consistency of the manual and text mining approaches in regarding the sentiment analysis.

| Theme | Manual (%) | Text mining (%) |
|---|---|---|
| Positive | 68.3 | 74.4 |
| Negative | 67.6 | 73.8 |

### Thematic content analysis

As is shown in Table 6.2, the text mining approach was more consistent when coding sentences related to experienced QoC and experiences. These were also the themes that occur most often in the interviews. On average the text mining approach was more consistent using the current metric. While, the results displayed a low consistency, it should be noted that only limited context was taken into account. This increased the perceived similarity of sentences and therefore decreases the consistency.

**Table 6.2:** Overview of the consistency of the manual and text mining approaches regarding various themes.

| Theme | Manual (%) | Text mining (%) |
|---|---|---|
| Experienced QoC | 51.8 | 58.9 |
| Experiences | 54.0 | 59.1 |
| Expectations | 59.5 | 61.8 |
| Context | 59.4 | 62.2 |
| Average | 56.2 | 60.5 |

## 3.3 Expert feedback

Overall, the research experts expressed a mixed-to-positive assessment of the analysis of the transcripts. While they were most positive about the manually coded transcript, they were unable to distinguish it from the transcript coded by the text mining algorithm in the training set. In contrast, the text mining approach in the validation set was recognized by the research experts as having a lower level of accuracy (e.g., smaller coded text segments compared to the manual codes). The research experts identified certain themes, such as "Context" and "Expectations," as posing greater difficulties for the algorithm, whereas other themes, such as "Experienced Quality of Care" and "Experiences," were coded more similarly by both the algorithm and the research experts. The experts acknowledged that coding was generally a challenging task.

> "I don't find the coding to be poor. I notice that the codes about which the text mining approach is wrong, we've also had deliberations."

> "[Text mining] isn't not all perfect, however it does allow us to analyse much more interviews."

The research experts were presented with a transcript from the validation set, where both the manual and text mining versions were presented side by side to enable the research experts to explain the differences between the approaches. Most of the feedback from the research experts focused on codes that were similar between the two approaches or where the text mining approach incorrectly coded something. However, according to the experts, some codes were coded correctly by the text mining approach, but not by the manual approach.

> "Yes, we've missed that one, seems logical to me."

> "Yes, [similar to the other] we missed that one as well."

Although the instances of text mining finding errors in the manual codes were few, they negatively impacted the accuracy analysis. This is because such codes were regarded as false positives. Additionally, there was at least one instance where the text mining algorithm had coded the same information at a different location in the text.

> "Here, the model applied the theme of quality of care [instead of where we coded it]."

## 4 Discussion

This study compared two approaches to coding text, a text mining approach and a manual approach, and carried out two types of analysis: a sentiment analysis and a thematic content analysis. The two approaches were compared in terms of their accuracy and consistency, and based on expert feedback. The results showed that for most text segments, the approaches were coded in a similar fashion. However, further analyses also showed that there were key differences in coding between the text mining approach and the manual approach in terms of accuracy and consistency.

The results of an accuracy analysis showed that the text mining models coded text with the same themes as the manual approach in more than 80% of cases. However, it was found that the number of false positives and false negatives were relatively high compared to the true positives. This indicates that the actual similarity (i.e. for text containing more coded segments) between the methods may be lower. One of reasons for the discrepancies between the manual and text mining approaches is that many manually coded text segments contain more than one theme; for example, 19% of all of the text coded with the theme 'Experiences' was also coded by the research experts with other themes, such as 'Experienced quality of care' or 'Expectations'. The presence of overlapping themes in text can pose a challenge for text mining models, as this makes it more difficult to accurately determine which text characteristics correspond to each theme. In addition, the complexity and variability of natural language and the current limitations of text mining algorithms may also contribute to the lower accuracy of text mining models [34–37]. The variance of the accuracy between transcripts shows that a possible reason for lower accuracies could be due to factors that vary between transcripts , such as the quality of the transcription, the nature of the language used by the participants, or contextual factors that were not taken into account by the text mining or manual approach.

The results of a consistency analysis suggested that the current text mining models were able to produce more consistent codes for semantically similar sentences across all interviews compared to the manual analyses. However, the measured difference in consistency between the approaches was less than 5% on average. This could be explained by the fact that the text mining approach learned from the manual codes, and hence the text mining models also exhibited the same type of inconsistencies to a certain degree [38, 39].

supplement to traditional qualitative analysis methods, and could provide a more efficient and objective way of analysing large amounts of text data [40, 41].. However, research experts were able to identify flaws in both methods of analysis. This could be because research experts had more knowledge about the subject of the analyses and could therefore recognise wider patterns [42, 43]. However, it was difficult for human experts to distinguish between the codes they had assigned manually and codes that were assigned by the text mining model. When the experts were able to compare the codes created by the text mining approach and their own manual codes, they reported that they had also missed certain text segments when they originally coded the interviews. These segments were discovered and coded by the text mining models. This finding suggests that text mining models could be helpful for manual analysis, as demonstrated using recent methods such as InstructGPT and MM-CoT [14–16]. These methods show that language models can aid in a variety of tasks, from writing cover letters to creating SPSS or Python scripts. However, these language models require human guidance to achieve the best results, as many of these tasks may be subject to human bias [38].

Using deep learning models, such as those highlighted in this study, offers a distinct advantage in terms of speed. While deep learning models can process and analyse data within seconds, manual analysis, depending on the complexity and volume of the data, can span weeks to even months [44]. However, it's essential to recognize that the results from deep learning models might not always align perfectly with those of manual analysis. As such, researchers might find the need to fine-tune the outputs generated by text mining models. Despite this, the integration of deep learning significantly accelerates the qualitative analysis process, offering a more efficient alternative to traditional methods.

Some methodological limitations must be acknowledged. Firstly, in large parts of the interviews, no codes were identified by either the text mining or manual approach. As a result, the average accuracy of the text mining models was higher than it might have been if the text were coded with a higher density. Secondly, it is important to consider the limitations of the algorithm used to calculate the sentence similarity. This algorithm has an accuracy that is limited to 66% for the classification of similar sentences [30]. This is also challenging, as it is therefore difficult to define which properties of a text segment are important in terms of the semantic similarity. For example, given four sentences regarding a resident, a nurse, a resident's family member, and a visiting doctor, it is possible to split them based on whether a person is a healthcare professional or not; however, it is also possible to split them based on whether a person is part of the nursing home staff or an outsider. Which property is more important to the similarity depends on factors such as the research question, and determining the similarity becomes more difficult with complex sentences. Moreover, it is important to consider the potential for human bias in qualitative analysis. Bias can arise from a variety of sources, including the research expert's own preconceptions and assumptions, the sampling and recruitment of participants, and the methods and techniques used to collect and analyse data [36, 37]. As the text mining model learns from inherently subjective data, it also learns to apply codes with the biases that exist in the data. While the expert feedback showed that few of these cases existed, such cases can negatively impact the evaluated accuracy of text mining models. Lastly, the analysis conducted in the current study only had context window of 512 words at most, which represents a technical limitation of the method [21, 22]. This limits the textual context that the models have access to. These issues can be mitigated using large language models that are better able to capture the nuances and complexities of natural language (e.g. GPT-3) [34, 45]. Such models can also handle a larger context of words. Whereas RobBERT has a maximum context length of 512, GPT-3 has a context of 4,096. However, such large language models cannot be used on most personal computers, as they require specialised hardware to run efficiently (i.e. GPUs or TPUs with large amounts of memory) [46]. Using these via online (cloud) systems could give rise to issues regarding the privacy of the interview participants. However, recent advances have shown that 'smaller' (i.e. more efficient) large language models can achieve similar results, and these models can be used on personal computers, unlike GPT-3 [17, 47].

### 4.1 Future Work

Future research could focus on applying a hybrid approach that combines the text mining and manual methods. Using this approach, a text mining algorithm could be used to pre-process the text data and identify potential themes and patterns, which could then be reviewed and refined by human experts. This would allow for an efficient and objective analysis of large datasets, while also allowing for the expertise and knowledge of human experts to be incorporated. Future research should investigate whether this approach could help to reduce the potential for bias and improve the accuracy of the results.

Future work could compare multiple novel text mining models such as GPT-4 and LLaMA to show whether larger models can generate results that are better with respect to the context and more similar to the manual analysis. Comparing different models side-by-side could offer a useful way to visualize the main features and capabilities of each model, and could also facilitate the identification of any common weaknesses or limitations that may exist across some or all of the models being investigated. This could also enable the identification of areas where specific models may excel relative to others.

## 5 Conclusions

The current study shows that text mining can be an effective tool for quickly and accurately identifying sentiment and thematic content from large amounts of textual data. Text mining can help to reduce the amount of time and resources needed to analyse textual data, making it a valuable tool for analysing large amounts of qualitative data. However, as shown in the current study, text mining has certain limitations regarding language understanding; in its current state, text mining is no substitute for manual coding, but can be seen as a helpful addition.

## 6 Acknowledgements

The authors would like to thank the Data Science Research Infrastructure (DSRI) at Maastricht University. Without the usage of their computational resources, the current study could not have been conducted. Moreover, thanks to the contributors of the “huggingface transformers” library, as the library provided many of the components for developing the ASR model in the current study. Lastly, a special thanks to Katya Sion, Audrey Beaulen and Erica de Vries, the research experts who manually coded the transcripts and gave their feedback.

## 7 Data Availability

The code is now available on Zenodo: `https://zenodo.org/doi/10.5281/zenodo.8391746`. Our interview data will not be publicly available due to the privacy of our participants. Upon request, our interview data may be provided with restrictions. Data are available from the Limburg Living Lab in Ageing and Long-Term Care for researchers who meet the criteria for access to confidential data.

## Bibliography

[1] J. Pols, "Enacting appreciations: Beyond the patient perspective," *Health Care Analysis*, vol. 13, no. 3, pp. 203–221, 2005.

[2] K. Sion *et al.*, "The feasibility of connecting conversations: A narrative method to assess experienced quality of care in nursing homes from the resident's perspective," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 5118, 2020.

[3] K. Y. Sion *et al.*, "Experienced quality of post-acute and long-term care from the care recipient's perspective–a conceptual framework," *Journal of the American Medical Directors Association*, vol. 20, no. 11, pp. 1386–1390, 2019.

[4] T. Delespierre, P. Denormandie, A. Bar-Hen, and L. Josseran, "Empirical advances with text mining of electronic health records," *BMC medical informatics and Decision Making*, vol. 17, no. 1, pp. 1–15, 2017.

[5] A. Strauss and J. Corbin, *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA, 1998.

[6] N. Norris, "Error, bias and validity in qualitative research," *Educational action research*, vol. 5, no. 1, pp. 172–176, 1997.

[7] P. Mackieson, A. Shlonsky, and M. Connolly, "Increasing rigor and reducing bias in qualitative research: A document analysis of parliamentary debates using applied thematic analysis," *Qualitative Social Work*, vol. 18, no. 6, pp. 965–980, 2019.

