MACHINE LEARNING APPLICATIONS IN HEALTHCARE

Ana Luiza Dallora Moraes

Blekinge Institute of Technology
Doctoral Dissertation Series No. 2020:06

Department of Health

Machine Learning Applications in Healthcare

Ana Luiza Dallora Moraes

Blekinge Institute of Technology Doctoral Dissertation Series
No 2020:06

Machine Learning Applications in Healthcare

Ana Luiza Dallora Moraes

Doctoral Dissertation in
Applied Health Technology

Department of Health
Blekinge Institute of Technology
SWEDEN

Department of Health
Publisher: Blekinge Institute of Technology
SE-371 79 Karlskrona, Sweden
Printed by Exakta Group, Sweden, 2020
ISBN: 978-91-7295-405-2
ISSN: 1653-2090
urn:nbn:se:bth-19513

To Mario Dallora and Del Prete Sobral Moraes.

“Por vezes sentimos que aquilo que fazemos não é senão uma gota de água no mar. Mas o mar seria menor se lhe faltasse uma gota”

— Madre Teresa de Calcuta

ABSTRACT

Healthcare is an important and high cost sector that involves many decision-making tasks based on the analysis of data, from its primary activities up till management itself. A technology that can be useful in an environment as data-intensive as healthcare is machine learning. This thesis investigates the application of machine learning in healthcare contexts as an applied health technology (AHT). AHT refers to application of scientific methods for the development of interventions targeting practical problems related to health and healthcare.

The two research contexts in this thesis regard two pivotal activities in the healthcare systems: diagnosis and prognosis. The diagnosis research context regards the age assessment of the young individuals, which aims to address the drawbacks in the bone age assessment research, investigating new age assessment methods. The prognosis research context regards the prognosis of dementia, which aims to investigate prognostic estimates for older individuals who came to develop the dementia disorder, in a time frame of 10 years. Machine learning applications were shown to be useful in both research contexts.

In the diagnosis research context, study I summarized the state of the art evidence in the area of bone age assessment with the use of machine learning, identifying both automated and non-automated approaches for age assessment. Study II investigated a non-automated approach based on the radiologists' assessment and study III investigated an automated approach based on deep learning. Both studies used magnetic resonance imaging. The results showed that the radiologists' assessment as input was not precise enough for the estimation of age. However, the deep learning method was able to extract more useful features from the images and provided better diagnostic performance for the age assessment.

In the research context of prognosis, study IV conducted a review on the relevant evidence in on the prognosis of dementia with machine learning techniques, identifying a focus on the research on neuroimaging studies dedicated to validating biomarkers for pharmaceutical research. Study V proposed a multifactorial decision tree approach for the prognosis of dementia in older individuals as to their development or not of dementia in 10 years. Achieving consistent performance results, it provided an interpretable prognostic model identifying possible modifiable and non-modifiable risk factors and possible patient subgroups of importance for the dementia research.

ACKNOWLEDGMENTS

The work done on this thesis was only possible due to the invaluable assistance of several people who, in one way or another, contributed to its completion. I extend my gratitude to each of them.

First of all, I would like to express my deepest gratitude towards my supervisors and mentors Dr. Peter Anderberg, Prof. Johan Sanmartin Berglund, Dr. Martin Boldt, Prof. Emilia Mendes and Prof. Ana Regina Rocha for the support, guidance, advice, critiques and fruitful collaboration, without all of these this thesis would not be possible.

I would also like to thank my colleagues from the Department of Health and fellow Ph.D. students from other departments for the friendly work environment, shared experiences, discussions and coffee chats. A special thanks to Jefferson Molléri, Raquel Ouriques, Thomas Sievert, Vinicius Barbosa, Ola Kvist and Martin Brogren.

My deepest gratitude goes to Alexandre Ribeiro and my family, my parents, Marcelo and Juliana; my brother, Guilherme; and grandmas, Marisa and Mariana for their unconditional love, support and patience throughout this journey. I would also like to thank many friends that, even from across the ocean, made me see the bright side in every situation and made me laugh even in the hardest times.

This work was supported by National Board of Health and Welfare of Sweden (Socialstyrelsen) and the Swedish National Graduate School on Ageing and Health (SWEAH).

PUBLICATIONS

The author participated actively in the planning, designing, conducting the experiments (except for study III), conducting analyses and writing the manuscripts, of the included studies in this thesis.

Study I :

Dallora, A. L., Anderberg, P., Kvist, O., Mendes, E., Ruiz, S. D., & Berglund, J. S. (2019). Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PloS one, 14(7).

Study II:

Dallora, A. L., Kvist, O., Berglund, J. S., Ruiz, S. D., Boldt, M., Flodmark, C. E., Anderberg, P., Predicting chronological age in living individuals with machine learning methods using bone age assessment staging and non-radiological aspects in a multifactorial approach. Submitted to JMIR Medical Informatics (2020).

Study III:

Dallora, A. L., Berglund, J. S., Brogren, M., Kvist, O., Ruiz, S. D., Dübbel, A., & Anderberg, P. (2019). Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach. JMIR Medical Informatics, 7(4), e16291.

Study IV:

Dallora, A. L., Eivazzadeh, S., Mendes, E., Berglund, J. S., & Anderberg, P. (2017). Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. PloS one, 12(6).

Study V:

Dallora, A. L., Minku, L., Mendes, E., Rennemark, M., Anderberg, P., Berglund, J. S. (2020). A decision tree multifactorial approach for predicting dementia in a 10 years' time. Manuscript

ABBREVIATIONS AND ACRONYMS

ADNI: Alzheimer's Disease Neuroimaging Initiative

AHT: Applied health technology

AUC: Area under the curve

CIND: Cognitive impairment no dementia

COVID-19: Coronavirus 2019

BAA: Bone age assessment

IBM: International Business Machines Corporation

MAE: Mean absolute error

MCI: Mild cognitive impairment

MLP: Multi-layer perceptron

MRI: Magnetic resonance imaging

PICO: Population, Intervention, Comparison and Outcome

RCT: Randomized clinical trial

RMSE: Root mean square error

ROI: Region of interest

SAAS: Swedish Age Assessment Study

SLR: Systematic literature review

SNAC: Swedish National study on Aging and Care

SVM: Support vector machines

WHO: World Health Organization

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

1 INTRODUCTION

Health research is beneficial to society as it can lead to important discoveries such as new treatments, tests, vaccines, prevention strategies and other general improvements that affect healthcare and its delivery to the general population. The information technology breakthrough into the health sciences was a significant advance in this field as it allowed clinically rich information to be available in a structured format allowing the development of information-based research in this area [1]. This type of research can be faster and cheaper in its execution, while allowing the analysis of big portions of data to identify unexpected scenarios, patterns in subgroups of subjects that would typically not be included in controlled experiments, work with complex illnesses which not much is known about it [1].

Modern healthcare experiences an expansion of health data in terms of volume, variety and availability, which reflects on the research's methodologies that demand more robust statistical techniques in order to make the most of this health data [2,3]. One instance of such robust techniques is machine learning [3]. Machine learning has been applied in the healthcare scenario for decades, in diverse areas such as cancer, diabetes and genomics research [4–6]. In more recent events, machine learning is being employed in one of the most severe public health crises in the last years, the outbreak of the new Coronavirus Disease 2019 (COVID-19). Machine learning helped scientists in the classification of COVID-19 genomes [7], identification of possible drug candidates to be used in trials [8], prediction of protein structures related to COVID-19 for the formulation of vaccines [9] and in other useful applications [10].

Machine learning has the potential to provide great value in the healthcare area with lots of possible applications. The practice of medicine is composed of four primary practices: diagnosis, prognosis, treatment, and prevention. This thesis will explore machine learning in two different contexts related to two of the pivotal practices in the delivery of healthcare: diagnosis and prognosis. These contexts are: (*i*) the assessment of the age of young individuals, based on their bone development; and (*ii*) the prognosis of dementia in the older population, regarding its development or not in healthy individuals after a time frame of 10 years, respectively.

2 SCOPE OF THE THESIS

In this thesis, the perspective taken on machine learning is as a technology used in the realm of Applied Health Technologies (AHT). To discern the field of AHT it is essential to consider its health technology aspect and its application aspect. Health technology is defined by The International Network of Agencies for Health Technology Assessment [11] as: "a*n intervention developed to prevent, diagnose or treat medical conditions; promote health; provide rehabilitation; or organize healthcare delivery. The intervention can be a test, medicine, vaccine, procedure, program or system*". The application aspect regards its applied science facet, which implies the use of the scientific method for practical purposes towards specific problems [12]. While considering these aspects, it is possible to delineate the work on AHT field as the employment of the scientific method to study technologies to develop interventions that target specific problems or applications in regards to health and healthcare. The Blekinge Institute of Technology Health Technology Research Lab [13] defines AHT as "*an interdisciplinary subject incorporating both studies on how health directly or indirectly relates to the application and results of technology*". This definition highlights the human context of the AHT field, which is essential considering the possibilities and limitations provided by health technologies, in order to support and improve health, and ease suffering [14,15]. Being a research area that interfaces health and technology sciences, the interdisciplinary factor plays a significant role in AHT. It allows the collective thinking about complex problems [16], like the ones commonly encountered in the healthcare field, in order to reach a new understanding and explanations that assist in finding better overall solutions, or into developing new and more insightful research questions [17].

The digitization (i.e. conversion of the analog to digital) of many healthcare functions enables the employment of new technologies for implementing novel ways of delivering healthcare [18]. Digitization is highly correlated to the field of AHT in the sense that it allows targeting specific problems in healthcare with the help of technologies, and in this thesis, the particular case of machine learning will be investigated. It is important to note that traditional works on machine learning are often algorithm-centered, focusing on implementations, optimizations, theories, and comparisons with benchmark datasets. However, this thesis focuses on the applications, i.e., the contexts in which a specific healthcare problem is identified, and the opportunity that machine learning offers in targeting them and offering novel solutions.

3 THESIS OUTLINE

This thesis is a synthesized framework based on five studies. The remainder of this work is organized as follows: chapter 4 (Background) presents important concepts and background information for the motivation of this work, also presenting the research contexts of this thesis; chapter 5 (Research Aims) details the overall and specific aims of this thesis; chapter 6 (Research context in diagnosis: age assessment of youth and young adults) presents the methodology used to achieved the research aims and main findings of the research context regarding diagnosis; chapter 7 (Research context in prognosis: prediction of the development of dementia in older individuals) presents analogous information to chapter 6, but for the research context regarding prognosis; chapter 8 (Ethical Considerations) discusses the ethical issues pertaining both research contexts; chapter 9 (Discussion) presents discussions about the main findings regarding the research contexts and the work as a whole, also this chapter presents a discussion about threats to validity; chapter 10 (Conclusion) provides the concluding remarks of the thesis work; finally, chapter 11 (Future Work) presents suggestions for possible future studies.

The studies included in this thesis are presented in chapters 13 to 17.

4 BACKGROUND

To better understand the application of machine learning in healthcare, first an explanation of its basic concepts is provided. From this point, the scenario of modern healthcare is presented, together with possible applications of machine learning to diagnosis and prognosis. Then, the two contexts of application, which concern this thesis are presented in detail, specifying the problems which machine learning aims to solve. Finally, the rationale for the use of machine learning is provided for both contexts.

4.1 Machine Learning

Artificial intelligence is a broad discipline that concerns the understanding and the designing of computer programs that exhibit key elements of intelligence [19]. The ability to learn is one of those critical elements and is explored in its sub-field named machine learning. The origin of the term machine learning dates back to the 1950s with its definition given by Arthur Samuel as: "*a field of study that gives computers the ability to learn without being explicitly programmed*" [20]. Samuel's research proposed to investigate if a computer program would be able to learn enough from past experiences to outperform an average person in the game of checkers. His algorithm was one of the earliest documented examples of machine learning and it showed that after about eight hours, the program learned to improve its checkers skills to eventually beat an average human opponent [21]. Decades later, in 1997, after many hardware and software developments, the International Business Machines Corporation's (IBM) Deep Blue machine learning program defeated the world chess champion Garry Kasparov, an event that was a historical landmark for machine learning [22]**.** Both examples illustrate the striking capability and evolution of machine learning algorithms that caused a paradigm shift in the Computer Science area, which primarily focused on '*how to manually program computers*', into developing a whole new field with the focus on *'how to get computers to program themselves'* (based on past experiences and basic initial settings) [23].

Being a field that has been gaining progressively more attention in the last decades, machine learning has a range of successful applications in many areas such as speech recognition (e.g. dictation systems), computer vision (e.g. face recognition), bio-surveillance (e.g. disease outbreak detection and tracking), robot control (e.g. helicopter flight stabilization), and a crescent interest in the empirical sciences by aiding the scientific discovery process in areas that are very data-intensive [23–25].