[8] M. Hofmann and A. Chisholm, *Text mining and visualization: case studies using open-source tools*. CRC Press, 2016, vol. 40.

[9] F. Popowich, "Using text mining and natural language processing for health care claims processing," *ACM SIGKDD Explorations Newsletter*, vol. 7, no. 1, pp. 59–66, 2005.

[10] U. Raja, T. Mitchell, T. Day, and J. M. Hardin, "Text mining in healthcare. applications and opportunities," *J Healthc Inf Manag*, vol. 22, no. 3, pp. 52–6, 2008.

[11] S. A. Moqurrab, U. Ayub, A. Anjum, S. Asghar, and G. Srivastava, "An accurate deep learning model for clinical entity recognition from clinical notes," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 10, pp. 3804–3811, 2021.

[12] A. H. Azeemi and A. Waheed, "Covid-19 tweets analysis through transformer language models," *arXiv preprint arXiv:2103.00199*, 2021.

[13] P. Lee, S. Bubeck, and J. Petro, "Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine," *New England Journal of Medicine*, vol. 388, no. 13, pp. 1233–1239, 2023.

[14] J. Thiergart, S. Huber, and T. Übellacker, "Understanding emails and drafting responses–an approach using gpt-3," *arXiv preprint arXiv:2102.03062*, 2021.

[15] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[16] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," *arXiv preprint arXiv:2302.00923*, 2023.

[17] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[18] B. Percha, "Modern clinical text mining: A guide and review," *Annual review of biomedical data science*, vol. 4, pp. 165–187, 2021.

[19] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[20] H. Song *et al.*, "In validations we trust? the impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis," *Political Communication*, vol. 37, no. 4, pp. 550–572, 2020.

[21] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[22] P. Delobelle, T. Winters, and B. Berendt, "RobBERT: A Dutch RoBERTa-based Language Model," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, Online: Association for Computational Linguistics, Nov. 2020, pp. 3255–3265.

[23] K. Sion *et al.*, "The validity of connecting conversations: A narrative method to assess experienced quality of care in nursing homes from the resident's perspective," *International Journal of Environmental Research and Public Health*, vol. 17, no. 14, p. 5100, 2020.

[24] K. Y. J. Sion *et al.*, "Listen, look, link and learn: A stepwise approach to use narrative quality data within resident-family-nursing staff triads in nursing homes for quality improvements," *BMJ Open Quality*, vol. 10, no. 3, 2021.

[25] V. Software, *Maxqda 2020 online manual*, 2019. [Online]. Available: `https://www.maxqda.com/help-mx20/welcome`.

[26] B. Yegnanarayana, *Artificial neural networks*. PHI Learning Pvt. Ltd., 2009.

[27] A. Hotho, A. Nürnberger, and G. Paaß, "A brief survey of text mining.," in *Ldv Forum*, Citeseer, vol. 20, 2005, pp. 19–62.

[28] Z.-H. Zhou, *Machine learning*. Springer Nature, 2021.

[29] S. B. Kotsiantis, I. Zaharakis, P. Pintelas, *et al.*, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering*, vol. 160, no. 1, pp. 3–24, 2007.

[30] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019.

[31] S. Schrauwen, “Machine learning approaches to sentiment analysis using the dutch netlog corpus,” *Computational Linguistics and Psycholinguistics Research Center*, pp. 30–34, 2010.

[32] W. Yin, J. Hay, and D. Roth, “Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach,” *CoRR*, vol. abs/1909.00161, 2019. arXiv: 1909.00161.

[33] N. Bölücü, B. Can, and H. Artuner, “A siamese neural network for learning semantically-informed sentence embeddings,” *Expert Systems with Applications*, vol. 214, p. 119 103, 2023.

[34] T. Brown *et al.*, “Language models are few-shot learners,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[35] L. Floridi and M. Chiriatti, “Gpt-3: Its nature, scope, limits, and consequences,” *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.

[36] K. L. Easton, J. F. McComish, and R. Greenberg, “Avoiding common pitfalls in qualitative data collection and transcription,” *Qualitative health research*, vol. 10, no. 5, pp. 703–707, 2000.

[37] M. Maycock, “‘i do not appear to have had previous letters’. the potential and pitfalls of using a qualitative correspondence method to facilitate insights into life in prison during the covid-19 pandemic,” *International Journal of Qualitative Methods*, vol. 20, p. 16 094 069 211 047 129, 2021.

[38] B. Kim, H. Kim, K. Kim, S. Kim, and J. Kim, “Learning not to learn: Training deep neural networks with biased data,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9012–9020.

[39] A. Goyal and Y. Bengio, “Inductive biases for deep learning of higher-level cognition,” *Proceedings of the Royal Society A*, vol. 478, no. 2266, p. 20 210 068, 2022.

[40] A. Wang *et al.*, “Superglue: A stickier benchmark for general-purpose language understanding systems,” *Advances in neural information processing systems*, vol. 32, 2019.

[41] Q. Zhong *et al.*, “Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue,” *arXiv preprint arXiv:2212.01853*, 2022.

[42] A. Fan, T. Lavril, E. Grave, A. Joulin, and S. Sukhbaatar, "Addressing some limitations of transformers with feedback memory," *arXiv preprint arXiv:2002.09402*, 2020.

[43] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," *ACM Computing Surveys*, vol. 55, no. 6, pp. 1–28, 2022.

[44] H. Wang *et al.*, "Efficient algorithms and hardware for natural language processing," Ph.D. dissertation, Massachusetts Institute of Technology, 2020.

[45] B. Workshop *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[46] S. Rajbhandari, O. Ruwase, J. Rasley, S. Smith, and Y. He, "Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning," in *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2021, pp. 1–14.

[47] R. Taori *et al.*, *Stanford alpaca: An instruction-following llama model*, `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

# CHAPTER 7

## General Discussion

The aim of this dissertation was twofold. Firstly, to develop tools and methodologies to support data collection and analyses in LTC by automating manual processes. Secondly, to investigate the possibilities of AI in analyzing textual data regarding the quality of care.

## 1 Main Findings

This dissertation first gave an overview of existing data science and AI methods in long-term care with regard to quality of care. The scoping review identified few data science methods used to study or improve quality of care. This is in contrast with data science in medicine, where research regarding AI is much more abundant [1, 2]. The review concluded that despite the potential of these methods to improve the quality of care, their use remains limited in a long-term care for older adults (LTC) setting. The findings suggest that overcoming barriers, such as lack of data infrastructure, expertise, and clear ethical guidelines, are essential steps toward leveraging data science to its full potential in LTC. The review underscores a need for broader adoption and more strategic implementation of data science in LTC.

In Part I of this dissertation, two methodologies were explored to enhance data collection in LTC settings, focusing on automating analysis processes traditionally performed manually. The first study developed an automatic speech recognition (ASR) model tailored to specific accents and speech impairments of older adults. This model significantly reduced the word error rate (WER) compared to general ASR models, demonstrating the efficacy of using domain-specific data to improve speech recognition accuracy with challenging accents. This advancement suggests that ASR technology can be a valuable tool for transcribing interviews and other speech data in LTC, potentially saving significant time and effort for researchers and care staff. The second study focused on the user-centered design (UCD) approach for developing a digital application, specifically transforming the Maastricht Electronic Daily Life Observation (MEDLO) tool from an Excel-based system to a more user-friendly app. This process involved continuous feedback and iterative design with the end-users, including researchers and care professionals. The result was a minimum viable product (MVP) that streamlined data collection and analysis, making it more accessible and less time-consuming for users. This UCD approach proved to be useful in improving the usability of the MEDLO tool. Moreover, it highlighted the importance of involving users in the design process to ensure the developed solutions meet their needs and preferences effectively.

In Part II of this dissertation, the focus shifted to exploring the potential of AI in analyzing data within an LTC setting, specifically looking at how text mining techniques can be applied to understand the quality of care from narrative data collected in nursing homes. Two main studies were presented that demonstrated the usefulness and possibilities of AI in analyzing qualitative data related to the quality of care. The first study explored the application of text mining to narrative data collected from interviews in the project 'Connecting Conversations' with clients, their family members, and care professionals. This study showed that text mining could extend our knowledge regarding the quality of care in a nursing home setting by efficiently processing large amounts of textual data to uncover valuable insights. The second study compared text mining to manual coding methods in terms of analyzing interviews about the quality of care. This comparison revealed that text mining could achieve high accuracy in sentiment analysis and thematic content analysis, with more than 80% of the text segments assigned the same themes and sentiments as manual coding. Additionally, text mining demonstrated higher consistency compared to manual coding, suggesting that AI could offer a more objective and efficient approach to analyzing qualitative data. However, expert feedback identified limitations in both the manual and the text mining approach, highlighting areas for improvement. For example, the text mining approach often coded shorter segments compared to manual coding. These findings suggest that both approaches have complementary strengths and limitations, which could be leveraged to achieve the best results. This shows that text mining can be a valuable tool for analyzing narrative data in LTC.

# 2 Methodological Considerations

This section is dedicated to a reflection on the overarching methodological aspects of this dissertation, specifically focusing on advancements in artificial intelligence, in particular large language models and natural language processing, and the limited testing by care professionals.

## 2.1 Advancements in Language Models

In the last few years, developments in the field of artificial intelligence (AI) have been groundbreaking. This has been shown in a variety of fields, from improvements to autonomous vehicles to AI that controls plasma in nuclear fusion, and AI models that aid in drug discovery [3–5]. Similarly, the subject of large language models (LLMs) has become particularly popular in recent years. LLMs have become available in the form of transformer models such as GPT-4 and Claude 3. These AI models are trained on much more data than the models used in this dissertation, and are thus expected to be able to learn better and faster from manually coded text than the models that are reported on in this dissertation [6]. However, advances in the field are shifting towards smaller models that are more parameter-efficient [7–9]. Parameter efficiency in LLMs refers to optimizing the number of parameters used in the model to achieve high performance while reducing computational power [10, 11]. By fine-tuning the parameters effectively, smaller LLMs can achieve efficiency in terms of speed and memory usage. This optimization is crucial for ensuring that the model is manageable and practical for real-world applications such as a LTC setting, where privacy and low latency are essential [6, 10]. Smaller open-source models such as LLaMA and Mistral have made their appearance [12, 13]. Smaller more parameter-efficient LLMs may be more usable in LTC settings, where the computational resources may be limited or expensive through software providers. These models can then be used in a variety of ways. For instance, language models have already been shown to answer medical questions with high accuracy; this learned medical knowledge could be combined with data from EHRs to provide personalized healthcare [14–16]. The challenge of biases and limitations in models, as noted in the broader context of NLP [17, 18], also applies here, underscoring the importance of carefully training and fine-tuning these models to avoid perpetuating biases in care delivery. Smaller models can also be employed more easily in practice, as they require fewer computational resources and can be more easily integrated into existing systems [7–9]. Despite advancements in current language models, the models used in this dissertation are still relevant, as they are trained on more specific data and are more easily accessible (i.e., due to their smaller size, they can be employed on a broader range of hardware). Even more importantly, this dissertation offers a baseline model for future research and development of LTC-specific models. Consequently, the methods shown in this dissertation can be applied to newer models, and it only requires small adjustments to the code to switch to a different foundation model (e.g., a new model architecture that has been trained on large datasets).