When talking about machine learning algorithms, a critical reflection to be made regards the question of *why* machines should learn instead of just designing them to perform as desired. The reasons for that have a lot to do with the limited human capacity to deal with large amounts of data, multiple variables at the same time and ever-changing environments [26]; all of which would make manual coding impractical (if not impossible), such as in the following tasks [27]:

- tasks that are too complex to be well defined, except by examples in a series of inputs and outputs;
- tasks that involve extracting meaning (relationships and correlations) in large quantities of data;
- tasks that involve on-the-job improvement, since not all of the characteristics of the environment are known;
- tasks with a very large amount of available knowledge that makes explicit coding impractical; and
- tasks that involve dynamic environments that would require constant redesign.

From a general perspective, a machine learning system operates by learning how to perform a task from a set of examples or observations, and by improving its performance while executing the defined task [27]. To say that learning happened means that the system in question changed its structure, programming or data, in order to improve its performance on the defined task, based on the examples or observations. The knowledge acquired on how to perform the task in question is encapsulated in the form of a model, so when given an entirely new set of data, this model can carry out the task within an acceptable performance level, assuming that the training data is representative of the new data [28].

For a better understanding of the works in this thesis, the basic machine learning terminology is defined in the following.

- **Data point**, **observation**, **example,** and **instance** refer to single and independent units of data. When referring to a set of data points the term **sample** will be used [29].
- **Training set** is a subset of data points used to develop models, while the **validation set** and the **test set** are used for evaluation purposes [29].
- **Model** is a representation of what the machine learning algorithm was able to learn from the training set [30].
- **Predictors**, **features**, **attributes** and **independent variables** refer to the model's input variables, which are employed in order to make predictions [30].
- **Outcome** and **dependent variable** refer to the target of the prediction [29].

- **Label** refers to the answer or result in regards to an example. A **labeled example** is an example that is constituted by features and a label [30].
- **Classes** are a set of enumerated label values which an outcome can assume [30].

There are several learning approaches within machine learning, which are used in different types of problems. The most common ones are supervised, unsupervised and reinforcement learning. However, in the studies that constitute this thesis only supervised approaches were used, so the remaining ones are considered out of the scope of this background explanation. In **supervised learning**, the examples or observations that are used to train the models have known labels that correspond to correct outputs, and the algorithm will execute in order to map the inputs to the outputs [26]. The most common types of problems in this learning approach are **classification**, in which the output is characterized by a finite number of classes; and **regression** in which the output is a real-value number [4].

Most machine learning problems tend to be well-constrained, presenting no significant challenges in representing the inputs into relevant variables to be used in building models [31]. However, complex real-world problems tend to involve data related to natural signals, natural sounds, language, natural images or visual scenes, which are difficult to be translated into descriptive means [31,32]. To deal with this challenge, a new modality of machine learning algorithms was introduced: deep learning. **Deep learning** can be employed with multiple learning approaches and has a fundamental difference from the most traditional machine learning modalities. While the latter requires the domain experts to hand-craft features, deep learning performs automatic feature engineering, meaning that it automatically finds the important informative features in the data first, and then performs the designated task (e.g. classification, regression) [33].

Both deep learning and more traditional forms of machine learning work in order to build models with the objective of making accurate predictions [29]. Another interest that can arise when building a model is interpreting it to understand why the model works and how are the features working together in order to estimate the target outcome [29]. Understanding the modeling priority is crucial since there is a critical tradeoff to be considered, since in most of the cases, the more complex the algorithms get, the less interpretable they are [34].

In brief, machine learning is not new and it is a field that has been evolving in the last decades, acting on more and more complex tasks. It offers the potential to solve problems that would be either impossible or impractical to be solved by manual coding and it is much present in modern life, from weather reports, internet search

engines and online shopping, to applications that have the potential to greatly affect individuals' lives, which is the case of healthcare.

4.2 Machine learning in healthcare

Healthcare is an essential and high-cost sector that, in the last 17 years, has been growing faster than the economy [35]. The global spending on healthcare in 2017 was estimated to have reached US$ 7.8 trillion, which corresponds to about 10% of the global gross domestic product in that year [35]. This increased expenditure is due to a multitude of factors related to population growth and aging, disease prevalence and incidence, increased demand for health services, and rise in the price of health services and pharmaceutical costs [36]. This upward trend in the costs of care brought difficult challenges related to efficiency and productivity to the healthcare systems [19] and was one of the driving factors for change in this sector that used to be primarily focused on hospital care, to give higher importance to preventative actions and outpatient care [37].

Healthcare systems are made of multiple information processing tasks. First, there is the screening and diagnosis, which consists of the classification of patient cases through the information given by examinations, investigations, and history. Then, there are the treatment and monitoring tasks, which comprise the planning, implementation, and managing of specific actions, based on the presented information, in order to provide future health outcomes for the patients [19]. Managing a healthcare system is also a very data-driven task since policy makers and managers need to, optimally, maintain, modify and allocate resources to the health system functions in order to achieve its goals of care, based on the information on the usage of services, bio-surveillance evidence etc [19]. In short, the healthcare sector comprises many decision-making tasks that are based on the analysis of large quantities of data, in different degrees of complexity, and multiple levels, ranging from its primary activities up till management itself. A technology that can be useful in such data-intensive environments is machine learning.

The employment of machine learning for health applications is not new, and it is experiencing a continuous expansion in both public and private sectors [38]. This expansion is made possible by the large volume of health data available in both structure form, as in the electronic medical records, and unstructured form, provided by clinicians' notes, reports, discharge summaries, medical images, audios and videos [37]. In terms of the applications themselves, they are numerous and are, in essence, driven by the basic definition of machine learning, which is a program that learns from experiences and derives knowledge from data. In this sense, machine learning in the health area works by acquiring knowledge from the decisions made by a vast

number of clinicians (and other health-related professionals) and the outcomes from a large number of patients, and have this collective experience be used in order to inform the care of new patients [39]. How machine learning can address two of the most pivotal activities in the realm of healthcare: diagnosis and prognosis, are further explored in the next sections.

4.2.1 Diagnosis

Diagnosis is the process responsible for explaining a patient's health condition through clinical reasoning and, when necessary, it establishes a possible treatment path [40]. In essence, it is a classification tool, in which the medical knowledge is used to eliminate disease complexes that are not related to the symptoms presented by the patient until a final list of possible conditions is reached [41,42].

The application of machine learning in medical diagnosis is very present in the area of computer-aided diagnosis (CAD). CAD systems use machine learning to assist the physicians' diagnosis by building models from a set of medical examinations and diagnoses outputs. Thus, data collected from routine care could be used in predicting a likely diagnosis during a consultation [43]. An area of diagnosis that is benefited by machine learning is medical imaging. With the advances of technology, new modalities of medical imaging were developed, such as Computer Tomography, Positron-Emission Tomography, and Magnetic Resonance Imaging, which allowed new forms of visualization and enabled more useful information to be gathered from the patients. These modalities produce a large number of images, which the health professionals must interpret in order to derive a diagnosis [44]. Machine learning systems in this area assist in organ and lesion segmentation; image fusion; image-guided therapy; image annotation; image retrieval; and in CAD systems [45].

4.2.2 Prognosis

Prognosis is a significant component of medicine, which refers to the estimation of the risk, for an individual, of developing specific outcomes based on clinical and non-clinical features [46]. It is commonly referred to as the expected outcome of a disease, but the prognosis is also useful at predicting the future outcome of healthy individuals, e.g., cardiovascular risk profiles to determine risks of heart disease in the general population [46].

In order to build a good prognostic estimate, a longitudinal view of the individuals should be considered [39], along with multiple features, which is an ideal setting for the application of machine learning techniques [39]. The way that machine learning can be useful for prognosis is by analyzing health trajectories of a large number of patients in order to identify patterns, which can be used in both individual and

population levels [39]. On the individual level, it helps patients understand the course of their condition so that they can have joint decisions with their healthcare providers, one of the pillars of patient-centered approaches [47]. On the population level, it acts as proactive support, by identifying persons at significant risk of developing a particular condition, and also in predicting the usage of healthcare services [39]. This information is useful for the allocation of resources in order to create and maintain preventive healthcare programs [47].

4.3 Research contexts

The research presented in this thesis is concerned with two applications of machine learning in healthcare, one related to diagnosis and the other related to prognosis in two different contexts. These are detailed in the following.

4.3.1 Research context in diagnosis: age assessment of youth and young adults

According to the United Nations Children's Fund (UNICEF), only half of the children in the developing world aged five or under have a registered birth certificate [48]. The consequences of not having a valid identification put minors in a vulnerable situation, which makes it easier for underage recruitment to fighting forces, underage marriage, and dangerous employment [48]**.** Moreover, with the increase of immigration, the consequence of various conflicts around the world, young asylum seekers with no means of proving their age can be wrongly considered as adults and be deprived of their rights granted by the United Nations Convention on the Rights of the Child [49]. The lack of valid documentation to prove age also affects numerous legal contexts, especially in cases related to adoption, pedopornography, criminal proceedings, and age fraud in sports competitions [50,51]. Processing children as adults in legal contexts have severe consequences since the application of the law is usually harsher for adults and can put vulnerable individuals through situations unfit for their maturity level, hindering the access to adequate support for reintegration to society. In contrast, the opposite situation benefits the adult individual with, arguably, unfair lenient sentences [48].

Bone age assessment (BAA) is a diagnostic tool currently used for assessing age in situations where valid documents are lacking [52]. It assesses the skeletal maturity of an individual through the analysis of the bones' primary (diaphysis) and secondary (epiphysis) ossification centers, where cartilage tissue gradually develops into bone tissue, which drives changes in the bones' size, shape and degree of mineralization [53]. The diaphysis and epiphysis continue to grow, while there is remaining cartilage

in the growth plate, and when it ceases, it is said that the growth plate is ossified (or fused) [53].

The most commonly used methods for BAA are the Greulich-Pyle (GP) and Tanner-Whitehouse (TW) methods, both of them are based on the analysis of diaphyses and epiphyses of the hand, through the use of radiographs. The first method is based on the comparison of radiograph images of a hand atlas, crafted in the 1940s, with the radiograph of the individual being assessed [54]. The TW method computes an aggregated score of skeletal maturity assessments of 13 short bones of the hand, ulna, and radius [55]. The TW method was developed in the 1950s and was further updated in 2001. Nonetheless, neither of these BAA methods were conceived for chronological age assessment purposes. These methods conveyed groundbreaking developments in numerous clinical applications, regarding diagnosis and in assessing the time for treatment in pediatric endocrinology, orthopedics, orthodontics, growth disorders, and final height estimations [56]. However, their initial intended use is guided at comparing the estimated bone age with the actual chronological age of the individual to deliver a diagnosis. A delayed or advanced bone age can point to endocrine disorders, nutritional deficiencies, chronic illnesses, multiple syndromes, and also be a consequence of the use of certain medications [57].

In regards to the use of bone age for the estimation of chronological age, most of its shortcomings come from the fact that it is done manually by radiologists. First, the activity of analyzing images is time-consuming and can be prone to inter and intra-rater variability [58,59]. Second, there is a serious ethical issue in exposing minors to radiation without any therapeutic purposes, and most of the methods currently in use employ radiographs [59].

In a clinical setting, physicians can work with secure thresholds in order to prescribe treatments and there is a continuous following until a therapeutic goal is reached. However, in a legal context, the act of assessing age is a one-time event that is almost absolute, i.e. "*is the subject a minor or not?*" and this can have life-long effects in the individual's life. Still, compared to other methods that aim at assessing maturation in individuals, BAA is the most trustable. The assessment of the dental age has a high degree of variation, the assessment of sexual characteristics as described in the Tanner Scale [60,61] could be subjective and it is restricted to teenagers, and the age at menarche is a one-off event that only regards women [62].

Thus, it is beneficial to research ways to improve the estimation of age through the assessment of bone age, and a technology that is able to augment the human capacity of extract information in such information-rich environments, such as images, is machine learning. Also, the possibility of automation that comes together with an algorithmic method can impact the time spent on assessment and the rater variability

problem. Lastly, the machine learning technology can also be used to explore other medical imaging modalities that do not make use of ionizing radiation and that are not very widespread in the BAA field, such as magnetic resonance imaging.

4.3.2 Research context in prognosis: prediction of the development of dementia in older individuals

In an article entitled "*The coming epidemic of dementia*", published in 1983, the author Henderson characterized the dementia disorder as '*an unexciting group of diseases occurring in an unexciting age group*' that was not receiving many research incentives, and to which the World Health Organization (WHO) had only begun to shift its attention to it [63]. More specifically about the dementia disorder, Henderson also stated, at that time, that it was a fairly common one, with unknown etiology, with only palliative treatment available, and with a significant social impact [63]. Almost four decades of research later, many advances were made in the field of dementia disorder. However, these four statements still hold true. To date, the WHO holds dementia as a public health priority, with alarming rates of an estimated 7.7 million new cases each year [64].

Dementia comprises a range of neurological disorders that are responsible for progressive cognitive deterioration and memory loss. As dementia progresses, the affected individuals suffer from an accumulation of disabilities and cognitive impairment so grave that it interferes with their social and professional functioning, which can get so severe as to lead to a complete loss of independence [65]. Further, little is known about its mechanisms, and no available symptomatic treatment was able to show relevant benefits in treating the deterioration in cognition [66]. Other severe dementia symptoms include disorientation, mood swings, intensified memory loss, confusion, behavioral changes, impaired gait, impaired speech, and difficulty in swallowing, all of which are responsible for the poor quality of life for the affected individuals [65].

The impact of dementia in society occurs on multiple levels. Besides the individuals diagnosed with this condition, their caregivers, who are usually their close family, are also directly affected by it. The financial loss that occurs in the household, caused by the reduction in income and costs of care, is usually accompanied by adverse health outcomes for them as well, such as high levels of strain, and risks of depression and alcohol-related problems, which are derived from the burden of care for such complex disorder [67,68]. Dementia also has a considerable impact on healthcare systems around the world. In 2015, it is estimated that the global cost of dementia, accounting for direct medical costs, social and informal care costs were US$ 818 billion, which corresponds to 1.1% of the global gross domestic product of that year [69]. Contrary

to other chronic illnesses, like diabetes, which 80-90% of the costs of care are directed to effective disease-modifying interventions, in dementia's case, it is estimated that circa 83% of the costs are directed to social and informal care that aims at compensating the cognitive consequences of the disorder, and roughly 1% was destined to pharmaceuticals, that have a modest effect on the symptoms [65]. Since age is considered to be the most significant risk factor for dementia and life expectancy is increasing worldwide, an increase in prevalence is expected in the future, which will cause an even more significant strain for governments, communities, families, individuals; and loss in productivity for the global economies [69].

In light of this scenario, measures for prevention and risk reduction are imperative [69], and prognostic estimates are vital in addressing the dementia epidemic. A good prognosis based on individuals and their development or not of dementia can identify predictors, and their relative importance [46]. Doing this prediction considering a time frame that is large enough for the application of interventions may lead to positive results as to prevent or delay the dementia onset. The problem with this approach lies in the hypothesis that dementia is a multifactorial disorder, and some factors that are believed to influence its development cannot be studied in traditional trials of control versus intervention (i.e., cardiovascular risk). However, having longitudinal and multifactorial data related to the older adult population enables machine learning to tackle this problem instead.

4.4 The rationale of the thesis

The increase in volume of health data and the necessity of tackling more and more complex problems, makes the use of machine learning in healthcare very beneficial. The rationale behind the use of machine learning in both contexts present in this thesis is detailed in the following.

In the research context that regards the estimation of the age of young individuals, even if the process underlying the bone development is known, its use for predicting the chronological age of an individual is not optimal, since BAA methods were not conceived for this use. Machine learning is employed in this context to uncover knowledge about how the maturation process that happens inside the bones can be used to diagnose the age of young individuals better. Besides, it can also obtain the advantages of an algorithmic approach, which also addresses the drawback in the methods in terms of rate variability and time spent in BAA activities.

On the other hand, in the research context that regards the prognosis of dementia in the older population, the opposite scenario is faced since dementia is a multifactorial disorder in which so little is known about its underlying mechanisms. In this context,

there is a large number of possible input variables at a baseline and respective outputs that regard the development of dementia years later, but no knowledge about how they relate to each other. Thus, machine learning is employed in order to find patterns in the patients' data and deliver a prognostic estimate.

In both research contexts, machine learning is employed in order to augment the human capability of tackling complex problems, by either extending the knowledge about a health concept or uncovering new knowledge in order to fill a research gap that could be beneficial to society.

5 RESEARCH AIM

The overall aim of this thesis was to investigate applications of machine learning in healthcare, concerning diagnosis and prognosis applications, in the two following contexts: age assessment of youth and prognosis of dementia in the elderly.

Regarding the diagnosis research context, the main goal was to investigate methods for the assessment of age in youth and young adults, which do not employ ionizing radiation and address the drawbacks in the bone age assessment research. This context was composed of three studies, and their individual aims are defined as the following:

- The aim of Study I was to perform a systematic literature review on BAA methods that makes use of machine learning techniques to present the state of the art evidence, trends and gaps in the research.
- The aim of Study II was to investigate an age assessment method for youth and young adults based on the radiologists' BAA and non-radiological features, using machine learning techniques.
- The aim of Study III was to investigate an automated age assessment method for youth and young adults based on deep-learning.

For the research context of the prognosis, the main goal was to propose prognostic estimates for older individuals, as to their development or not of dementia, in a 10 years' time, considering modifiable risk factors. This context was built on two studies and their individual aims are defined as the following:

- The aim of Study IV was to present a systematic literature review of the state of the art, trends and gaps in the research on the prognosis of dementia that makes use of machine learning and microsimulation techniques.
- The aim of Study V was to investigate the prognosis of dementia in a cohort of older individuals as to their development or not of dementia in a time frame of 10 years, in a broad multifactorial approach, using decision trees.

6 RESEARCH CONTEXT IN DIAGNOSIS: AGE ASSESSMENT OF YOUTH AND YOUNG ADULTS

The works on the research context in diagnosis regarded the age assessment of youth and young adults and were composed of three studies. The starting point was the execution of a systematic review of the literature (SLR) and meta-analysis in order to contextualize the relevant published evidence in this area (study I). Based on the results of the SLR, two machine learning experiments were designed in order to propose chronological age assessment methods: one based on the radiologists' assessments of images (study II), and one based on the automatic assessment of images through deep learning (study III).

6.1 Materials and Methods

A detailed account of the methodologies of the studies I, II and III is presented in the following.

6.1.1 Systematic literature review

Study I was performed as an SLR and meta-analysis in the Pubmed, Web of Science and Scopus databases, which aimed at answering the central question: "*How machine learning techniques are being employed in studies concerning youth age assessment (10 to 30 years)?*". This research question was answered in regards to the machine learning techniques, the data characteristics being considered, the type of medical imaging, and the regions of interest (ROI) being explored in the literature. A meta-analysis was carried out on the performance of the proposed age assessment methods. A review protocol[1] was made, beforehand, in order to define the inclusion and exclusion criteria (see table 1), a search string (see table 2), a quality assessment checklist (based on Kitchenham and Charters' [70] guidelines) and procedures for data extraction and synthesis. The search for studies was performed on March 21st of 2018 and 6th of February of 2019. The study selection process followed the steps shown in figure 1, except that for the second search no snowballing was performed. The retrieved studies were divided in thirds and had their titles and abstracts assessed by the participants of the study. The selected studies had their references assessed in a one iteration backward snowballing in order to find additional studies. All selected

[1] The protocol regarding the SLR of the study IV is available at http://tiny.cc/4wuw8y

studies were fully read for eligibility and quality assessment, then selecting the final set of included papers. Summary tables were built in order to compile the information for the SLR, while the meta-analysis of the performances was calculated on the average performance weighted by sample sizes.

Table 1. Inclusion and exclusion criteria for the selection of studies in study I

Inclusion Criteria	Exclusion Criteria:
• Be a primary study in English; AND • Published in the last 10 years; AND • Address age assessment using medical imaging; AND • Propose age assessment models using machine learning techniques; AND • Age assessment is analyzed through growth zones in joints; AND • Be a study regarding age assessment in living individuals.	• Be a secondary or tertiary study; OR • Have been published before 2007; OR • Written on a language other than English; OR • Do not address research on age assessment using medical imaging; OR • Do not propose models for the purpose of age assessment; OR • Address height prediction, or a specific syndrome or disease that affects normal growth; OR • Age assessment is not analyzed through growth zones in joints; OR • The study population is out of the range of 10 to 30 years old by 3 years; OR • Use of post-mortem material.

Table 2. Population, intervention, comparison and outcome (PICO) components, which guided the building of the search string used in the SLR of study I

Component	Description
Population	Studies involving age assessment in youth.
Intervention	Use of medical imaging
Comparison	No comparison was defined since the study aimed to characterize the research, therefore there was no comparison with other interventions.
Outcome	Machine learning models for age assessment.

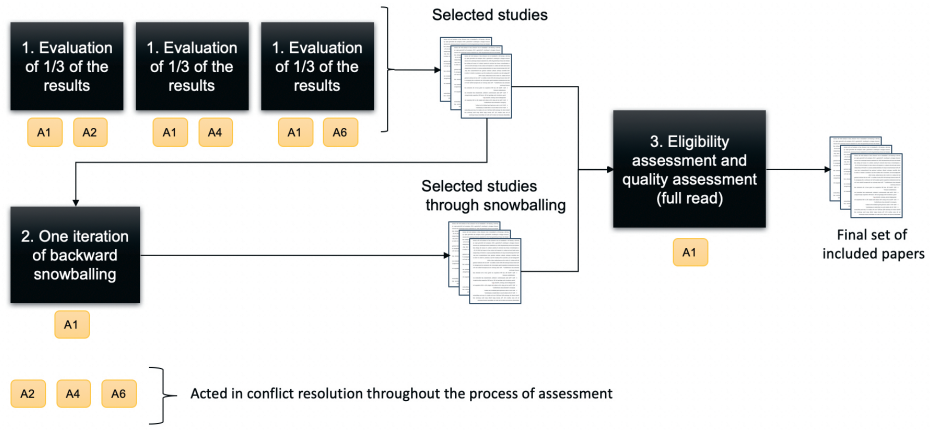


Figure 1. Study selection process used in the SLR of study I. Abbreviations**: A1, A2, A4** and **A6** refer to the first, second, fourth and sixth authors of the published paper from study IV, respectively.

6.1.2 Experiments

Both Studies II and III investigate methods for both: (i) estimation of chronological age; and (ii) classification of minors and adults, regarding the threshold of 18 years. The results from the SLR and meta-analysis of study I showed that the proposed methods for BAA consisted of automated and non-automated approaches in regards to the need of human input to the system.

Study II investigates a non-automated approach of age assessment based on the radiologists' assessment of medical images. Study II also addresses drawbacks and identified gaps in the literature related to: (i) the use of non-ionising radiation modality of medical imaging, with the use of magnetic resonance imaging (MRI); (ii) the employment of three regions of interest (ROI) for the BAA: wrist, knee and foot; and (iii) the inclusion of non-radiological factors in the assessment.

Study III investigates an automated approach for age assessment based on deep learning, which was employed in most of the automated methods identified by study I. Study III also addresses other identified gaps in the literature and drawbacks in the traditional methods used for BAA related to: (i) the use of non-ionising radiation modality of medical imaging, by the use of MRI; (ii) the analysis of the knee region, which was scarcely explored in the literature; (iii) mitigates the risk of rater variability, since there is no human input in the method; and (iv) it reduces the time-consumed in the assessment, by proposing an automated approach.

6.1.3 Population and data

Studies II and III were part of the Swedish Age Assessment Study (SAAS) - Study to Deepen the Knowledge of Magnetic Camera Examinations as a Method for Medical Age Assessment project, which aims to: "*provide, on the basis of best available knowledge, suggestions for methods of medical age assessment of whether a person is over or under 18 years of age*" [71]. This project conducted MRI examinations on healthy volunteer subjects in the age range of 14 to 21 years, during the period of 2017 and 2018. All MRI examinations were performed on 1.5 Tesla whole-body MRI scanners, and followed the same protocol with settings of 256 x 256 pixel resolution and 160 by 160 mm field of view. The inclusion and exclusion criteria that determined the participation in study are shown in table 3.

Table 3. Inclusion and exclusion criteria for the participation in the SAAS project

Inclusion Criteria	Exclusion Criteria
• Have been born in Sweden; AND • have a birth certificate verified by the Swedish national authorities.	• History of bilateral fractures or trauma near the regions of assessment; OR • history of chronic disease; OR • use of long-term medications; OR • noncompliance during the examination; OR • have resided outside Sweden for more than six consecutive months; OR • past or current pregnancy (all female subjects were tested).

The data for study I consisted of 455 male and 467 female subjects (see table 4), who had T2-weighted cartilage dedicated exposure MRI images taken from the foot, knee and wrist in regards to five growth zones: Calcaneus (foot), Distal Tibia (foot), Proximal Tibia (knee), Distal Femur (knee) and Radius (wrist). These images were independently assessed by two pediatric radiologists (plus one for conflict resolution), blinded to age and gender. Each of the five growth plates was attributed to a stage based on the appearance of cartilage signal intensity of the growth plate, according to the staging systems proposed by Dedouit et al. [72] and Kellinghaus et al. [73] with minor modifications. The non-radiological data was gathered by a questionnaire given to the subjects, and measures of weight and height at the examination sessions. The non-radiological data used in the models were: type of residence during upbringing, daily level of physical activity, parents' origin, Self-assessed Tanner Scale for pubertal growth [60,61] and Body Mass Index (BMI) .