It is important to note that while many LLMs are openly available and can be used for free, it is very expensive (i.e., both in carbon emissions and monetarily) to develop so-called ”foundation models” from scratch due to the required computation resources [19]. A foundation model is ”any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks” [19]. This does create a dependency on technology companies that have the resources to train these models, therefore it is important to keep in mind that these models may not be neutral, and that they can contain biases [19]. That is why it is important for academic institutions to also invest in training and fine-tuning these foundation models.

Despite the ongoing fast-paced research of LLMs, recent research has shown that nuanced expression of sentiment (i.e., subtle differences on a scale from negative to positive) in text is not always recognized by the model [20]. While a recognition of positive or negative sentiment can be recognized with accuracies above

90%, when presented with 5 labeling options from positive to negative, the accuracy falls to around 60% [20]. This suggests that nuanced expressions of sentiment are more difficult to classify. Smaller models, such as RoBERTa, are shown to give similar performance [20]. This may indicate that larger models are not immediately better at analyzing nuanced text, such as interviews in LTC. This could potentially be resolved by using high-quality nuanced data. If this problem is solved, the resulting models may be universally better at coding than humans.

## 2.2 Limited Testing by Care Professionals

While, many of the methods in this dissertation have been directed towards the use of LTC, they have only been tested in a research setting. This limitation poses a significant gap in understanding the practical applicability and usability of the developed methodologies and tools [21–23]. While the involvement of researchers is invaluable for theoretical validation and initial usability testing, the real test of these technologies lies in their application within an actual long-term care setting [21, 22]. Care professionals interact on a daily basis with the complexities and nuances of long-term care environments, that aren't fully captured in a research environment [24]. Their direct feedback and insights are crucial for refining these AI tools and methodologies to ensure they meet the practical needs, workflow integration, and usability standards required for effective deployment in long-term care settings [25].

## 2.3 Limits of Manual Coding Interviews

This thesis used manual coding as the golden standard to develop automatic coding. The process of manual coding, where human researchers categorize and annotate interview data, plays a crucial role in qualitative research [26]. However, this approach is subject to limitations, particularly regarding consistency and reliability. Studies have highlighted that inter-rater reliability, the degree to which different coders agree on the codes assigned to the same data, often hovers around 70% [27–29]. This variability in manual coding introduces challenges, especially when attempting to utilize this data to train large language learning models. Essentially, if researchers disagree on the classification of data, it becomes challenging for an algorithm to align with both perspectives simultaneously, thus imposing a ceiling on the potential accuracy of any model trained on such data.

To mitigate inconsistencies between the coding results of different researchers, one method is to refine the coding process by ensuring that researchers discuss and reach consensus on the codes [30]. An alternative solution could involve expanding the coding team beyond two individuals and adopting a majority vote system for determining the final codes [31, 32]. This method could enhance the overall accuracy of the coding, which can be useful for training AI models. However, it is important to recognize that increasing the number of coders also escalates the resources required for data annotation, both in terms of time and financial cost.

# 3 Theoretical Considerations

The theoretical contributions of this dissertation lie in its exploration of AI's integration in long-term care. This work can be valuable in long-term care for older adults (LTC), as many tasks in LTC are still performed manually, and because much of the data is currently not being used to make predictions. However, there are also various ethical considerations, as well as considerations regarding the explainability of models.

## 3.1 Quality of Care in Nursing Homes

Humans possess an ability to detect nuanced emotions and themes that is currently better than LLMs [20]. This is essential in contexts like LTC where understanding emotional cues can significantly impact quality of care. This innate capability allows for a deep, empathetic understanding of individuals' needs and experiences, which is particularly crucial in LTC. However, the primary challenge with relying exclusively on human capabilities for these tasks is the issue of scale and time. As the volume and type of data on quality of care indicators in long-term care (e.g., interviews, daily observations, and interactions) increases, the task of manually processing and analyzing this information becomes impractical, if not impossible, due to time constraints and the potential for human error [6].

Automatic speech recognition (ASR) and natural language processing (NLP) are examples of technologies that offer a compelling solution for efficiently storing data and retrieving information from it [33]. By automating the transcription and preliminary analysis of spoken or written communication, these technologies can handle vast amounts of data efficiently, making it feasible to gain insights from large amounts of data that would be unmanageable manually. For example, data from multidisciplinary meetings can be transcribed, summarized and added to EHRs automatically, reducing the administrative burden on care professionals. ASR technology can, for example, transcribe hours of interviews in a fraction of the time it would take a human, while NLP can identify patterns, themes, and sentiments in textual data, providing a basis for further analysis [6, 33]. Although these technologies may not yet match the nuanced understanding of humans fully, they significantly expand the capacity to process and analyze data, enabling a broader and more systematic examination of quality of care indicators. For example, by using AI models on larger volumes of data stemming from various sources (e.g., qualitative and quantitative data from EHRs and data from technological devices such as sensors and wearables), this could lead to more informed decision-making, improved care planning, and ultimately better care outcomes for clients in LTC settings.

## 3.2 Responsible AI and Ethics

While AI can help provide valuable insights into the quality of care, it is important to note that AI can be problematic if used improperly. That is why many businesses and researchers are now advocating for 'responsible AI' [19]. Responsible AI describes the ethical and moral considerations that should be taken into account when developing and deploying AI systems [19]. This term has been discussed frequently since the publication of models such as GPT-3.5 [34, 35]. These models have large, generalized capabilities and can be used for a wide variety of tasks including, for example, answering questions based on provided context, or rendering artistic depictions based on a text description. However, proposed legislation does seem to contain problems. For example, one type of regulation that was proposed, was to force companies that develop AI to have certification (i.e., similar to how ISO certifications work) for their AI models [36, 37]. However, this could increase the difficulty for universities and smaller companies to research and develop models, as certification often involves a payment from the AI model developer to the certification authority. This could increase the threshold for companies – and especially for research groups – to enter the field of AI. This was, for example, a criticism of earlier versions of the EU AI Draft; however, the final version of the EU AI Act has been more lenient in this regard [36, 37].

The deployment of AI in LTC introduces significant challenges concerning the scale of errors. Unlike human errors, which might affect a limited number of individuals, a single mistake by an AI model has the potential to impact thousands due to its scalability [38–41]. For example, the Dutch tax authorities ('Belastingdienst') had many (i.e., more than 30k) incidents where parents who received child care benefits were wrongfully marked as fraudulent (i.e., 'de toeslagenaffaire') [42, 43]. This scalability underscores the need for rigorous safeguards and continuous monitoring to prevent and mitigate the consequences of errors.

There are inherent biases in the use of AI. If a system is too sensitive (i.e., has many false positives), users may become desensitized to that system. For example, in a nursing home setting, a term called 'alarm fatigue' is used to describe the phenomenon where care professionals become desensitized to alarms because they go off too often [44]. If a bed-alarm keeps going off, care professionals might start to ignore the alarm, which could lead to a dangerous situation. On the other hand, there is also the possibility of 'automation bias,' where care professionals might assume the technology is infallible [45]. Automation bias can lead to overreliance on AI decisions without sufficient critical evaluation, potentially overlooking or undervaluing human judgment and expertise [40, 46]. Continuing the example of the 'Belastingdienst,' the parents involved who pointed out that mistakes were made, were repeatedly told that they were wrong and that the system was correct [42, 43]. Addressing these challenges involves not only implementing stringent data protection measures but also educating stakeholders about the limitations of AI, creating a balanced approach that integrates AI insights with human oversight in an LTC setting.

### 3.3 Explainability of AI Models

Most of the AI models discussed in this dissertation predominantly exhibit a 'black-box' nature. This term refers to the opaqueness in how these models process input data and arrive at their conclusions or outputs [47, 48]. The internal workings and decision-making processes of these models are often not easily understandable, even to experts in the fields of data science [47, 48]. This 'blackbox' characteristic poses a challenge in fields where understanding the rationale behind decisions is important, such as in long-term care for older adults [49, 50]. The blackbox nature of these models can impede trust and acceptance among users, particularly when the decisions made by these systems have implications regarding care outcomes [47–50]. For example, if an AI model is responsible for planning the time spent with a client, it is crucial that care professionals understand how the model arrived at its recommendations to ensure that the care provided is appropriate and aligned with the client's needs [49, 50]. For this reason, the area of explainable AI (XAI) has been gaining traction in recent years, aiming to make AI models more transparent and interpretable to users [49–52].

Research shows an inherent trade-off between designing AI models for greater explainability and achieving optimal performance in terms of accuracy and other metrics [51–54]. Increasing the transparency and interpretability of models often requires simplifying their architectures or incorporating additional mechanisms to track and elucidate decision paths [51, 52]. These modifications can potentially compromise the model's complexity and its ability to capture nuanced patterns in data, thereby affecting its overall performance. This trade-off is a consideration in AI development, especially in applications where both high accuracy and understanding of model decisions are crucial [55, 56]. For example, in a nursing home setting, an AI model that recommends care plans for clients must balance the need for accurate predictions with the necessity of providing explanations for these recommendations to ensure that care professionals can trust and act on the model's suggestions.

While AI models, particularly large language models (LLMs), are often criticized for their lack of explainability, it is noteworthy that humans also often struggle to fully articulate their decision-making processes [57–59]. The complexity of human thought and the subconscious influences on our decisions often imply that complete self-explanation is unattainable [60, 61]. Interestingly, recent developments in AI have introduced models capable of a form of self-reflection. These LLMs can generate explanations of their reasoning, shedding light on how they arrive at certain conclusions [62–65]. Although these explanations may not reveal the entire complexity of the underlying processes, they represent a significant step towards bridging the gap between AI decision-making and human interpretability [64, 65].

# 4 Recommendations

## 4.1 Future Research

In the current dissertation, the focus was on specific thematic content analysis (i.e., assigning themes to text segments) and sentiment analysis of interviews. However, large language models (LLMs) can offer greater generalization for both deductive (i.e., where themes are decided beforehand based on pre-existing literature) and inductive (i.e., where themes are derived from the text being analyzed) analysis. Future research could explore the potential of LLMs in analyzing interviews in long-term care (LTC) settings. LLMs can be used to generate summaries of interviews, identify patterns in the data, and even generate new insights based on the content of the interviews [6]. These abilities can be used to provide deeper insights into the quality of care in LTC settings, enabling researchers to uncover hidden insights and trends that may not be immediately apparent through manual coding. Additionally, LLMs can be used to analyze other types of data, such as care plans, medical records, and observational data, to provide a more holistic view of the quality of care in LTC settings. By leveraging LLMs in this way, researchers can gain a deeper understanding of the factors that influence the quality of care and identify areas for improvement.

The automatic speech recognition (ASR) model in this dissertation did not include any way of determining the number of speakers or the current speaker identity. Future research could explore the potential of speaker diarization in ASR models for LTC settings. Speaker diarization is the process of segmenting an audio recording into speaker-specific segments, enabling researchers to identify who is speaking at any given time [66]. By incorporating speaker diarization into ASR models, researchers can gain a more nuanced understanding of the interactions between clients, caregivers, and family members in LTC settings. This approach could provide valuable insights into the dynamics of communication within these environments, enabling researchers to identify patterns in speech and interaction that may influence the quality of care.

### 4.2 Daily Practice in Long-Term Care

It would be beneficial to develop AI assistants that can give nurses advice based on client context. For example, in the case of an infection, the AI assistant could advise on the best course of action, taking into account medical history and potential allergies. This support could significantly reduce nurse workload and decrease the likelihood of errors by helping to prevent them before they occur. Moreover, integrating predictive functionalities in these AI systems could greatly enhance the quality of care. AI could be programmed to anticipate potential incidents such as falls, pressure ulcers, or behavioral issues, allowing healthcare professionals to proactively address these risks. Predictive insights into quality of care outcomes would enable timely interventions, potentially preventing adverse events. In terms of implementation, embedding such systems directly into existing electronic health records (EHRs) may be more feasible than developing standalone voice assistants. This integration could streamline the adoption process, mitigating implementation challenges associated with more complex AI interfaces. This comprehensive approach could improve the quality of life for clients by personalizing and enhancing care activities.

An AI system could assist in filling out medical records. For example, when a nurse is filling in a medical record, the AI system could provide suggestions for the next steps, based on the current context. This would reduce the time nurses spend on administrative work and decrease the number of errors. Additionally, this would aid in providing a more complete medical record, as the AI system could help to fill in gaps in the medical record.

## 5 Conclusion

This dissertation has demonstrated the potential of AI in analyzing data regarding the quality of care for older adults. It has shown that AI can be used to collect and analyze data in nursing homes, thereby providing a more comprehensive understanding of the quality of care. The research presented in this dissertation can serve as a foundation for future studies, offering a framework for AI application in long-term care settings.

## Bibliography


[1] P. Hamet and J. Tremblay, "Artificial intelligence in medicine," *Metabolism*, vol. 69, S36–S40, 2017.

[2] J. Holmes, L. Sacchi, R. Bellazzi, *et al.*, "Artificial intelligence in medicine," *Ann R Coll Surg Engl*, vol. 86, pp. 334–8, 2004.

[3] D. Parekh *et al.*, "A review on autonomous vehicles: Progress, methods and challenges," *Electronics*, vol. 11, no. 14, p. 2162, 2022.

[4] J. Degrave *et al.*, "Magnetic control of tokamak plasmas through deep reinforcement learning," *Nature*, vol. 602, no. 7897, pp. 414–419, 2022.

[5] J. Abramson *et al.*, "Accurate structure prediction of biomolecular interactions with alphafold 3," *Nature*, pp. 1–3, 2024.

[6] W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[7] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[8] C.-Y. Hsieh *et al.*, "Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes," *arXiv preprint arXiv:2305.02301*, 2023.

[9] Y. Fu, H.-C. Peng, L. Ou, A. Sabharwal, and T. Khot, "Specializing smaller language models towards multi-step reasoning," 2023.

[10] H. Touvron *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[11] A. Ruoss *et al.*, "Grandmaster-level chess without search," *arXiv preprint arXiv:2402.04494*, 2024.

[12] H. Touvron *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[13] A. Q. Jiang *et al.*, "Mistral 7b," *arXiv preprint arXiv:2310.06825*, 2023.

[14] E. Harris, "Large language models answer medical questions accurately, but can't match clinicians' knowledge," *JAMA*, 2023.

[15] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[16] S. Tian *et al.*, "Opportunities and challenges for chatgpt and large language models in biomedicine and health," *Briefings in Bioinformatics*, vol. 25, no. 1, bbad493, 2024.

[17] Z. Li *et al.*, "Split and merge: Aligning position biases in large language model based evaluators," *arXiv preprint arXiv:2310.01432*, 2023.

[18] H. Jin *et al.*, "Llm maybe longlm: Self-extend llm context window without tuning," *arXiv preprint arXiv:2401.01325*, 2024.

[19] R. Bommasani *et al.*, "On the opportunities and risks of foundation models," *CoRR*, vol. abs/2108.07258, 2021. arXiv: 2108.07258.

[20] Z. Li *et al.*, "Label supervised llama finetuning," *arXiv preprint arXiv:2310.01208*, 2023.

[21] G. Rolfe, "The theory-practice gap in nursing: From research-based practice to practitioner-based research," *Journal of advanced nursing*, vol. 28, no. 3, pp. 672–679, 1998.

[22] A. Le May, A. Mulhall, and C. Alexander, "Bridging the research–practice gap: Exploring the research cultures of practitioners and managers," *Journal of advanced nursing*, vol. 28, no. 2, pp. 428–437, 1998.

[23] L. A. Bero, R. Grilli, J. M. Grimshaw, E. Harvey, A. D. Oxman, and M. A. Thomson, "Closing the gap between research and practice: An overview of systematic reviews of interventions to promote the implementation of research findings," *Bmj*, vol. 317, no. 7156, pp. 465–468, 1998.

[24] W. Berta *et al.*, "Spanning the know-do gap: Understanding knowledge application and capacity in long-term care homes," *Social science & medicine*, vol. 70, no. 9, pp. 1326–1334, 2010.

[25] M. Chen and M. Decary, "Artificial intelligence in healthcare: An essential guide for health leaders," in *Healthcare management forum*, SAGE Publications Sage CA: Los Angeles, CA, vol. 33, 2020, pp. 10–18.

[26] A. Strauss and J. Corbin, *Basics of qualitative research techniques*. Sage publications Thousand Oaks, CA, 1998.

[27] L. Burla, B. Knierim, J. Barth, K. Liewald, M. Duetz, and T. Abel, "From text to codings: Intercoder reliability assessment in qualitative content analysis," *Nursing research*, vol. 57, no. 2, pp. 113–117, 2008.

[28] M. J. Belotto, "Data analysis methods for qualitative research: Managing the challenges of coding, interrater reliability, and thematic analysis," *The Qualitative Report*, vol. 23, no. 11, pp. 2622–2633, 2018.

[29] J. Belur, L. Tompson, A. Thornton, and M. Simon, "Interrater reliability in systematic review methodology: Exploring variation in coder decision-making," *Sociological methods & research*, vol. 50, no. 2, pp. 837–865, 2021.

[30] C. O'Connor and H. Joffe, "Intercoder reliability in qualitative research: Debates and practical guidelines," *International journal of qualitative methods*, vol. 19, p. 1 609 406 919 899 220, 2020.

[31] G. James, "Majority vote classifiers: Theory and applications," 1998.

[32] L. Lam and S. Suen, "Application of majority voting to pattern recognition: An analysis of its behavior and performance," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 27, no. 5, pp. 553–568, 1997.

[33] Y. Kuratov, A. Bulatov, P. Anokhin, D. Sorokin, A. Sorokin, and M. Burtsev, "In search of needles in a 10m haystack: Recurrent memory finds what llms miss," *arXiv preprint arXiv:2402.10790*, 2024.

[34] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv preprint arXiv:2112.10752*, 2021.

[35] L. Ouyang *et al.*, "Training language models to follow instructions with human feedback," *arXiv preprint arXiv:2203.02155*, 2022.

[36] M. Veale and F. J. Z. Borgesius, "Demystifying the draft eu artificial intelligence act — analysing the good, the bad, and the unclear elements of the proposed approach," *Computer Law Review International*, vol. 22, 2021.

[37] M. Pencina, B. Goldstein, and R. D'Agostino, "Prediction models - development, evaluation, and clinical application.," *The New England journal of medicine*, vol. 382 17, 2020.

[38] D. Char, M. Abràmoff, and C. Feudtner, "Identifying ethical considerations for machine learning healthcare applications," *The American Journal of Bioethics*, vol. 20, 2020.

[39] C.-J. Wang, S. Liu, H. Yang, J. Guo, Y. Wu, and J. Liu, "Ethical considerations of using chatgpt in health care," *Journal of Medical Internet Research*, vol. 25, 2023.

[40] U. Pagallo, "Designing data protection safeguards ethically," *Inf.*, vol. 2, 2011.

[41] I. Barclay and W. Abramson, "Identifying roles, requirements and responsibilities in trustworthy ai systems," *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, 2021.

[42] A. Brenninkmeijer and A. Marseille, "Een dialoog met de raad van state na de toeslagenaffaire," *Nederlands Juristenblad NJB*, vol. 2021, no. 8, pp. 601–608, 2021.

[43] P. Huisman, "Hoe de toeslagenaffaire kon gebeuren," *Management Kinderopvang*, vol. 26, no. 2, pp. 36–37, 2020.

[44] S. Sendelbach, "Alarm fatigue," *Nursing Clinics of North America*, vol. 47, no. 3, pp. 375–382, 2012.

[45] T. Araujo, N. Helberger, S. Kruikemeier, and C. H. De Vreese, “In ai we trust? perceptions about automated decision-making by artificial intelligence,” *AI & society*, vol. 35, pp. 611–623, 2020.

[46] B. Yeager, K. Yeager, and J. Muenzer, “Avoiding the ai minefield: Privacy implications for the use of ai and emerging technologies in healthcare,” *Pediatrics*, 2020.

[47] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should i trust you?”: Explaining the predictions of any classifier,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[48] M. Ananny and K. Crawford, “Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability,” *New Media & Society*, vol. 20, 2018.

[49] M. Sujan *et al.*, “Human factors challenges for the safe use of artificial intelligence in patient care,” *BMJ Health & Care Informatics*, vol. 26, 2019.

[50] P. Srinivasu, N. Sandhya, R. Jhaveri, and R. Raut, “From blackbox to explainable ai in healthcare: Existing tools and case studies,” *Mobile Information Systems*, 2022.

[51] Z. C. Lipton, “The mythos of model interpretability,” *Communications of the ACM*, vol. 61, 2016.

[52] L. H. Gilpin, D. Bau, B. Z. Yuan, A. Bajwa, M. A. Specter, and L. Kagal, “Explaining explanations: An overview of interpretability of machine learning,” *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, 2018.

[53] B. Kim *et al.*, “Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav),” 2017.

[54] D. Bau, B. Zhou, A. Khosla, A. Oliva, and A. Torralba, “Network dissection: Quantifying interpretability of deep visual representations,” *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[55] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, “Explainable ai: A review of machine learning interpretability methods,” *Entropy*, vol. 23, 2020.

[56] A. Datta, S. Sen, and Y. Zick, “Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems,” *2016 IEEE Symposium on Security and Privacy (SP)*, 2016.

[57] P. Das and L. Varshney, “Explaining artificial intelligence generation and creativity: Human interpretability for novel ideas and artifacts,” *IEEE Signal Processing Magazine*, vol. 39, 2022.

[58] M. G. Haselton *et al.*, “Adaptive rationality: An evolutionary perspective on cognitive bias,” *Social Cognition*, vol. 27, no. 5, pp. 733–763, 2009.