Table 4. Demographics of study II subjects

Age Group	14	15	16	17	18	19	20	21	Total
Number of female subjects	59	58	57	60	59	57	57	60	**467**
Number of male subjects	58	56	60	58	53	58	53	59	**455**

The data for study III consisted of 221 male and 181 female subjects (see table 5) who underwent MRI examinations for T1-weighted images with bone dedicated exposure, of the knee region. No radiologists' assessment was employed in study III.

Table 5. Demographics of study III subjects

Age Group	14	15	16	17	18	19	20	21	Total
Number of female subjects	22	21	30	27	20	12	25	24	**181**
Number of male subjects	22	26	31	25	24	25	35	33	**221**

6.1.4 Study II experimental design

Cohen's kappa coefficient [74] was calculated in order to measure the inter-observer agreement between the pediatric radiologists regarding the assessment of the MRI images in all growth plates separately. Two types of age assessment models were investigated in study II: binary classification of minors and adults regarding a threshold of 18 years; and a multi-class classification of individuals into one of eight classes of age (14 to 21 years). The data preparation procedure consisted of applying the K-nearest Neighbors (KNN) Imputation to deal with missing data (which in none of the variables was higher than 1.9%). The machine learning algorithms were chosen based on the findings regarding the non-automated approaches in study I and an additional search in the literature. The following machine learning algorithms were chosen to build the age estimation models: Decision Tree, Random Forest, Multi-layer Perceptron, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors.

All experiments were performed in a nested cross-validation setup (5-fold outer, 3-fold inner) with stratified data splits. Additionally, a grid search was performed in order to find the best hyperparameters. The evaluation metrics used to measure model performance were: Area Under the Curve (AUC), Accuracy, Precision, Recall, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The latter two were only used in the multi-class classification models. These metrics were chosen based on guidelines for binary and ordinal multiclass classification problems [75,76]. The Performance for each of the employed algorithms was calculated in terms of the mean

and standard deviations of the appropriate metrics on the outer cross-validation test sets. The best performing algorithms had their performance results detailed for each of the outer cross-validation test sets. The choice of the final models is given by the one which gives the median results in order to minimize the risk of selecting over optimistic models.

6.1.5 Study III experimental design

The proposed method for the estimation of age in study III is based on deep learning and comprises two Convolutional Neural Networks (CNN) models, as shown in figure 2.

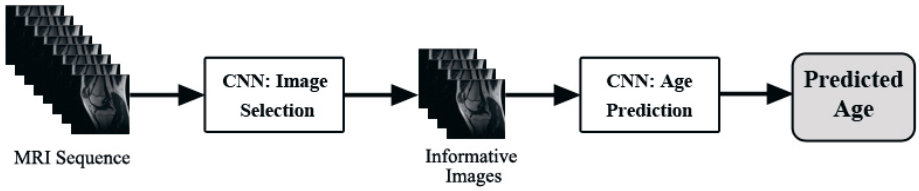


Figure 2. Overview of the proposed automated age assessment method

To train the 'CNN: Image Selection', one image from each of the MRI sequences was labeled 'informative' (i.e. contain anatomical structures of interest) and one 'noninformative' (i.e. anatomical structures of interest occluded). The output of this CNN model is the confidence level of the two classes (informative and noninformative), for a given MRI image of the sequence. The confidence level is a continuous value that ranges between 0 (lowest) and 1 (highest), with the confidence levels of the two classes sum up to 1. The GoogLeNet [77] CNN architecture was used in order to build this model.

To train the 'CNN: Age prediction' module seven different CNN architectures were considered: GoogLeNet [77], ResNet-50 [78], Inception-v3 [79], VGG [80], AlexNet [81], DenseNet [82] and U-Net [83]. The final classification layer of these networks was replaced with a linear scalar output to provide the age estimation. The input of the age prediction CNN model is an MRI image with N channels, created from an informative image as the center, and a subset of N MRI images around this image is extracted from the MRI sequence.

The Caffe deep learning framework [84] with the Amazon Web Services (AWS) on an EC2 p3.2xlarge with a Tesla V100 Nvidia GPU, was used for training and evaluation. The Adam optimizer [85] was used to minimize the cross entropy loss when training the classifier and the Euclidean loss when training the regressor. All experiments were performed using 6-fold cross-validation, in which four parts were used for training, one part was used for validation during training and one part was used to evaluate the model's performance. The data splits were done in a stratified manner and all of the images from a subject were assigned to a single fold. A sparse grid search was performed on the validation set for each model to tune the hyper parameters. Transfer learning with models pre-trained on the ImageNet database [86] and data augmentation approaches were employed in training the CNNs in order to deal with smaller datasets and avoid overfitting. Data augmentation was performed on all training samples, which were randomly cropped, shifted, rotated (at a maximum of five degrees) and scaled (up to 20%).

The estimation of age on the test set used all subject images with a confidence level of at least 0.95 (up to 10 images per subject). Data augmentation was applied on these 15 times, using the same procedures as in the training set. Age was estimated for all augmented images and the median of the estimated ages for each subject comprised the final prediction. Performance results are given in terms of the Mean Absolute Error (MAE).

Experiments on the classification of minors versus adults in the threshold of 18 years were also carried out, however, no new training of models was done. This classification was performed by applying a threshold to the estimated age from the best models trained in the age estimation experiments.

6.2 Results

The results provided by the studies I, II and III are presented in the following.

6.2.1 Study I

After the study selection process, 26 studies were included for data extraction and analysis.

The results show an interest in the research in proposing automated BAA approaches, however, a considerable amount of studies proposed non-automated approaches. In the first case the system operates in a fully automated way, without human intervention, which implies automatic feature extraction and analysis. In the non-automatic approaches some kind of human intervention was employed, e.g. manual location of regions of interest, assessment of the image by radiologists as to specific

stages of ossification or tissue analyses. Both address in a higher or lower degree the problems of subjectivity, time spent on or costs of the assessment. Most of the studies investigated BAA approaches that involved the analysis of radiographs of the hand ROI. Studies' samples were mostly from North America and West Europe, however ethnic aspects were scarcely present in studies. Additionally, socioeconomic or physical activity aspects were not considered. Most of the age ranges in the studies bordered or contained the age of 18 years, showing particular interest in this age.The most frequently employed machine learning techniques evidenced in the included papers were Regression-based methods (13 studies), followed by Artificial Neural Networks (4 studies), Convolutional Neural Networks (4 studies) and Support Vector Machines (5 studies), in which the later three were mostly associated with the automated approaches. Other less frequent techniques featured in the studies are: Bayesian Networks (2 studies), Decision Trees (1 study) and K-Nearest Neighbors (1 study).

The included studies presented a high heterogeneity in terms of age-ranges, dataset sizes and employed performance metrics that hindered the comparison of most of the studies for the meta-analysis. The average performance weighted by the sample size of the 7 comparable studies (with age ranges somewhat in the range of 0-19 years) resulted in a MAE of 9.96 months.

6.2.2 Study II

The MRI assessments of the growth plates by the radiologists showed that stages 1 and 2 did not show evidence in any of the subjects. Stage 3 was present scarcely, for the Calcaneus and Radius growth plates. For the female subjects, nearly all or most of the sample was already on the last stage of ossification (stage 5) for all of the assessed growth plates (see table 6), and a considerable amount of female subjects, from the age of 17, had already all growth plates in the final stage (stage 5) (see table 7). For the male subjects, these numbers slightly change (see tables 6 and 7). The independent assessment between the radiologists presented substantial agreement in all ROIs, according to the guidelines by Landis and Koch [87].

Table 6. Percentage of the sample in stage 5, by growth plate

	Calcaneus	**Distal Tibia**	**Proximal tibia**	**Distal Femur**	**Radius**
Female Subjects	94.6%	90.8%	81.6%	74.5%	65.5%
Male Subjects	80.4%	70.1%	57.6%	54.9%	47.4%

Table 7. Percentages over each age group of subjects with all of the growth plates in stage 5

Age group	Female subjects	Male subjects
14	3.3%	0%
15	13.7%	0%
16	40.3%	5%
17	73.3%	22.4%
18	89.8%	58.4%
19	100%	86.2%
20	100%	94.3%
21	100%	100%
Total	**65.1%**	**45.2%**

The classification of minor and adults aimed at discriminating subjects at a threshold of 18 years. For the male subjects, the algorithms which presented the best mean and standard deviation on the performance metrics was the Random Forest, which achieved 0.90 ± 0.01 Accuracy , 0.90 ± 0.01 AUC, 0.87 ± 0.03 Precision and 0.94 ± 0.04 Recall. The Random Forest model which produced the median results on the outer cross-validation test sets achieved the following results: 0.90 Accuracy, 0.90 AUC, 0.83 Precision, 1.00 Recall (Minors class) and 0.80 Recall (Adults class). For the female subjects, the best performing algorithm was also the Random Forest, with the following results: 0.83 ± 0.02 Accuracy, 0.83 ± 0.01 AUC, 0.76 ± 0.02 Precision and 0.97 ± 0.01 Recall. The median Random Forest model for the female subjects provided the following results: 0.84 Accuracy, 0.84 AUC, 0.77 Precision, 0.96 Recall (Minors class) and 0.72 Recall (Adults Class).

The age estimation models aimed at the multi-class classification of subjects into eight age groups (14 to 21 years). In the male subjects' case, the algorithm which produced the best mean and standard deviation performance metrics was the Multi-Layer Perceptron (MLP) with the following results: 0.98 ± 0.08 MAE (years), 0.33 ± 0.02 Accuracy, 1.32 ± 0.13 RMSE (years) 0.84 ± 0.01 AUC, 0.65 ± 0.27 Precision and 0.61 ± 0.31 Recall. The median MLP model provided the following results: 0.95 MAE (years), 0.33 Accuracy, 1.29 RMSE (years), 0.83 AUC, 0.91 Recall and 0.48 Precision. Even with moderate results in terms of the MAE metric, the results discriminated by age regarding this model (see figure 3) show a clear trend of overestimation of ages for male subjects in general, limited capacity of classifying subjects over the age of 16. In the female subjects' case, the algorithm which produced the best mean and standard deviation performance metrics was the Support Vector Machines (SVM) with results as follows: 1.21 ± 0.06 MAE (years), 0.32 ± 0.04

Accuracy, 1.68 ± 0.06 RMSE (years) 0.80 ± 0.01 AUC, 0.55 ± 0.07 Precision and 0.71 ± 0.11 Recall. The median SVM model presented the following results: 1.24 MAE (years), 0.37 Accuracy, 1.75 RMSE (years), 0.79 AUC, 0.75 Recall and 0.56 Precision. As in the male subjects' case, there is a clear overestimation of ages for the female subjects in the results discriminated by age (see figure 4).

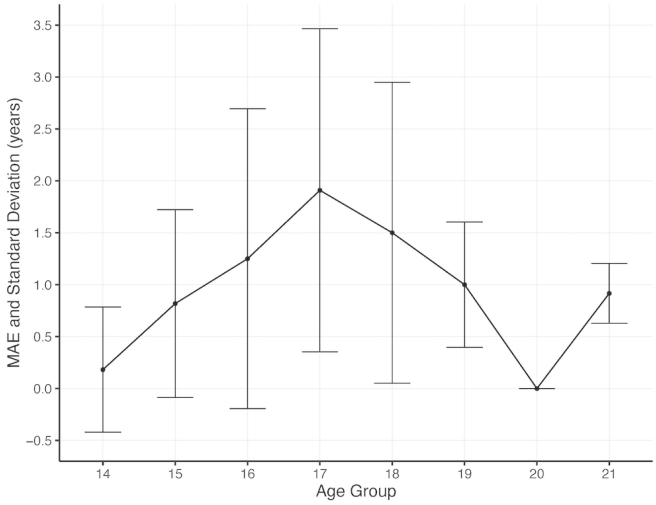


Figure 3. Mean absolute error (years) and standard deviations for the male Multi-Layer Perceptron model

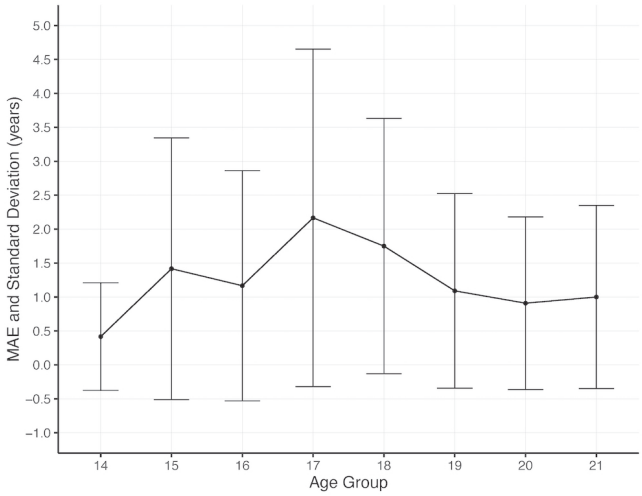


Figure 4. Mean absolute error (years) and standard deviations for the female Support Vector Machine model

The results regarding the age estimation models for both male and female subjects were negatively affected by the fact that a significant part of the sample (65.5% for females and 45.2% for males) had all of the growth plates assessed as to stage 5, which meant that for a considerable part of the sample, the classification would be dependent only on the non-radiological variables (self-assessed Tanner Scale, BMI, residence during upbringing, physical activity and parents origin). This shows that for age assessment purposes, the BAA provided by radiologists was not precise enough for predicting the exact age, but showed good results in the minor and adults classification.