[59] M. Hilbert, "Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making.," *Psychological bulletin*, vol. 138, no. 2, p. 211, 2012.

[60] K. S. Bowers, "Knowing more than we can say leads to saying more than we can know: On being implicitly informed," in *Toward a psychology of situations*, Psychology Press, 2013, pp. 189–204.

[61] R. Schank, *Explanation patterns: Understanding mechanically and creatively*. Psychology Press, 2013.

[62] M. Narayanan, E. Chen, J. He, B. Kim, S. Gershman, and F. Doshi-Velez, "How do humans understand explanations from machine learning systems? an evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1802.00682*, 2018.

[63] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. O. Riedl, and Y. Choi, "Reframing human-ai collaboration for generating free-text explanations," 2021.

[64] I. Lage *et al.*, "An evaluation of the human-interpretability of explanation," *arXiv preprint arXiv:1902.00006*, 2019.

[65] H. Zhao *et al.*, "Explainability for large language models: A survey," *ACM Transactions on Intelligent Systems and Technology*, 2023.

[66] Q. Wang, Y. Huang, G. Zhao, E. Clark, W. Xia, and H. Liao, "Diarizationlm: Speaker diarization post-processing with large language models," *arXiv preprint arXiv:2401.03506*, 2024.

# SUMMARY

This dissertation represented a comprehensive exploration of the uses of Artificial Intelligence (AI) in long-term care (LTC), with a particular emphasis on the quality of care for older adults. It was an in-depth study that navigated through the opportunities of data in LTC. The dissertation focused on two main challenges. Firstly, it developed tools and methodologies to support data collection in LTC, automating manual processes. Secondly, it investigated the possibilities of AI in analyzing textual data regarding the quality of care.

**Chapter 1** set the stage for the dissertation by establishing the essential role of resident perspectives in evaluating the quality of care in LTC settings. It highlighted the increasing demand for incorporating these perspectives to enhance the quality of care and overall well-being of residents. The chapter underscored the challenges posed by the vast amounts of data generated from various sources, such as questionnaires, sensors, interviews, and electronic health records, and the need for advanced analytical methods to manage and interpret this data effectively.

**Chapter 2** presented a scoping review aimed at understanding the current application of data science techniques in LTC. The review included 27 studies, highlighting a variety of techniques such as natural language processing (NLP), regression models, support vector machines, random forest models, and convolutional neural networks. These techniques were primarily used to predict adverse outcomes like falls, pressure ulcers, and infectious diseases, and to identify risk factors and improve care practices. Despite the demonstrated effectiveness of these methods, the review found an under-use of data science in LTC settings. The findings indicated a need for more strategic use of data science to exploit its potential fully, emphasizing its value to gaining insights into the quality of care, informing decision-making processes, and addressing challenges in LTC.

**Chapter 3** studied the usefulness of Automatic Speech Recognition (ASR) and the potential for improvement of ASR systems in LTC, especially for groups such as older adults or individuals with distinct accents. Through iterative training with diverse interview data from residents, family members, and care professionals, the WER was significantly reduced, achieving around 80% accuracy. The final ASR model demonstrated substantial improvements in transcription speed, making it a valuable tool for reducing the time and effort required for verbatim transcription, though it still required some manual corrections.

**Chapter 4** focused on the development of an app-based version of the Maastricht Electronic Daily Life Observation tool (MEDLO), using a User-Centered Design (UCD) approach. End-users were involved in the development process to ensure the tool's practicality and effectiveness in LTC settings. The chapter described the iterative design process.

**Chapter 5** explored the application of text-mining techniques to narrative data collected in nursing homes, presenting a novel approach to analyzing qualitative data in long-term care. This study used various text-mining techniques, including word frequency analysis, bigram frequencies, correlation analysis, sentiment analysis, and topic clustering. By applying these methods to interview data from nursing home residents, family members, and care professionals, the study identified patterns and insights related to the quality of care. The findings demonstrated that text-mining could effectively process and analyze large volumes of narrative data, revealing key themes and sentiments that contributed to understanding and improving care in nursing homes.

**Chapter 6** compared text mining and manual coding, specifically focusing on sentiment analysis and thematic content analysis in the context of LTC. The chapter studied a variety of criteria such as accuracy, consistency, and expert feedback. The findings revealed that text mining not only matched but in some cases surpassed manual coding in terms of consistency. The chapter discussed the implications of these findings, suggesting that text mining had the potential to become an integral part of the coding process. However, it also acknowledged the limitations of text mining, noting areas where manual analysis still held value. The chapter concluded by advocating for a balanced approach that leveraged the strengths of both text mining and manual coding to achieve the best outcomes in LTC data analysis.

**Chapter 7** discussed the key findings from the entire dissertation, including methodological and theoretical reflections. It discussed advances being made in AI and their effects on LTC, and it addressed the limited amount of testing with care professionals. The chapter discussed the importance of responsible and explainable AI in LTC. Additionally, it discussed privacy risks of applying AI in a real-world setting and how the energy usage of large AI systems had an influence on the environment. The chapter concluded with the opportunities that automatic speech recognition and applications of NLP could bring to LTC.

In conclusion, this dissertation was a comprehensive and pioneering exploration of the integration of AI in LTC. It provided valuable insights into how AI could be used for data collection and analysis. By presenting innovative approaches and highlighting the need for responsible and ethical AI use, this dissertation contributed significantly to the field of LTC and set the stage for future research and applications in this important area.

# SAMENVATTING

Dit proefschrift is een beschrijving van een uitgebreide verkenning van de toepassing van kunstmatige intelligentie (AI) in de langdurige ouderenzorg, in het bijzonder op de kwaliteit van zorg in de langdurige ouderenzorg. Deze verkenning was een diepgaande studie die de mogelijkheden van het gebruik van data in langdurige ouderenzorg onderzocht. Het proefschrift richtte zich op twee doelstellingen. Ten eerste het ontwikkelen van tools en methodologieën ter ondersteuning van dataverzameling in de langdurige ouderenzorg, waarbij handmatige processen van zorgprofessionals werden geautomatiseerd. Ten tweede de toepassing van AI bij het analyseren van tekstuele data met betrekking tot de kwaliteit van zorg.

**Hoofdstuk 1** schets de context van het proefschrift door de essentiële rol van bewonersperspectieven bij het evalueren van de kwaliteit van zorg in verpleeghuizen vast te stellen. Het benadrukt de toenemende noodzaak voor het integreren van deze perspectieven om de kwaliteit van zorg en het welzijn van bewoners te verbeteren. In dit hoofdstuk worden de uitdagingen benadrukt die werden gevormd door de enorme hoeveelheid data die werden gegenereerd uit verschillende bronnen, zoals vragenlijsten, sensoren, interviews en elektronische patiëntendossiers, en de noodzaak van geavanceerde analysemethoden om deze data effectief te beheren en te interpreteren.

**Hoofdstuk 2** presenteert een verkennende review gericht op het begrijpen van de huidige toepassing van datatechnieken in de langdurige ouderenzorg. De review omvatte 27 studies en belichtte verschillende technieken zoals text mining, NLP, regressiemodellen, support vector machines, random forest-modellen en convolutionele neurale netwerken. Deze technieken werden voornamelijk gebruikt om ziektebeelden en gezondheidsproblemen als vallen, decubitus en infectieziekten te voorspellen, risicofactoren te identificeren en zorgpraktijken te verbeteren. Ondanks dat de artikelen in deze review aantoonden dat de AI methodes effectief zijn, bleek ook dat AI technieken relatief weinig worden ingezet bij het analyseren van data afkomstig uit de langdurige ouderenzorg. De bevindingen wezen op de noodzaak van een meer strategisch gebruik van datatechnieken om hun volledige mogelijkheden te benutten, en benadrukten hun waarde bij het verkrijgen van inzichten in de kwaliteit van zorg, het leveren van informatie voor besluitvormingsprocessen en het aanpakken van uitdagingen in de langdurige ouderenzorg.

**Hoofdstuk 3** worden de bruikbaarheid van automatische spraakherkenning (ASR) en de mogelijkheden voor verbetering van ASR-systemen in de langdurige ouderenzorg onderzocht, met name voor groepen zoals ouderen of mensen met verschillende accenten. Door iteratieve training met diverse interviewdata van bewoners, familieleden en zorgprofessionals werd de WER (woordfoutpercentage) aanzienlijk verminderd, met een nauwkeurigheid van ongeveer 80%. Het uiteindelijke

ASR-model leidde tot aanzienlijke verbeteringen in transcriptiesnelheidaan, waardoor het een waardevol hulpmiddel werd om de tijd en moeite die nodig waren voor letterlijke transcriptie te verminderen, hoewel er nog steeds enige handmatige correcties nodig waren.

**Hoofdstuk 4** richtte zich op de ontwikkeling van een app van de Maastricht Electronic Daily Life Observation Tool (MEDLO), met behulp van een een op de gebruiker gerichte benadering (UCD). Eindgebruikers werden betrokken bij het ontwikkelproces om de bruikbaarheid en effectiviteit van dit hulpmiddel in de langdurige ouderenzorg te waarborgen. Het hoofdstuk beschrijft het iteratieve ontwerpproces. Dit resulteerde in een app met verbeterde bruikbaarheid, die werd gepresenteerd als een casestudy van de adaptatie van technologie in de langdurige ouderenzorg.

**Hoofdstuk 5** verkent de toepassing van text mining-technieken op narratieve data die in verpleeghuizen zijnverzameld, en presenteert een nieuwe benadering voor het analyseren van kwalitatieve data in de langdurige zorg. In deze studie zijn verschillende text mining-technieken gebruikt, waaronder woordfrequentieanalyse, bigram-frequenties, correlatieanalyse, sentimentanalyse en topic clustering. Door deze methoden toe te passen op interviewdata van verpleeghuisbewoners, familieleden en zorgprofessionals, zijn patronen en inzichten met betrekking tot de kwaliteit van zorg geïdentificeerd. De bevindingen toonden aan dat text mining een effectief middel is om grote hoeveelheden narratieve data te verwerken en analyseren. Hierdoor werden belangrijke thema's en sentimenten blootgelegd, die bijdroegen aan een beter begrip en de verbetering van de zorg in verpleeghuizen.

In **Hoofdstuk 6** worden text mining en handmatige codering vergeleken, met een specifieke nadruk op sentimentanalyse en thematische inhoudsanalyse in de langdurige ouderenzorg. Het hoofdstuk bestudeerteen verscheidenheid aan criteria, zoals nauwkeurigheid, consistentie en feedback van deskundigen. Uit deze vergelijking bleek dat text mining qua consistentie niet alleen overeenkwam met handmatige codering, maar het in sommige gevallen zelfs overtrof. Het hoofdstuk bespreekt de implicaties van deze bevindingen en toont aan dat text mining het potentieel heeft om een integraal onderdeel van het coderingsproces te worden. Het erkent echter ook de beperkingen van text mining en wijst op situaties waar handmatige analyse nog steeds van waarde is. In de conclusie van dit hoofdstuk wordt gepleit voor een evenwichtige benadering, waarbij gebruik wordt gemaakt van de sterke punten van zowel text mining als handmatige codering, om zodoende de beste resultaten uit gegevens van de langdurige ouderenzorg te behalen.

**Hoofdstuk 7** worden de belangrijkste bevindingen van het gehele proefschrift gepresenteerd, waarbij zowel methodologische als theoretische reflecties worden besproken. Het gaat in op de huidige vooruitgang in AI en de gevolgen hiervan voor de langdurige ouderenzorg Ook wordt er aandacht besteed aan de beperkte hoeveelheid testen met zorgprofessionals en het belang van verantwoord en verklaarbaar AI-gebruik in de langdurige ouderenzorg. Bovendien worden de privacyrisico's van het toepassen van AI in een real-world setting benoemd en is er aandacht voor de invloed van het energieverbruik van grote AI-systemen op het milieu. Het hoofdstuk sloot af met de kansen die automatische spraakherkenning en toepassingen van NLP de langdurige ouderenzorg kunnen bieden.

Tot slot biedt dit proefschrift een eerste verkenning van het gebruik van AI-technieken in het analyseren van data die in de langdurige ouderenzorg is verzameld. Het biedt hierbij waardevolle inzichten in hoe AI kan worden gebruikt voor dataverzameling en -analyse. Door het gebruik van vernieuwende benaderingen en door het belang van verantwoord en ethisch AI-gebruik te benadrukken, draagt dit proefschrift bij aan het gebied van de langdurige ouderenzorg en heeft het de weg vrijgemaakt voor toekomstig onderzoek en toepassingen op dit belangrijke gebied.

# IMPACT

In 2020, the 'Limburg Living Lab in Aging and Long-Term Care' recognized the importance of using larger datasets and researching innovative data analysis methods. The 'Core Group Data science' was established to collaborate with care organizations on this topic. This working group consists of data analysts, innovation mangers and business intelligence experts from care organizations in Limburg (MeanderGroep Zuid-Limburg, Sevagram, Envida, Cicero Zorggroep, Zuyderland, Vivantes, De Zorggroep, Land van Horne & Proteion) as well as researchers from Zuyd Hogeschool and Maastricht University. This marks a significant milestone in advancing research. This dissertation, one of the first major projects to emerge from this initiative, focuses on the data analysis of developed measurement tools within the Living Lab.

In this dissertation, we show that the use of AI and data science in nursing homes can lead to significant scientific advancements and advancements in the way that LTC is conducted. The development of tools like the MEDLO app and the Transcription Tool represents major progress in data collection and analysis, offering faster and user-friendly methods for collecting nursing home data. This enhances the quantity of research that can be conducted in nursing homes, because it enables more people collected data in a structured manner. The dissertation has contributed in various ways, including the development of automated tools for data collection and analysis, the enhancement of research methodologies, including automatic speech recognition (ASR) and natural language processing (NLP) approaches.

## Societal Impact

This dissertation has had many developments that could aid in quality of care analysis. One key development is the transcription program based on a refined automatic speech recognition (ASR) model. This tool enables researchers to easily convert collected data from elderly care into text for subsequent analysis. By automating the transcription process, the ASR model significantly reduces the time and effort required for manual data conversion, thereby allowing researchers to focus on more in-depth analysis and improving overall data accuracy. This tool is particularly valuable for converting interviews and observations into analyzable text, facilitating a more comprehensive understanding of the experiences of older adults in care. Furthermore, this thesis has contributed to the use of natural language processing (NLP), advancing research regarding the automation of quality of care analysis. We have shown various ways in which NLP can be used to analyze data, including the development models that can automatically code data. We have shown that these models are capable of analyzing with accuracies of above 80%. These models can be used in applications to streamline the coding process, making it more efficient and less prone to human error. Both the ASR model and the NLP model developed in this dissertation are currently being used to further advance the Connecting Conversations method. These offer an advantage compared to general models (e.g., Llama 3, GPT 4), because our models were trained on text and speech data from a LTC setting. A specific module was developed that automatically translates speech into text of older adults, their family caregivers, and staff in nursing homes regarding their perspective on quality of care. In addition, by using the developed NLP techniques, the collected data can be automatically analyzed. This empowers frontline staff and management in nursing homes to gain better insight in the quality of care themselves and use narrative data to improve care processes. Connecting Conversations is now implemented on a national level.

The creation of the MEDLO app, based on the existing Maastricht Electronic Daily Life Observation (MEDLO) method, has greatly simplified data collection in the field. This app is designed to be user-friendly, making it easier for researchers to gather data on the quality of life of persons with dementia. By transforming the traditionally cumbersome Excel-based system into a streamlined digital application, the MEDLO app enhances the efficiency of data collection and analysis. This improvement not only aids researchers but also ensures that the data collected is more accurate and representative of the real-life conditions of residents. Furthermore, it enhances the use of the app by nursing staff themselves, also employing it as a quality improvement strategy or tool. In addition, the research demonstrates the capability of automated coding to a significant extent. This advancement allows researchers to analyze data more quickly and extensively. Automated coding reduces the time-consuming nature of manual data analysis, enabling the examination of larger datasets and providing deeper insights into the quality of care in LTC settings. This capability enhances the scope of research, allowing for more robust and comprehensive studies that can drive meaningful improvements in care practices.

## Scientific Impact

This dissertation has demonstrated the significant potential of artificial intelligence (AI) and data science in improving long-term care (LTC) for older adults. The research showcases how advanced technologies can be integrated into LTC settings to enhance data collection, analysis, and ultimately, the quality of care provided. One of the major contributions of this work is the development and validation of AI models that can automatically code qualitative data related to the quality of care. These models have achieved high accuracy rates, significantly reducing the reliance on manual coding, which is often time-consuming and subject to human error. This advancement enables researchers to handle larger datasets more efficiently, allowing for more comprehensive studies that can uncover deeper insights into care practices. Additionally, the dissertation highlights the importance of not solely relying on computer-generated analyses. While AI tools offer significant advantages, the research emphasizes the need for a balanced approach where human oversight remains crucial. This ensures that the data and insights derived are both accurate and contextually relevant. Another notable scientific contribution is the detailed process for converting interview recordings into analyzable data. By employing automatic speech recognition (ASR) technology to transcribe audio data and then utilize natural language processing (NLP) techniques for analysis, this process streamlines data collection while providing a rich source of information. This methodology not only enhances the efficiency of research but also enriches the understanding of care practices and outcomes in LTC settings. The dissertation also maps out the possibilities in the field of AI and LTC, with a significant portion of the research originating from this work. This comprehensive exploration of AI applications in LTC provides a valuable framework for future research, guiding the development of new tools and methodologies that can further improve the quality of care for older adults.

## Dissemination of Findings

The findings of this dissertation have been widely disseminated through international conferences and reputable journal publications, ensuring broad visibility and impact. Presentations at conferences such as the Nordic Congress of Gerontology 2022, NVG-KNOWS 'Nationaal Gerontologiecongres' 2022, SANO 'Wetenschapsdag' 2023, and GSA 2023 were conducted for groups including scientists and care professionals, innovation managers, and data/ICT specialists working in long-term care. These conferences provided platforms to showcase the practical applications of tools like the automatic speech recognition (ASR) model and the Maastricht Electronic Daily Life Observation (MEDLO) app, highlighting their potential to streamline data collection and enhance the quality of care in nursing homes. Feedback from these events has been invaluable in refining methodologies and addressing implementation challenges.

In addition to conference presentations, several high-impact journal publications have further disseminated the research findings. Four of the chapters were published in high-impact peer-reviewed journals, including **Chapter 2**, **Chapter 3**, **Chapter 5**, and **Chapter 6**. These publications demonstrate the growing interest in innovative AI applications in LTC regarding quality of care.

These dissemination facilitated a more receptive environment for adopting innovative AI technologies in LTC-based research and in LTC in general. By engaging with the broader scientific and professional communities through these channels, the dissertation has contributed to advancing the understanding and application of AI in improving the quality of care for older adults, laying the groundwork for future innovations and practical applications in this critical field.

# SOURCE CODE

Open science, as demonstrated throughout the dissertation, exemplifies the integration of accessible, comprehensive knowledge with practical application. This dissertation stands out as they are accompanied by code, offering a level of detail and practical understanding that often exceeds the constraints of traditional academic papers. Academic papers typically face word limits, which can restrict the depth of explanation and detail they can provide. In contrast, the accompanying code serves as a more expansive and interactive form of knowledge dissemination. This dissertation is meticulously crafted to resonate with a broader, non-technical audience, ensuring that the insights and methodologies they contain are not just reserved for experts but are accessible and comprehensible to a wider public. This approach not only enhances the transparency and applicability of scientific research but also bridges the gap between complex scientific concepts and everyday understanding. Links to the code repositories can be found in Figure 7.1.

**(a)** The code corresponding with **Chapter 3**.

**(b)** The code corresponding with **Chapter 5**.

**(c)** The code corresponding with **Chapter 6**.

**Figure 7.1:** QR Codes and URLs for GitHub Repositories

# DANKWOORD

Het moment is eindelijk daar: het boekje is af. Ik heb hier lang naar uitgekeken, maar nu het zover is, voelt het toch een beetje gek. De afgelopen jaren heb ik veel geleerd en heb ik veel leuke, interessante en bijzondere mensen mogen leren kennen en ontmoeten. Ik heb veel mensen om mij heen gehad die mij hebben geholpen, en ik wil graag iedereen bedanken die hieraan heeft bijgedragen.

Dit onderzoek was niet mogelijk geweest zonder de Academische Werkplaats Ouderenzorg Limburg (AWO-L). Ik wil daarom graag iedereen bedanken die betrokken is bij de AWO-L. Helaas begon mijn onderzoek net toen we vanwege COVID-19 allemaal thuis moesten werken, waardoor ik veel van jullie pas later in het echt heb ontmoet. Dankzij het harde werk van de dataverzameling, die al was gedaan voordat ik begon, kon ik toch meteen aan de slag met het analyseren van de data.

Beste team, ik wil jullie graag bedanken voor de fijne samenwerking. Ik heb veel van jullie geleerd en ik hoop dat jullie ook wat van mij hebben opgestoken (alhoewel ik weet dat ik het jullie soms ook erg lastig heb gemaakt met alle technische termen).

Jan, jij hebt mij altijd gesteund in mijn onderzoek en me de vrijheid gegeven om mijn eigen koers te varen. Mijn resultaten en programma's waren soms wel erg technisch; misschien dat je me daarom ook een "tovenaar" bent gaan noemen. We hebben soms erg ons best moeten doen om elkaar te begrijpen, maar ik denk dat we daar goed in zijn geslaagd, en het resultaat mag er zijn. Nu het proefschrift af is en ik dit dankwoord schrijf, ben ik er nog steeds niet achter wat sommige van de hiërogliefen in je handgeschreven feedback betekenen. Ik wens je veel succes, en ik hoop dat je nog veel mooie dingen mag doen.