6.2.3 Study III

The best results were achieved using a confidence threshold C of 0.95 in the Image Selection CNN when choosing the most informative MRI images, number of channels set to 1, and the GoogLeNet architecture. The age prediction model achieved a MAE of 0.793 years for male subjects and 0.988 years for female subjects. Table 8 shows the performance results for male and female subjects on the GoogLeNet architecture, discriminated by age group. The MAE increases substantially from the age of 21 for male subjects and 20 for female subjects, which suggests that after the ages of 20.5 for men and 19.5 for women, no additional information regarding age estimation can be extracted from the MRI image data (in the knee region). The 0.5 increments regards the data collection procedure, which collected data from the subjects within six months of their birthday date.

Table 8. Results from the experiments with the best performing models on GoogLeNet

	MAE ± SD (years)							
Age Group	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**
Male	0.74 ± 0.50	0.73 ± 0.80	0.99 ± 1.07	0.98 ± 1.17	1.14 ± 1.19	0.55 ± 0.70	0.51 ± 0.49	1.37 ± 0.59
Female	0.75 ± 0.56	0.89 ± 1.05	1.57 ± 1.08	1.09 ± 1.34	0.61 ± 0.70	0.54 ± 0.55	1.25 ± 0.73	1.75 ± 0.62

Regarding the minors versus adults classification experiment, three strategies were explored for setting the threshold:

1. Increase the accuracy for minors at the expense of the accuracy for adults classification.
2. Setting the threshold to get as equal accuracy as possible for adults and minors.
3. Using the threshold of 18 year without any modification

The results obtained from these strategies for male and female subjects are presented in tables 9 and 10, respectively.

Table 9. Accuracies for minor vs adult classification of male subjects

Strategy	Threshold in years	Accuracy for minors	Accuracy for adults
1	Threshold 18.73	98.1	88.0
2	Threshold 18.38	93.3	93.2
3	Threshold 18.00	90.4	95.7

Table 10. Accuracies for minor vs adult classification of female subjects

Strategy	Threshold in years	Accuracy for minors	Accuracy for adults
1	Threshold 19.11	95.0	45.7
2	Threshold 18.2	85.0	85.2
3	Threshold 18.00	77.0	88.9

7 RESEARCH CONTEXT IN PROGNOSIS: PREDICTION OF THE DEVELOPMENT OF DEMENTIA IN OLDER INDIVIDUALS

The research context that regards the prognosis of dementia comprises two studies and as in the previous research context, had the execution of an SLR as the starting point in order to summarize the relevant published evidence in the area of the prognosis of dementia (study IV). Based on the information provided by the SLR, a machine learning experiment was designed and performed in order to propose the prognostic estimates for the dementia disorder (study V). The remainder of this chapter presents the methodologies and summary of the results pertaining to the studies IV and V.

7.1 Materials and Methods

A detailed account of the methodologies of the studies IV and V is presented in the following.

7.1.1 Systematic literature review

Study IV performed an SLR on the Pubmed, Web of Science and Scopus databases. The review protocol[2] stated the following main research question: "*How are the machine learning and microsimulation techniques being employed by the research on the prognosis of dementia and comorbidities?*". This question was answered in terms of the employed machine learning and microsimulation techniques; the data characteristic considered in the models; the goal of the studies; the handling of data censoring; and the prognosis focus being on an individual or populational level. The review protocol also defined the criteria for inclusion and exclusion of studies (see table 11), search string (components shown in table 12), quality assessment checklist (based on Kitchenham and Charters' [70] guidelines) and for data extraction and synthesis procedures. After building and tailoring the search string, the searches were performed on October 23rd of 2015. The study selection process followed the steps shown in figure 5. It started with an evaluation round of titles and abstracts with a random set of 100 studies, for creating a common ground of assessment between the

[2] The protocol regarding the SLR of the study IV is available at http://goo.gl/6Jddw3.

participants. Then, the titles and abstracts of the remaining studies were assessed. The third activity in the study selection process was a one-iteration backward snowballing on the references of the selected studies in order to find additional evidence. All selected studies were then fully read for eligibility and quality assessment. Also, before the quality assessment, an evaluation round with 10 random selected studies was performed in order to ensure consistency for the quality assessment. The synthesis of results was done through summary tables.

Table 11. Inclusion and exclusion criteria for the selection of studies in the SLR of study IV

Inclusion Criteria	Exclusion Criteria
• Be a primary study in English; AND • address research on dementia and comorbidities; AND • use of at least one machine learning or microsimulation technique; AND • address the prognosis related to dementia and comorbidities.	• Be a secondary or tertiary study; OR • be written in another language other than English; OR • do not address a research on dementia and comorbidities; OR • do not address at least one machine learning or microsimulation technique; OR • do not address a prognosis related to dementia and comorbidities.

Table 12. Population, intervention, comparison and outcome (PICO) components, which guided the building of the search string used in the SLR of study IV

Component	Description
Population	Studies on dementia and its comorbidities.
Intervention	The use of machine learning or microsimulation techniques.
Comparison	No comparison was defined since the study aimed to characterize the research, therefore there was no comparison with other interventions.
Outcome	A prognostic estimate of dementia and its comorbidities.

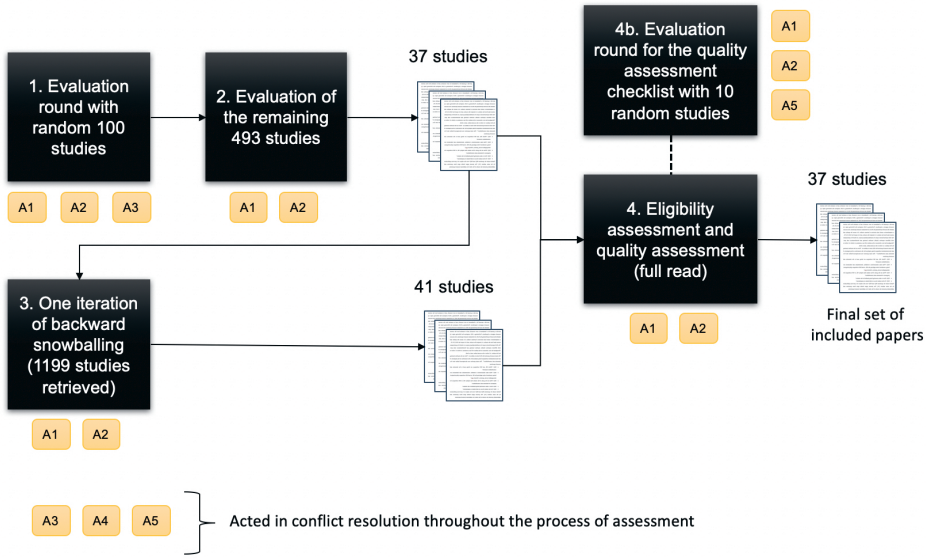


Figure 5. Study selection process followed on the SLR of study IV. Abbreviations**: A1, A2, A3, A4** and **A5** refer to the first, second, third, fourth and fifth authors of the published paper from study IV, respectively.

7.1.2 Experiment

Study V investigates prognostic estimates of older individuals as to their development, or not, of dementia in a time frame of 10 years. Additionally, study V presents the following attributes that address some of the identified gaps in the literature, by study IV:

- A longitudinal population-based approach which considered a time frame for prediction of 10 years, which could open possibilities for the investigation of long-term interventions, which focus on measures to prevent or delay the onset of dementia, instead of pharmaceutical treatments.
- A broad multifactorial approach which considers demographic, social, lifestyle, medical history, biochemical tests, physical examination, psychological assessment and multiple health instruments currently used in addressing dementia; instead of employing unidimensional approaches.

The population of Study V consisted of the baseline examination (collected from 2001 to 2003) of the Swedish National Study on Aging and Care (SNAC) at the Blekinge site. The SNAC project consists of a longitudinal cohort that collects multifactorial data from the older adult population in Sweden, in specific age groups (60, 66, 72, 78, 81, 84, 87, 90, 93 and 96+ years), to be used in the research on aging and care [88]. The following criteria determined exclusion from the study: (i) subjects who already had dementia at baseline; (ii) subjects who had missing values at the

outcome variable (dementia diagnosis at the 10 year mark of the study); (iii) subjects who presented more than 10% of missing values in the input variables; (iv) subjects deceased before the 10 years study mark; and (v) subjects who were diagnosed with dementia before the 10 year mark, since they could already have advanced progress of dementia at baseline. The final sample of the study comprised 726 subjects (313 males and 413 females), of which 91 (12.5%) were given a diagnosis of dementia at the 10-year mark.

In regards to data, study IV did not succeed in identifying important predictor variables, since the research has a substantial focus on neuroimaging. The input variables to be used by the decision tree for the prediction of the subjects who will develop dementia in 10 years were selected by senior researchers specialized in geriatrics and gerontology, considering a multifactorial approach for the prognosis of dementia (see table 13). The outcome variable was the diagnosis of dementia at the 10 year mark of the SNAC Blekinge study (2010 to 2013). This outcome variable is binary and can take the values of 'no dementia' or 'dementia'.

A decision tree approach was employed in order to build the prognostic estimates, due to its interpretability and the fact that it is a technique that has been used in medical contexts to investigate prognostic subgroups [99]. Data preparation consisted of the application of the KNN Imputation to handle the missing data, that was performed separately for each class (the K value was set to 3). Recursive Feature Elimination feature selection approach was applied in order to find an optimal subset of variables that maximize a performance [29]. To avoid bias by the decision tree classifier towards the majority class ('no dementia'), a cost-sensitive learning approach was employed, which applied a heavier penalty to the misclassification of the minority class ('dementia') [100]. All of the experiments were performed on a 5-fold outer, 4-fold inner stratified nested cross-validation approach. The reduced number of folds in the cross-validation was due to the high class imbalance, in order to have enough examples of the minority class in every fold. Performance results are reported for each of the outer cross-validation test sets, and the choice of the model is given by the one which gives the median results in order to minimize the risk of selecting over optimistic models. The metrics used for the performance evaluations were: Area Under the Curve (AUC), Accuracy, Recall and Precision.

Table 13. Input variables selected by the specialized senior researchers

Variable Type	Variables
Demographic	Age, Gender,
Social	Education, Holds or not a Religious Belief, Participation in Religious Activities, Voluntary Association, Social Network, Support Network, Loneliness
Lifestyle	Light Exercise, Alcohol Consumption, Alcohol Quantity, Working State at 65 years, Physical Workload, Present Smoker, Past Smoker Number of Cigarettes a Day, Social Activities, Physically Demanding Activities, Leisure Activities
Medical History	Number of Medications, Family History of Importance, Myocardial Infarction, Arrhythmia, Heart Failure, Stroke, TIA/RIND, Diabetes Type 1, Diabetes Type 2, Thyroid Disease, Cancer, Epilepsy, Atrial Fibrillation, Cardiovascular Ischemia, Parkinson's Disease, Depression, Other Psychiatric Diseases, Snoring, Sleep Apnea, Hip Fracture, Head Trauma, Developmental Disabilities, High Blood Pressure
Biochemical Test	Haemoglobin Analysis, C-Reactive Protein Analysis
Physical Examination	Body Mass Index (BMI), Pain in the last 4 weeks, Heart Rate Sitting, Heart Rate Lying, Blood Pressure on the Right Arm, Hand Strength in Right Arm in a 10s Interval, Hand Strength in Left Arm in a 10s Interval, Feeling of Safety from Rising from a Chair, Assessment of Rising from a Chair, Single-Leg Standing with Right Leg, Single Leg Standing with Left Leg, Dental Prosthesis, Number of Teeth
Psychological	Memory Loss, Memory Decline, Memory Decline 2, Abstract Thinking, Personality Change, Sense of Identity
Health Instruments	Sense of Coherence [89], Digit Span Test [90], Backwards Digit Span Test [90], Livingston Index [91], EQ5D Test [92], Activities of Daily Living [93], Instrumental Activities of Daily Living [94], Mini-Mental State Examination [95], Clock Drawing Test [96], Mental Composite Score of the SF-12 Health Survey [97], Physical Composite Score of the SF-12 Health Survey [97], Comprehensive Psychopathological Rating Scale [98]

7.2 Results

This section presents the results achieved by studies IV and V, which are presented in the following.