Hilde, het is bijna moeilijk om te geloven dat er dingen zijn die jij niet leuk vindt, want alles wat ik liet zien van ons onderzoek werd door jou begroet met: "Wat leuk!". Het is een deugd om jou te zien stralen als je over je onderzoek praat (vooral op gouden hakken), en ik hoop dat ik later in mijn carrière net zo vol passie over mijn werk kan vertellen. Ik ben je dankbaar voor je enthousiasme en je inzet voor ons onderzoek. Ik wens je dan ook nog veel succes met je de rest van je onderzoek.

Sil, jouw toewijding en enthousiasme zijn een grote bron van inspiratie voor me geweest. Jij weet als geen ander hoe je de wetenschappelijke lat hoog kunt leggen. Ik ben je dankbaar voor je grondige feedback en het meedenken over details. Ik heb ontzettend veel geleerd van jouw kennis en ben blij met de fijne samenwerking die we hebben opgebouwd. Dank je wel voor alle tijd en moeite die je in mijn onderzoek hebt geïnvesteerd.

Katya, Audrey en Erica, jullie hulp was onmisbaar tijdens mijn onderzoek. Katya, jouw precisie en vastberadenheid om alles goed te organiseren hebben de dataverzameling enorm soepel laten verlopen. Audrey, jouw geduld en oog voor detail zorgden ervoor dat het testen van mijn programma's zonder problemen verliep. Erica, jouw scherpe inzichten en vermogen om met creatieve oplossingen te komen hielpen bij het verbeteren van de efficiëntie van mijn programma's. Ik wil jullie alle drie hartelijk bedanken voor de fijne samenwerking en toewijding.

Sander, jij was een geweldige sparringpartner op het gebied van AI. Onze gesprekken over de nieuwste LLM's, de vooruitgang in machine learning en de richting waarin de technologie zich beweegt, hebben mij enorm geïnspireerd. Je stond altijd open voor het delen van nieuwe ideeën en je hebt me geholpen om steeds kritisch naar mijn eigen werk te blijven kijken. Bedankt voor alle stimulerende gesprekken en inzichten die je met mij hebt gedeeld.

Beste collega's van de AWO-L, dankzij jullie heb ik de universi-tijd van mijn leven gehad. Ik weet dat jullie het zullen missen, dus ik vertrouw erop dat jullie de woordgrappen in dit stuk kunnen waarderen. We hebben vaak samen lekker geluncht, geborreld en koffie gedronken. Die momenten hebben niet alleen bijgedragen aan een fijne werksfeer, maar ook aan een hechte band binnen de vakgroep. De sfeer was altijd warm en open, ondanks het gebrek aan een verwarming in 'Dub30'. Jullie betrokkenheid en steun hebben mijn onderzoekservaring extra bijzonder gemaakt. Heel erg bedankt voor alle mooie herinneringen en gezelligheid!

## 6 Vrienden

Yannick, ik heb je leren kennen bij de dansvereniging en ik had toen al veel bewondering voor het feit dat je voor elke dag van het jaar een ander pak lijkt te hebben. Al ben ik ook wel trots dat ik, als een van de weinigen, je een keer in een T-shirt heb mogen zien. Er is veel bier doorheen gegaan in onze tijd samen, ook al drinken we allebei tegenwoordig wat minder. Als een van ons een keer hulp nodig had, was een telefoontje met "Hey, heb je zin in KFC?" genoeg om de ander te overtuigen om te komen helpen. Bovendien heb ik dankzij jou geleerd hoe waardevol en duurzaam avocadopitten kunnen zijn. Ik ben blij dat ik je heb leren kennen, en ik hoop dat we nog veel biertjes zullen drinken in de toekomst.

Larissa, je bent een dappere vrouw. Terwijl ik af en toe gefrustreerd was met mijn PhD en twijfelde of ik nog wel wilde doorgaan met mijn onderzoek, heb ik jou toch weten te overtuigen om zelf een PhD te gaan doen en in Maastricht te blijven. Ik weet nog dat ik tegen je zei: "Do it, but don't do it." Ik ben blij dat je toch deze stap hebt gezet en dat je het ook leuk vindt. Hoewel je het tegenwoordig wel heel druk hebt met je PhD, hoop ik dat er genoeg zonnige dagen zullen zijn waarop we samen een "Paso Doble" kunnen dansen in het park. Nog veel succes met je PhD.

Joris, wij kennen elkaar natuurlijk ook al een hele tijd via de dansvereniging. We hebben samen in het bestuur gezeten, waar we allebei wat meer feestjes hadden willen geven. Daarna ben je met je hotel begonnen, waar ik je graag bij geholpen heb. Het is jammer dat ”Fris.” er niet meer is, maar ik ben blij dat we nog steeds contact hebben, zelfs nu je in het buitenland woont samen met Berenice. Ik hoop dat jullie snel weer naar Nederland komen, want ik kijk erg uit naar jullie bruiloft.

Marc, ik ken jou natuurlijk als de broer van Yannick, maar de laatste tijd zijn we steeds vaker samen wat gaan drinken. Die avonden eindigen vaak bij de frituur, waar je niet vies van bent. Ondanks je passie voor Grieks en je verfijnde smaak voor exotisch eten, ben je altijd in voor wat vettigs—zeker als ik je een keer help met statistiek. Je bent ontzettend vrijgevig, en ik voel me regelmatig schuldig als jij weer de rondjes betaalt. En dan is er nog je stijl: hoewel Yannick meestal in pak verschijnt, ben jij toch de meest stijlvolle van de twee met je kraanvogel-smoking. Jammer dat je misschien weggaat uit Maastricht, maar ik wens je wel nog heel veel geluk.

Marco, ”Donde está la biblioteca?” Natuurlijk heb jij de universiteitsbibliotheek gevonden, want je bent ook bijna klaar met je PhD. Dat heb je zeker niet alleen gedaan: er was veel salsa, tequila en spicy food om je te helpen. Wanneer ik bij ”With Love Burrito” bestel, vind ik de sour cream al pittig genoeg, maar jij lijkt geen enkele hoeveelheid pepers op te merken. Dat doet me wel een beetje twijfelen aan je geloofwaardigheid als voedingswetenschapper. Tussen een Mexican en een Mexi-can't, ben jij zeker een Mexi-can. Ik hoop dat je snel je PhD afrondt en dat je blijft salsa dansen met ons.

Celine, je bent er natuurlijk ook bijna, net als Marco. Misschien twijfel je daar af en toe aan, maar weet dat wij er allemaal zijn om je te steunen. Ik denk niet dat ik iemand ken die zo ontzettend lief is als jij. Vandaar dat we je ook af en toe zeggen dat je een beetje zelfzuchtiger mag zijn. Je hebt wel bewezen dat je dit kan, bijvoorbeeld toen we een keer Thais gingen eten en je ter plekke besloot om een gong te bestellen. Het resultaat mag er zijn, want hij klinkt fantastisch, maar niet zo fantastisch als jouw Franse hoorn. Dat je muzikaal bent, is natuurlijk ook niet zo gek, want ook met salsa dansen ga je helemaal los. Ik wens je nog heel veel succes met je PhD en hoop dat je nog veel mooie dingen mag doen.

## 7 Familie

Oma, bedank veur alle leefde en zörg ummer. Ich höb vaak bie uch en opa maoge logere. Dao door weit ich waal ouch dat ich mot oetkieke eine sopteier, maar dat is gelökkig allemaol good gekomme. Mit kaesjmes maak geer nog ummer lekkere pasteitjes en de knien in ’t zoer is nog ummer get wo ich nao oetkiek. Geer zeet ummer erg greuëtsj op mich gewaes en ich bèn uch hiël dankbaar veur wie geer mich ummer gesteund höb. Een dikke knoevel van mich.

Opa, jaomer genóg koos geer neet mièr bie de promotie zin, maar geer hub nao 3,5 lestige jaore in de rolsjtool toch uch rös gevònje. Ich weit nog dat veer ummer same appele gingen haole bie Mia en dat ich dan ouch op de weagsjaol maogde staon. Van tummere koos geer waal get, maar geer koos ouch genóg klómmele, want ich höb uch toch waal vaak gezeen met akelige roäi pleisters óm de vingere. Boete dat, koos geer ouch good klómmele mit de computer, wo ich ouch vaak genóg rommel van höb opgeruimd. ’T waor ummer leuk om in de zomer door de tuin te renne en te zeen wie geer aan ’t böttele waort. Geer zeet ummer hiel trots op mich gewaes mit mien studie en PhD. Ich mis uch.

Mama, gehr steunt mich altied in alles wat ich doon. Geer höb mich erg gesteund in ’t vinge van ein hoes, en geer hulp mich ouch mit poetsen. Ich huër uch dan waal knotere, omdat ich te väöl rotzooi höb. Geer zeet ummer thoes gewaes om veur Roel en mich te zörge. Och deeg geer os euveral brenge en hoale om dao nao spek mit ei, bótramme en een beaker chocomel te maken. Geer zeet ouch ein stök henjiger dan papa. Vreuger höb geer mich ouch väöl gehollepe mit hoeswerk. Ich bèn uch erg dankbaar veur alles wat geer veur mich doot. Ich haop dat ich uch trots maak mit mien PhD.

Papa, geer waord en zeet nog altied ein groat voorbeeld veur mich. Uch hulp mit mien hoeswerk en uche raod höbbe mich altied gesjtimuleerd om door te zètte. Netuurlik haet Socrates mich waal ein bietje op mien paad gezat mit de IT,maar ouch Game Maker haet mich väöl gelièrd. Van DOOM en Heretic höb ich dankzij uch ouch genóg mit gekrege. En nao dat geer al ein aantal ultra-trails hub gerend hub ouch ich eindelijk eine marathon gelaupe. Ich bèn uch erg dankbaar veur alles wat geer veur mich doot.

Roel, dich bès mit dien PhD in optics bezig. ’T is netuurlik waal een ganse rit om eedere daag nao Eindhoven te gaon, maar de deis ’t toch ondanks dat de treine neet ummer op tied rieje. Boete dien PhD bès dich netuurlik ouch de Pi- ano Man van de femilie, en ’t is leuk dat veer dich af en toe ouch maoge huëre zonger dat ze diene koptelefoon op höbs. De bès netuurlik ouch de langste en sjlankste van os twiè. Dan bèn ich toch maar de kleine dieke. Ich ving ’t leuk dats doe sjtraks bie mien promotie bes en ich verheug mich op dae van dich.

# ABOUT THE AUTHOR

Coen Hacking, born on 22 October 1995, is an accomplished Dutch software developer renowned for his expansive skill set and a deep-rooted passion for advancing technology frontiers. His journey in the tech realm began with video game programming, a foundation that paved the way for his later ventures in application development. His academic pursuits introduced him to new challenges and opportunities, particularly in the field of artificial intelligence. Coen engaged in cutting-edge research projects, These projects are testament to his commitment to groundbreaking research and innovation. More recently, Coen has channeled his expertise into the healthcare technology sector. A notable achievement in this field is his development of a transcription program that utilizes advanced AI models for Dutch speech. This innovation has made significant contributions to research in the area of long-term care for older adults. Moreover, he has been involved in researching the user experience of software in long-term care settings. This work highlights his dedication to creating secure, user-centric software solutions that address real-world challenges. Despite his involvement in a diverse range of impactful projects, Coen remains committed to exploring the vast landscape of software engineering. He has continued to focus on automation, though the area of has shifted toward the creation of software tools to create aeronautical charts, which are vital in the safety of air travel.

# SCIENTIFIC PUBLICATIONS

## International Publications

Hacking, C., Verbeek, H., Hamers, J. P., Sion, K., & Aarts, S. (2022). Text mining in long-term care: Exploring the usefulness of artificial intelligence in a nursing home setting. PLOS ONE, 17(8), e0268281.

https://doi.org/10.1371/journal.pone.0268281

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2022). The development of an automatic speech recognition model using interview data from long-term care for older adults. Journal of the American Medical Informatics Association, 30(3), 411-417.

https://doi.org/10.1093/jamia/ocac241

Golz, C., Aarts, S., Hacking, C., Hahn, S., & Zwakhalen, S. M. G. (2022). Health professionals' sentiments towards implemented information technologies in psychiatric hospitals: a text-mining analysis. BMC Health Services Research, 22(1), 1426.

https://doi.org/10.1186/s12913-022-08823-4

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2023). Comparing text mining and manual coding methods: Analysing interview data on quality of care in long-term care for older adults. PLOS ONE, 17(8), e0268281.

https://doi.org/10.1371/journal.pone.0292578

Hendriks, A., Hacking, C., Verbeek, H., & Aarts, S. (2024). Data science techniques to gain novel insights into quality of care: A scoping review of long-term care for older adults. Explorations in Digital Health Technology, 2(1), 67–85.

https://doi.org/10.37349/edht.2024.00012

## Conference Contributions

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2022). HuBERTien: automatic listening to interview recordings in long-term.care. Nordic Congress of Gerontology, Odense, Denmark.

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2022). Een goudmijn aan tekst: Een vergelijking tussen text mining en manuele analyses in de langdurige ouderenzorg. NVG-KNOWS, Den Bosch, The Netherlands.

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2023). A goldmine of text: comparing text mining and manual coding methods in long-term care for older adults regarding quality of care. GSA, Tampa, Florida, USA.

Hacking, C., Verbeek, H., Hamers, J. P., & Aarts, S. (2023). Een goudmijn aan tekst: Een vergelijking tussen text mining en manuele analyses in de langdurige ouderenzorg. SANO, Amsterdam, The Netherlands.

# LIMBURG LIVING LAB IN AGING AND LONG-TERM CARE

This thesis is part of the Limburg Living Lab in Ageing and Long-Term Care, a formal and structural multidisciplinary network consisting of Maastricht University, nine long-term care organizations (MeanderGroep Zuid-Limburg, Sevagram, Envida, Cicero Zorggroep, Zuyderland, Vivantes, De Zorggroep, Land van Horne & Proteion), Intermediate Vocational Training Institutes Gilde and VISTA college and Zuyd University of Applied Sciences, all located in the southern part of the Netherlands. In the Living Lab we aim to improve quality of care and life for older people and quality of work for staff employed in long-term care via a structural multidisciplinary collaboration between research, policy, education and practice. Practitioners (such as nurses, physicians, psychologists, physio- and occupational therapists), work together with managers, researchers, students, teachers and older people themselves to develop and test innovations in long-term care.

## Academische Werkplaats Ouderenzorg Limburg

Dit proefschrift is onderdeel van de Academische Werkplaats Ouderenzorg Limburg, een structureel, multidisciplinair samenwerkingsverband tussen de Universiteit Maastricht, negen zorgorganisaties (MeanderGroep Zuid-Limburg, Sevagram, Envida, Cicero Zorggroep, Zuyderland, Vivantes, De Zorggroep, Land van Horne & Proteion), Gilde Zorgcollege, VISTA college en Zuyd Hogeschool. In de werkplaats draait het om het verbeteren van de kwaliteit van leven en zorg voor ouderen en de kwaliteit van werk voor iedereen die in de ouderenzorg werkt. Zorgverleners (zoals verpleegkundigen, verzorgenden, artsen, psychologen, fysio- en ergotherapeuten), beleidsmakers, onderzoekers, studenten en ouderen zelf wisselen kennis en ervaring uit. Daarnaast evalueren we vernieuwingen in de dagelijkse zorg. Praktijk, beleid, onderzoek en onderwijs gaan hierbij hand in hand.

## PhD-Theses Limburg Living Lab in Ageing and Long-Term Care / Proefschriften Academische Werkplaats Ouderenzorg Limburg

Coen Hacking. Text mining in long-term care: Exploring the usefulness of artificial intelligence in a nursing home setting. 2025.

Merel van Lierop. Keep on learning! Fostering continuous learning and improvement in long-term nursing care. 2025.

Niklaus Stefan Bernet. Fair and Meaningful Comparison of Inpatient Fall Rates in Acute Care Hospitals. 2025.

Johanna Rutten. Are we walking or just talking? Enhancing relationship-centered care in nursing homes. 2024.

Ines Mouchaers. Managing Everyday Life. Exploring the Essential Components of Reablement and User Experiences. 2024.

Lindsay Groenvynck. The transition from home to a nursing home: The perspectives and experiences of older people with dementia, informal caregivers and professional caregivers. 2024.

Svenja Cremer. Undervalued & Unexplored. Underpinning and Guiding Nursing Care in Activities of Daily Living. 2024.

Katinka Pani – Harreman. The place to be. Guiding the activation of a community to facilitate ageing in place. 2024.

Vincent Moermans. Struggling with Daily Care Dilemmas. Insights in Involuntary Treatment Use among Persons Living with Dementia Receiving Professional Home Care. 2023

Chandni Khemai. There is an I in WE. Collaborative Awareness in the dedicated palliative care approach for persons with dementia. 2023.

Amal Fakha. Improving the implementation of transitional care innovations. 2023.

Tom Vluggen. Towards improved multidisciplinary stroke care for older people: assessing feasibility and effectiveness of an integrated multidisciplinary geriatric rehabilitation programme for older stroke patients. 2023.

Priscilla Attafuah. Quality of life, health, and social needs of slum-dwelling older adults in Ghana. 2023.

Ron Warnier. Frailty screening in older hospitalized patients. 2023.

Megan Davies. Tri-national ethnographic multi-case study of person-centred care and quality of life in long-term residential care. 2023.

Christoph Golz. Technostress among health professionals: The blame game between health professionals and technology. 2023.

Simone Paulis. Dehydration in the nursing home. Research into the prevalence, risk factors, diagnosis, roles and current collaboration between (in)formal caregivers and nursing home residents. 2023.

Sheizi Sari. Pressure injuries in Indonesian community-dwelling older adults: prevalence, prevention and treatment by the (in)formal support system. 2022.

Teuni Rooijackers. Supporting older adults to STAY ACTIVE AT HOME. Process, effect and economic evaluation of a reablement training program for homecare staff. 2022

Anne van den Bulck. Differences that matter: Understanding case-mix and quality for prospective payment of home care. 2022.

Marlot Kruisbrink. Towards enhanced management of fear of falling in older people. Unravelling interventions and measuring related avoidance of activity. 2022.

Ruth Vogel. Nurses in the Lead: empowering community nurse leaders to implement evidence into practice. 2022.

Fabian Groven. The bed bath with or without water? It's a wash! Experiences with the washing without water intervention used for the bed bath. 2021.

Roy Haex. Take a look through my eyes: The development of an experienced quality measure with clients, informal, and formal caregivers in Dutch home care. 2021.

Sascha Bolt. The fundamentals of a DEDICATED palliative approach to care for people with dementia. 2021.

Angela Mengelers. To risk or to restrain? Involuntary treatment use in people with dementia living at home. 2021.

Katya Sion. Connecting Conversations. Experienced quality of care from the resident's perspective: a narrative method for nursing homes. 2021.

Linda Hoek. Change begins with choice. Supporting the autonomy of nursing home residents with dementia through partnership. 2020

Mirre den Ouden. Every step counts. Daily activities of nursing home residents and the role of nursing staff. 2018.

Theresa Thoma-Lürken. Innovating long-term care for older people. Development and evaluation of a decision support app for formal caregivers in community-based dementia care. 2018.

Eveline van Velthuijsen. Delirium in older hospitalised patients: diagnosis and management in daily practice. 2018.

Bram de Boer. Living at a green care farm. An innovative alternative for regular care in nursing homes for people with dementia. 2017.

Nienke Kuk. Moving forward in nursing home practice. Supporting nursing staff in implementing innovations. 2017.

Irma Everink. Geriatric rehabilitation. Development, implementation and evaluation of an integrated care pathway for older patients with complex health problems. 2017.

Ramona Backhaus. Thinking beyond numbers. Nursing staff and quality of care in nursing homes. 2017.

Martin Van Leen. Prevention of pressure ulcers in nursing homes, a big challenge. 2017.

Mariëlle Daamen-Van der Velden. Heart failure in nursing home residents. Prevalence, diagnosis and treatment. 2016.

Armand Rondas. Prevalence and assessment of (infected) chronic wounds. 2016.

Hanneke Beerens. Adding life to years. Quality of life of people with dementia receiving long-term care. 2016 (Cum Laude).

Donja Mijnarends. Sarcopenia: a rising geriatric giant. Health and economic outcomes of community-dwelling older adults with sarcopenia. 2016.

Tanja Dorresteijn. A home-based program to manage concerns about falls. Feasibility, effects and costs of a cognitive behavioral approach in community-dwelling, frail older people. 2016.

Basema Afram. From home towards the nursing home in dementia. Informal caregivers' perspectives on why admission happens and what they need. 2015.

Noemi Van Nie-Visser. Malnutrition in nursing home residents in the Netherlands, Germany and Austria. Exploring and comparing influencing factors. 2014.

Silke Metzelthin. An interdisciplinary primary care approach for frail older people. Feasability, effects and costs. 2014.

Jill Bindels. Caring for community-dwelling frail older people: a responsive evaluation. 2014.

Esther Meesterberends. Pressure ulcer care in the Netherlands versus Germany 0-1. What makes the difference? 2013.

Math Gulpers. EXBELT: expelling belt restraints from psychogeriatric nursing homes. 2013.

Hilde Verbeek. Redesigning dementia care. An evaluation of small-scale homelike care environments. 2011.

Judith Meijers. Awareness of malnutrition in health care, the Dutch perspective. 2009.

Ans Bouman. A home visiting program for older people with poor health. 2009.

Monique Du Moulin. Urinary incontinence in primary care, diagnosis and interventions. 2008.

Anna Huizing. Towards restraint free care for psychogeriatric nursing home residents. 2008.

Pascalle Van Bilsen. Care for the elderly, an exploration of perceived needs, demands and service use. 2008.

Rixt Zijlstra. Managing concerns about falls. Fear of falling and avoidance of activity in older people. 2007.

Sandra Zwakhalen. Pain assessment in nursing home residents with dementia. 2007.