7.2.1 Study IV

The results from the synthesis of the final 37 selected studies presented the research context of the prognosis of dementia with the use of machine learning and microsimulation techniques, in addition to identifying gaps in this field.

The results show that the focus of the research on the prognosis of dementia is on the use of neuroimaging for predicting the development of Alzheimer's from MCI, mostly using the data from the ADNI database. This indicates that there is a very strong focus of the research in validating brain biomarkers to be used in pharmaceutical research. While efforts in prevention were inexistent, additionally, this also shows that the research is focused on individuals who already present some kind of cognitive deterioration and not how healthy individuals can come to develop dementia in the future. The time frame for the estimations is mostly up to 36 months before the development of Alzheimer's Disease, which may raise uncertainties regarding if this time frame would be adequate to apply interventions. Also, there is an absence of lifestyle and socioeconomic variables in the assessed models, which could be considered a research gap. Data censoring is not addressed in the vast majority of the studies, which may raise questions if right censoring was considered in these studies. The focus of the research is mostly on individual estimations and little attention was given to populational projections about dementia. 35 out of the 37 selected studies employed machine learning methods. In terms of the techniques, the most frequent used were: Support Vector Machines (30 studies), Decision Trees (6 studies), Bayesian Networks (6 studies) and Artificial Neural Networks (3 studies). Note that one study can feature more than one machine learning technique. Only two microsimulation techniques were evidenced: time-to-event and grade of membership.

7.2.2 Study V

The results of the median model, given by the median AUC on the outer cross-validation test sets, are the following: an AUC of 0.735, Accuracy of 0.745, Recall of 0.722 and Precision of 0.289. Figure 6 shows the resulting decision tree of the median model, the main findings regarding it are pointed out in the following.

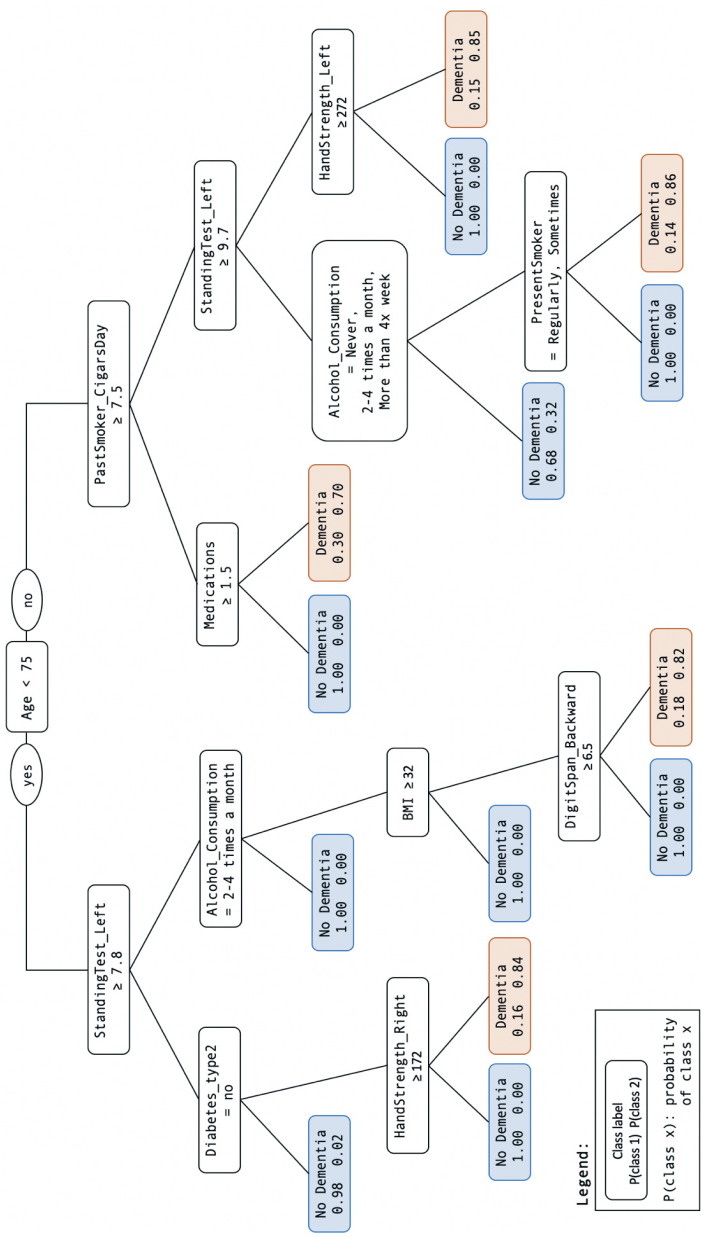


Figure 6. Decision Tree of the median model

The most important predictive factor chosen by the decision tree algorithm was age, which is already believed to be a major risk factor for dementia [65]. Most of the variables present in the decision tree regard modifiable risk factors. These are: physical strength, present smoking, BMI, diabetes type 2 and alcohol consumption. However, without further investigation by trials there is no saying in regards to these factors being protective for preventing or delaying the dementia onset. Physical strength was present in every single branch of the tree. Additionally, there was a lack of variables related to the health instruments currently in use to perform psychological assessments for dementia, which may indicate that these instruments are insensitive for long-term prediction and little sensitive to mild cases. The only exception regarding health instruments was the Backwards Digit Span Test, which measures working memory.

Regarding the performance results, it is important to notice that the lower value for the precision metric is a direct consequence of the cost-sensitive learning approach employed in order to give higher importance to the minority class ('dementia'). The effect of such an approach is a higher misclassification of the majority class, but since the priority relied on identifying factors that lead to dementia in 10 years, this consequence was considered acceptable, especially since the identified predictive factors in the decision tree are very relatable to health advice already given to the general population in order to maintain good health.

8 ETHICAL CONSIDERATIONS

Both research contexts employed activities of data collection of human subjects. These were conducted in accordance with the Declaration of Helsinki with study protocols approved by research ethics committees, as in the following:

- For the research context in diagnosis, the study was reviewed and approved by the Central Ethical Review Board in Stockholm (diary numbers: 2017/4-31/4, 2017/1184-32, 2017/1773-32).
- For the research context in prognosis, the study was evaluated and approved by the Research Ethics Committee at Lund (diary numbers: LU 604-00 and LU 744-00).

In all data collection procedures, written informed consent was collected from the subject; and their legal guardian, in case of minors under the age of 18. In the case of older subjects that presented disabilities, if proxy interviews with relatives were necessary, this was done upon the informed consent of the subject. The recruitments in the studies were done on a voluntary basis and subjects could opt out of the study at any time.

Data privacy procedures were employed in order to anonymize the data, which means that the researchers participating in the study at no point had access to any personal information of the recruited subjects.

For the MRI examinations, regarding the research context on age assessment, some additional ethical issues need to be addressed, mainly because the subjects that partake this research include minors. The international guidelines of the World Health Organization pose the condition that the "risks should be minimized to no more than minimal" [101]. This means that the subjects participating in the research should not be put through a risk that corresponds to a minimal increment than the risk partaken in their daily lives. Two classes of risks could arise from the use of the MRI machinery: physical harm and psychological harm. The risk of psychological harm is derived from the prolonged immobility and the permanence in an enclosed narrow space. To minimize this risk, it was established a maximum duration for the examination (approximately 30 minutes), and the selected regions to be the target of the examination took into consideration the stress levels for the subjects that never enter their head inside the MRI machine, so examinations of the clavicle and arm were not considered for the studies. Additionally, the subject had access to a means to immediately stop the examination in case of any discomfort. In the case of risk of

physical harm in an MRI examination (as a result of the "missile effect" or malfunctioning of medical devices), a screening procedure was employed to assess if any existing condition would make it unsafe for the subject to go through the MRI examination; and also, the machinery went through regular inspections so to ensure safe functioning. In addition to these mitigation procedures, the conduction of MRI examinations was held by radiologists with years of experience and are trained to deal with complications that may arise. During the whole process of data collection, no undesirable irregular events came to happen.

9 DISCUSSION

This chapter presents the discussion regarding the results achieved by the research contexts on diagnosis and prognosis, the application of machine learning in healthcare, and threats to validity on the studies included in this thesis.

9.1 Diagnosis

In the research context regarding diagnosis, an SLR was first performed in order to summarize the relevant evidence on the area of BAA with machine learning techniques (study I). Based on its results, experiments were designed for building models of age estimation and discrimination between minors and adults; in regards to two approaches: one based on the radiologists' assessment (study II) and one based on deep learning (study III). Both studies made use of MRI as a non-ionising radiation imaging modality, acquired by the SAAS project. Study II achieved MAEs of 0.95 years for male and 1.24 years for female subjects on the age estimation experiment (showing limited prediction power for ages older than 15) and accuracies of 90% for male and 84% for female subjects in the classification of minors and adults experiment. Study III achieved MAEs of 0.79 years for males (age range 14-20 years) and 0.99 years (age range 14-19 years) for female subjects on the age estimation experiment, and accuracies of 98% for males and 95% for female subjects in the classification of minors and adults experiment.

One of the main findings of the work on this research context was that the deep learning approach showed superior results in comparison with the one based on the radiologists' assessment, in both types of models, even with the radiologist-based approach considering 5 ROI instead of one. This finding can be associated with the realization that the radiologists' assessment in stages was not precise enough for the estimation of age, since most of the female sample and a great portion of the male sample was on the last stage of the grading scale, in all assessed ROI. The hypothesis is that the deep learning method was able to identify more features in the images, than what is considered in the grading scale that the radiologists used, which focused only on the appearance of the cartilage tissue in the growth plates area. In addition to this, there is a possibility that the deep learning method could have taken into consideration other image elements, like the other ossification centers in the knee area that were not assessed by the radiologists, which are the fibula and the patella. However, being a black-box method, the deep learning approach does not provide an accessible solution, so confident inferences cannot be made as to exactly what was considered

in the model. There are some methods which were proposed in order to '*open'* black box methods. However, these provide low abstraction level of explanation (e.g. which *pixels* are relevant to the estimation); and they explain only 'first-order' information, in which it is possible to identify which features are relevant to the estimation, but not the relation between then, or whether they are important alone or together [102]. These methods were not considered in the works in this thesis, due to the limitations as mentioned above.

Most of the chosen machine learning methods, on both approaches, were black-box methods, with the only exceptions being the decision tree and naïve bayes, 2 out of the 6 methods explored in study II. This choice of methods was motivated by the SLR results and reflects the characteristics of the context of the application, which is the proposition of a diagnostic tool model for age assessment. Ideally, it is desired that a proposed model would present great performance and high interpretability, however, in real world problems this is not often possible and a trade-off between performance and interpretability needs to be thought out. It is crucial for a radiological diagnostic tool that is evaluating minors to be the most accurate as possible, especially considering the gravity of the outcomes for misclassification of a minor for an adult. Thus, the tradeoff favoured the performance.

One important consequence of the lack of interpretability in machine learning models in the health studies is that it hinders the possibility of advances in the medical field. Machine learning provides a great potential in uncovering new patterns in data and interactions between features, which could generate new and interesting hypotheses to be further investigated in future pure science works.

The importance of the findings in this research context regards that it is feasible to perform age assessment in youth and young adults without the use of ionising radiation, like in the traditional methods for BAA, with the employment of MRI. This finding addresses an important ethical issue of exposing minors to radiation without therapeutic purposes, which is also a gap in the research as identified by the SLR. The proposed automated approach also addresses the time spent on the assessment and inherent subjectivity of the traditional BAA.

9.2 Prognosis

In the work on the research context regarding prognosis, an SLR was first performed in order present the state of the art evidence on the studies on the prognosis of dementia with machine learning, then based on its results, an experiment was designed and carried out in order to investigate prognostic estimates for older individuals who come to develop, or not, dementia in a time frame of 10 years, using a decision tree approach. The proposed model achieved an AUC of 0.745 and recall

of 0.722. The points of novelty in this work include the 10 year interval for prediction and the broad multifactorial approach which considered 75 variables from multiple categories, including: demographic, social, lifestyle, medical history, biochemical tests, physical examination, psychological assessment and evaluation of multiple health instruments.

The main findings in this research context regard the identification of possible modifiable (i.e. physical strength, present smoking habit, BMI, diabetes type 2 and alcohol consumption) and non-modifiable (i.e. past smoking habit, number of regular medications, Backwards Digit Span test) risk factors, and interactions between them, which could suggest patient subgroups of importance in regards to the dementia risk. The latter being a factor that is not often considered in studies involving risk factors [103]. Possible subgroups were identified for ages younger and older than 75 years, with the physical strength factor being important across all ages.

One could argue about the achieved AUC of 0.745, and if the employment of more complex machine learning methods would provide better results. To address this issue, like in the previous research context, the trade-off between interpretability and performance was carefully considered. First of all, dementia is a disorder that results in very grave outcomes for the individuals, which gradually lose independence and functioning, presenting poorer and poorer health indicators with time. Second, there is no known cure or disease-modifying treatments that can act on the symptoms. The early prediction of dementia just for the sake of correctly identifying individuals at risk, without knowing why they are in this situation, is not the most useful in this current state of affairs. Adding up to this, there is also evidence that points out a risk of depression, suicide and requests for euthanasia in individuals who receive an early diagnosis of dementia [104]. This situation is discussed as a serious issue, since an individual is being diagnosed with a highly debilitating condition, without the possibility of treatment or even a procedure for the alleviation of the symptoms, could be argued to be against the medical ethical conduct. Given all the aforementioned facts, it was established that the main focus on investigating prognostic estimates of dementia would be the interpretability of the model, and the uncovering of possible modifiable factors that are able to be acted upon in efforts of delaying or preventing the dementia onset.

This research context managed to achieve reasonable results, considering the time for prediction, which are highly interpretable and provide ground for new hypotheses to be further investigated in trials in regards to modifiable risk factors and important patient subgroups. The importance of these results rely on the fact that they do not concern pharmacological efforts in regards to individuals which already present some degree of cognitive impairment, which are the focus of the research on the prognosis of dementia (evidenced by the SLR).

9.3 Machine learning in healthcare

In both research contexts, machine learning technology was employed in order to augment the human capacity in some way. In the diagnosis context, the process involving bone development is already known and understood by physicians, however, when technologies such deep learning are employed it can explore new features in the images without any constraint imposed by a manual method, leading to a better performance in the estimation of age. In the prognosis case, the factors involved in the process of dementia in the older population (and possible interplays between them) are unknown and the application of machine learning is done in order to uncover patterns in the data, considering a large number of factors that would be too impractical to be manually analyzed.

Both contexts highlight the importance of the application aspect in the various decision-making processes that will drive the choices of the technology that will be used in order to act on the problems at hand. The research contexts that are part of this thesis presented different priorities in terms of the necessity of a higher performance or a higher interpretability, which were given by the scenario and particularities of both cases. In the applied health technology field, the healthcare problem is both the beginning and the end for the employment of a certain technology.

Another important aspect of the application of machine learning in healthcare problems is the multidisciplinary aspect, in which health and information technology researchers interplay. Using machine learning to build and validate models is a straightforward process, however it is the domain experts are able to create and select high quality examples that will allow the employment of machine learning. Also, the domain expert needs to be in the loop in order to interpret, validate and identify possible points of improvement or failure.

One important methodological consideration to be made in regards to the use of machine learning in the healthcare domain is its robust exploratory features and could be considered an inductive type of research, but it differs from the '*pure induction*'. In the '*pure induction*' there is a hypothesis *h* and a body of evidence *e* and the problem is if *e* justifies *h*. However, in the machine learning's case the *e* is known, but there is no *h* to explain it. So, the objective is to use methods to obtain a suitable *h* from *e* [105]. Machine learning works by being a mechanical (or automated) way to discover new hypotheses in a way that "i*s not only taken as the inference from observations to given general rules. It includes the search for these rules in a large set of possibilities*" [106]. These new uncovered hypotheses need to be further investigated and validated by trials before any new application is translated to a clinical application.

Machine learning is a technology that can expand the human limitations in regards to tasks that are impractical or rather impossible to be manually performed. It offers great potential in the healthcare scenario by uncovering new hypotheses in complex scenarios like the ones that pertain to human health.

9.4 Threats to Validity

The threats to validity will be discussed in the following in regards to internal, external, construct and conclusion validities in the included studies in this thesis, along with actions taken in order to mitigate risks.

9.4.1 Internal validity

The internal validity refers to experimental designs and means that the observed outcomes are related to the studied interventions, and not other possible causes [107]. A possible internal validity threat that can be related to the experiments in this thesis is selection threat, which is the case in which the selection of the population can affect the outcomes [107]. The mitigation of this threat on the experiments in this thesis was done by recruiting subjects on a voluntary basis on studies II and III, also the data used by study V from the SNAC project consists of a randomized sample of the older population. On the experimental level it was done by the employment of cross-validation approaches.

Another source of threat to internal validity is the choice of machine learning methods that were employed to build the models in the experiments. To mitigate this threat, in both research contexts that pertain to this thesis, SLRs were performed, on beforehand, in order to identify the machine learning techniques being used in the research, so the choices were made based on previous published works.

9.4.2 Construct Validity

Construct validity refers to whether the measures that are being studied actually represent what the study intends to investigate [108]. This validity threat is mostly related to study II, where radiologist assessments and data from questionnaires, filled by the subjects, were the source of input data. To mitigate the first, the radiologists responsible for the assessment were specialized in pediatric radiology, were trained on the grading scale on beforehand, and a triangulation approach was used, in which more than one radiologist assessed the images of each subject. In regards to the questionnaire, it was validated in previous studies, in similar populations [109]. In study V, the data collected by the SNAC project employed the Good Clinical Practice

(GCP) standards for clinical studies and all of the examinations and tests were performed by trained physicians, nurses and psychologists.

9.4.3 External Validity

The external validity regards the extent to which the achieved results are generalizable [110]. For the SLRs of studies I and IV, the execution followed a systematic approach pre-defined by a protocol and an inclusive approach was used that even if all of the inclusion criteria were not present on the abstract of a paper, it would be included to be fully read and re-checked. Studies II and III built predictive models on data from Swedish healthy young subjects as specified in their respective methods, therefore results may not be generalizable to individuals who present health conditions that affect bone development, or originated from places with socioeconomic conditions that disparate from the Swedish one. Likewise, the estimation model produced in study V was built on data from the older population in Sweden, and may not generalize to other older populations with different socioeconomic conditions.

9.4.4 Conclusion Validity

Conclusion validity threats concern the risk of drawing inaccurate conclusions from the data [107]. In regards to this threat to validity, it is important to note that the experiments performed in the studies II, III and V, do not establish cause-effect conclusions. Especially in the field of health applications, it is important that the results are investigated in randomized clinical trials (RCT). The conclusion from the experiments presented in this thesis provides new hypotheses to be tested in RCTs, which is a very important step in the clinical translation process. Also, on an experimental design level, the employment of mechanisms to avoid overfitting (approaches in which data used for model optimization is not used to assess performance) mitigated the chance of producing erroneous models. In study II, statistical testing was not used in order to choose between machine learning models, which could be considered a threat to conclusion validity, however study II did not produce good results for the age estimation in the explored techniques due to the data used, the radiologists assessment, not being precise enough for the classification (one of the main conclusions of the diagnosis research context).

10 CONCLUSION

This thesis comprised five research papers that aimed at investigating the application of machine learning in healthcare. This application considers machine learning as an applied health technology, which is driven to the solution of specific practical problems. Two research contexts were considered in this thesis: (i) diagnosis, with the problem of age estimation of youth and young adult individuals; and (ii) prognosis, with the problem of the prognosis of dementia. The development of solutions in these contexts was dependent on the collaboration of a multidisciplinary team of experts and presented different priorities in regards to interpretability and performance.

In the research context of diagnosis, studies I, II and III investigated age assessment methods, using MRI and addressing the drawbacks in the traditional methods of BAA. Study I was an SLR that provided the state of the art evidence of BAA methods with the use of machine learning, which indicated that the research investigated automated and non-automated methods in regards to the human input to the system, most of the methods made use of radiographs of the hand area and non-radiological data was scarcely explored. Study II investigated a non-automated approach for the assessment of age, which was based on the radiologists' assessment and non-radiological data. The age assessment models presented limited prediction power for ages older than 15, achieving MAEs of 0.95 years for male and 1.24 years for female subjects on the age estimation experiment and accuracies of 90% for male and 84% for female subjects on the classification of minors and adults. Study III proposed an automated method for age assessment, based on deep learning, achieving MAEs of 0.79 years for male (age range 14-20 years) and 0.99 years (age range 14-19 years) for female subjects on the age estimation experiment, and accuracies of 98% for male and 95% for female subjects on the classification of minors and adults. It was concluded that the radiologist's assessment was not precise enough for the age estimation and that the deep learning method was able to extract more features from the MRIs, providing better performance.

In the research context of prognosis, studies IV and V investigated prognostic estimates for the dementia disorder. Study IV summarized the relevant evidence in this through conducting an SLR on the prognosis of dementia with machine learning techniques. The results showed that the focus of the research lies on models for the prediction of MCI patients who developed Alzheimer's disease, usually in 36 months, through the analysis of neuroimaging. Study V proposed a multifactorial decision tree approach for the prognosis of dementia in older individuals as to their development

or not of dementia in 10 years. The proposed method achieved an AUC of 0.75 and recall of 0.72, identifying possible modifiable and non-modifiable risk factors, and possible patient subgroups of importance.

11 FUTURE WORK

Proposition for future works should revolve around clinical studies in order to validate and investigate further the results obtained by the estimation models built on both research contexts in clinical trials, which constitutes the next step in the process of transportability of findings to the clinical practice.

In the research context of diagnosis, future efforts could be directed to explore the explainable artificial intelligence methods in order to *'open'* the deep learning black box, and have its results evaluated in conjunction with radiologists in a quantitative-qualitative mixed methods approach.

In the research context regarding prognosis, it would be interesting to perform validation studies on the other sites of the SNAC project. The model presented in study V was built on the data from the Blekinge site.

12 REFERENCES

1. Nass SJ, Levit LA, Gostin LO. Beyond the HIPAA Privacy Rule: Enhancing Privacy, Improving Health Through Research. Washington (DC): National Academies Press (US); 2010.

2. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. Journal Of Big Data. 2014. p. 2. doi:10.1186/2196-1115-1-2

3. Crown WH. Potential application of machine learning in health outcomes research and some statistical cautions. Value Health. 2015;18: 137–140.

4. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13: 8–17.

5. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16: 321–332.

6. Kavakiotis I, Tsave O, Salifoglou A, Maglaveras N, Vlahavas I, Chouvarda I. Machine Learning and Data Mining Methods in Diabetes Research. Comput Struct Biotechnol J. 2017;15: 104–116.

7. Randhawa GS, Soltysiak MPM, El Roz H, de Souza CPE, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS One. 2020;15: e0232391.

8. Ge Y, Tian T, Huang S, Wan F, Li J, Li S, et al. A data-driven drug repositioning framework discovered a potential therapeutic agent targeting COVID-19. bioRxiv. 2020. Available: https://www.biorxiv.org/content/10.1101/2020.03.11.986836v1.abstract

9. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577: 706–710.

10. Alimadadi A, Aryal S, Manandhar I, Munroe PB, Joe B, Cheng X. Artificial intelligence and machine learning to fight COVID-19. Physiol Genomics. 2020;52: 200–202.

11. International Network of Agencies for Health Technology Assessment. Health Technology Assessment (HTA) Glossary. HTA Glossary. 2006. Available: http://htaglossary.net/HomePage

12. Feibleman JK. Pure Science, Applied Science, Technology, Engineering: An Attempt at Definitions. Technol Cult. 1961;2: 305–317.

13. Health Technology Research Lab. Applied health technology at BTH. 2017. Available: http://healthtechnology.se/research/

14. Tell J. Implementation and use of Web-based National Guidelines in Child Healthcare. Blekinge Tekniska Högskola. 2019. Available: https://www.diva-portal.org/smash/record.jsf?pid=diva2:1265310

15. Olander E, Nilsson L. Applied Health Technology – a New Research Discipline at Blekinge Institute of Technology. 2009. Available: https://www.diva-portal.org/smash/get/diva2:833695/FULLTEXT01.pdf

16. Lélé S, Norgaard RB. Practicing Interdisciplinarity. Bioscience. 2005;55: 967–975.

17. Hofflander M. Implementing video conferencing in discharge planning sessions : leadership and organizational culture when designing IT support for everyday work in nursing practice. Blekinge Tekniska Högskola. 2015. Available: http://www.diva-portal.org/smash/record.jsf?pid=diva2:817200

18. Anderberg P. Gerontechnology, digitalization, and the silver economy. XRDS. 2020;26: 46–49.

19. Panch T, Szolovits P, Atun R. Artificial intelligence, machine learning and health systems. J Glob Health. 2018;8: 020303.

20. Lee A, Taylor P, Kalpathy-Cramer J, Tufail A. Machine Learning Has Arrived! Ophthalmology. 2017;124: 1726–1728.

21. Samuel AL. Some Studies in Machine Learning Using the Game of Checkers. IBM J Res Dev. 1959;3: 210–229.

22. IBM100 - Deep Blue. In: IBM Corporation [Internet]. 7 Mar 2012 [cited 15 Apr 2020]. Available: https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/

23. Mitchell TM. The discipline of machine learning. School of Computer Science. Carnegie Mellon University, Pittsburg, PA; 2006.

24. De La Paz S. Composing via Dictation and Speech Recognition Systems: Compensatory Technology for Students with Learning Disabilities. Learn Disabil Q. 1999;22: 173–182.

25. Boubela RN, Kalcher K, Huf W, Našel C, Moser E. Big Data Approaches for the Analysis of Large-Scale fMRI Data Using Apache Spark and GPU

Processing: A Demonstration on Resting-State fMRI Data from the Human Connectome Project. Front Neurosci. 2015;9: 492.

26. Kotsiantis SB, Zaharakis I, Pintelas P. Supervised machine learning: A review of classification techniques. Emerging artificial intelligence applications in computer engineering. 2007;160: 3–24.

27. Nilsson Nils J. Introduction To Machine learning. Robotics Laboratory Department of Computer Science Stanford University. 1998;1.

28. Louridas P, Ebert C. Machine Learning. IEEE Software. 2016. pp. 110–115. doi:10.1109/ms.2016.114

29. Kuhn M, Johnson K. Applied Predictive Modeling. Springer, New York, NY; 2013.

30. Google. Machine Learning Glossary. 2020. Available: https://developers.google.com/machine-learning/glossary

31. Deng L, Yu D. Deep Learning: Methods and Applications. Found Signal Process Commun Netw. 2014;7: 197–387.

32. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2018;19: 1236–1246.

33. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 2017;21: 4–21.

34. Rowe M. An Introduction to Machine Learning for Clinicians. Acad Med. 2019;94: 1433–1436.

35. World Health Organization. Global spending on health: a world in transition. 2019. Available: https://www.who.int/health_financing/documents/health-expenditure-report-2019/en/

36. Dieleman JL, Squires E, Bui AL, Campbell M, Chapin A, Hamavid H, et al. Factors Associated With Increases in US Health Care Spending, 1996-2013. JAMA. 2017;318: 1668–1678.

37. Bhardwaj R, Nambiar AR, Dutta D. A Study of Machine Learning in Healthcare. 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC). 2017. pp. 236–241.

38. World Health Organization. Big Data and Artificial Intelligence for Achieving Universal Health Coverage: An International Consultation on Ethics. 2018. Available: https://www.who.int/ethics/publications/big-data-artificial-intelligence-report/en/

39. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 2019;380: 1347–1358.

40. Holmboe ES, Durning SJ. Assessing clinical reasoning: moving from in vitro to in vivo. Diagnosis (Berl). 2014;1: 111–117.

41. Ledley RS, Lusted LB. Reasoning foundations of medical diagnosis; symbolic logic, probability, and value theory aid our understanding of how physicians reason. Science. 1959;130: 9–21.

42. Jutel A. Sociology of Diagnosis: A Preliminary Review. Advances in Medical Sociology. 2011. pp. 3–32. doi:10.1108/s1057-6290(2011)0000012006

43. Li M, Zhou Z. Improve Computer-Aided Diagnosis With Machine Learning Techniques Using Undiagnosed Samples. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans. 2007;37: 1088–1098.

44. Suzuki K. A review of computer-aided diagnosis in thoracic and colonic imaging. Quant Imaging Med Surg. 2012;2: 163–176.

45. Suzuki K. Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and analysis. Suzuki K, editor. Medical Information Science reference (an inprint of IGI Global); 2012.

46. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009;338: b375.

47. Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform. 2001;34: 428–439.

48. United Nations Children's Fund. Age assessment practices: a literature review & annotated bibliography. 2011. Available: https://www.unicef.org/protection/files/Age_Assessment_Practices_2010.pdf

49. United Nations Children's Fund. Convention on the Rights of the Child. 1989. Available: https://www.unicef.org/child-rights-convention/convention-text

50. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, et al. The problem of aging human remains and living individuals: a review. Forensic Sci Int. 2009;193: 1–13.

51. Schmidt S, Vieth V, Timme M, Dvorak J, Schmeling A. Examination of ossification of the distal radial epiphysis using magnetic resonance imaging. New insights for age estimation in young footballers in FIFA tournaments. Sci Justice. 2015;55: 139–144.

52. Ekizoglu O, Hocaoglu E, Inci E, Can IO, Aksoy S, Kazimoglu C. Forensic age

estimation via 3-T magnetic resonance imaging of ossification of the proximal tibial and distal femoral epiphyses: Use of a T2-weighted fast spin-echo technique. Forensic Sci Int. 2016;260: 102.e1–102.e7.

53. Gilsanz V, Ratib O. Hand Bone Age: A Digital Atlas Of Skeletal Maturity. Springer Science & Business Media; 2005.

54. Greulich W, Pyle SL. Radiographic atlas of skeletal development of the hand and wrist. Am J Med Sci. 1959;238.

55. Tanner JM, Whitehouse RH, Cameron N. Assessment of skeletal maturity and prediction of adult height (TW3 Method). 3rd edn. WB Saunders: London; 2001.

56. Satoh M. Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol. 2015;24: 143–152.

57. Creo AL, Schwenk WF 2nd. Bone Age: A Handy Tool for Pediatric Providers. Pediatrics. 2017;140. doi:10.1542/peds.2017-1486

58. Mansourvar M, Ismail MA, Herawan T, Raj RG, Kareem SA, Nasaruddin FH. Automated bone age assessment: motivation, taxonomies, and challenges. Comput Math Methods Med. 2013;2013: 391626.

59. Hjern A, Brendler-Lindqvist M, Norredam M. Age assessment of young asylum seekers. Acta Paediatr. 2012;101: 4–7.

60. Marshall WA, Tanner JM. Variations in pattern of pubertal changes in girls. Arch Dis Child. 1969;44: 291–303.

61. Marshall WA, Tanner JM. Variations in the pattern of pubertal changes in boys. Arch Dis Child. 1970;45: 13–23.

62. Cox LA. The biology of bone maturation and ageing. Acta Paediatr Suppl. 1997;423: 107–108.

63. Henderson AS. The coming epidemic of dementia. Aust N Z J Psychiatry. 1983;17: 117–127.

64. World Health Organization. Dementia: A public health priority. 2012. Available: https://www.who.int/mental_health/publications/dementia_report_2012/en/

65. Winblad B, Amouyel P, Andrieu S, Ballard C, Brayne C, Brodaty H, et al. Defeating Alzheimer's disease and other dementias: a priority for European science and society. Lancet Neurol. 2016;15: 455–532.

66. Furiak NM, Klein RW, Kahle-Wrobleski K, Siemers ER, Sarpong E, Klein TM. Modeling screening, prevention, and delaying of Alzheimer's disease: an early-

stage decision analytic model. BMC Med Inform Decis Mak. 2010;10: 24.

67. Jennings LA, Reuben DB, Evertson LC, Serrano KS, Ercoli L, Grill J, et al. Unmet needs of caregivers of individuals referred to a dementia care program. J Am Geriatr Soc. 2015;63: 282–289.

68. Burns A. The burden of Alzheimer's disease. Int J Neuropsychopharmacol. 2000;3: 31–38.

69. World Health Organization. Global action plan on the public health response to dementia 2017–2025. 2017. Available: https://www.who.int/mental_health/neurology/dementia/action_plan_2017_2025/en/

70. Kitchenham B, Charters S. Guidelines for Performing Systematic Literature Reviews in Software Engineering, Keele University. UK EBSE-2007-12007. 2007.

71. Larsson B. Swedish Age Assessment Study (SAAS). 2017. Available: https://clinicaltrials.gov/ct2/show/NCT03242811

72. Dedouit F, Auriol J, Rousseau H, Rougé D, Crubézy E, Telmon N. Age assessment by magnetic resonance imaging of the knee: a preliminary study. Forensic Sci Int. 2012;217: 232.e1–7.

73. Kellinghaus M, Schulz R, Vieth V, Schmidt S, Pfeiffer H, Schmeling A. Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int J Legal Med. 2010;124: 321–325.

74. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960;20: 37–46.

75. Gaudette L, Japkowicz N. Evaluation Methods for Ordinal Classification. Advances in Artificial Intelligence. Springer Berlin Heidelberg; 2009. pp. 207–210.

76. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45: 427–437.

77. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. pp. 1–9.

78. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 770–778.

79. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 2818–2826.

80. Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv [cs.CV]. 2014. Available: http://arxiv.org/abs/1409.1556

81. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012. pp. 1097–1105.

82. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp. 4700–4708.

83. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Springer International Publishing; 2015. pp. 234–241.

84. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. Caffe: Convolutional Architecture for Fast Feature Embedding. Proceedings of the 22Nd ACM International Conference on Multimedia. New York, NY, USA: ACM; 2014. pp. 675–678.

85. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv [cs.LG]. 2014. Available: http://arxiv.org/abs/1412.6980

86. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. pp. 248–255.

87. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33: 159–174.

88. Lagergren M, Fratiglioni L, Hallberg IR, Berglund J, Elmståhl S, Hagberg B, et al. A longitudinal study integrating population, care and social services data. The Swedish National study on Aging and Care (SNAC). Aging Clin Exp Res. 2004;16: 158–168.

89. Antonovsky A. The structure and properties of the sense of coherence scale. Soc Sci Med. 1993;36: 725–733.

90. Wechsler Adult Intelligence Scale (All Versions). SpringerReference. Berlin/Heidelberg: Springer-Verlag; 2011.

91. Livingston G, Blizard B, Mann A. Does sleep disturbance predict depression in elderly people? A study in inner London. Br J Gen Pract. 1993;43: 445–448.

92. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37: 53–72.

93. Katz S. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. J Am Geriatr Soc. 1983;31: 721–727.

94. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist. 1969;9: 179–186.

95. Folstein MF, Folstein SE, McHugh PR. Mini-mental state': A practical method for grading the cognitive state of patients for the clinician. J Psychiatry Res 1975; 12: 189--198. External Resources Pubmed/Medline (NLM) CrossRef (DOI) Chemical Abstracts Service (CAS). 1962.

96. Agrell B, Dehlin O. The clock-drawing test. Age Ageing. 1998;27: 399–403.

97. Jenkinson C, Layte R. Development and testing of the UK SF-12. J Health Serv Res Policy. 1997;2: 14–18.

98. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134: 382–389.

99. A. N, Kudus A. Decision Tree for Prognostic Classification of Multivariate Survival Data and Competing Risks. Recent Advances in Technologies. 2009. doi:10.5772/7429

100. Chen C, Liaw A, Breiman L, Others. Using random forest to learn imbalanced data. University of California, Berkeley. 2004;110: 24.

101. World Health Organization, Council for International Organizations of Medical Sciences. International ethical guidelines for health-related research involving humans. Geneva: Council for International Organizations of Medical Sciences; 2016. Available: https://cioms.ch/wp-content/uploads/2017/01/WEB-CIOMS-EthicalGuidelines.pdf

102. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R. Explainable AI: Interpreting, Explaining and Visualizing Deep Learning. Springer Nature; 2019.

103. Tang EYH, Harrison SL, Errington L, Gordon MF, Visser PJ, Novak G, et al. Current Developments in Dementia Risk Prediction Modelling: An Updated Systematic Review. PLoS One. 2015;10: e0136181.

104. Draper B, Peisah C, Snowdon J, Brodaty H. Early dementia diagnosis and the risk of suicide and euthanasia. Alzheimers Dement. 2010;6: 75–82.

105. Gillies D. Artificial Intelligence and Scientific Method. OUP Oxford; 1996.

106. Bergadano F, Gunetti D. Inductive Logic Programming: From Machine Learning to Software Engineering. MIT Press; 1996.

107. Robson C, McCartan K. Real World Research. John Wiley & Sons; 2016.

108. Runeson P, Host M, Rainer A, Regnell B. Case Study Research in Software Engineering: Guidelines and Examples. John Wiley & Sons; 2012.

109. Andö P. Living Conditions Surveys (ULF/SILC). 2013. Available: https://www.scb.se/contentassets/b7fedb7ce4434d90a525891b7f08d9d6/le0101_bs_2012_en.pdf

110. Zhou X, Jin Y, Zhang H, Li S, Huang X. A Map of Threats to Validity of Systematic Literature Reviews in Software Engineering. 2016 23rd Asia-Pacific Software Engineering Conference (APSEC). 2016. doi:10.1109/apsec.2016.031

Study I

13 Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis

PLOS ONE

RESEARCH ARTICLE

Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis

Ana Luiza Dallora[1]*, Peter Anderberg[1‡], Ola Kvist[2☯], Emilia Mendes[3‡], Sandra Diaz Ruiz[2☯], Johan Sanmartin Berglund[1‡]

1 Department of Health, Blekinge Institute of Technology, Karlskrona, Sweden, **2** Department of Pediatric Radiology, Karolinska University Hospital, Stockholm, Sweden, **3** Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work
* ana.luiza.moraes@bth.se

Citation: Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J (2019) Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PLoS ONE 14(7): e0220242. https://doi.org/10.1371/journal.pone.0220242

Editor: Ruxandra Stoean, University of Craiova, ROMANIA

Received: April 2, 2019

Accepted: July 11, 2019

Published: July 25, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Background

The assessment of bone age and skeletal maturity and its comparison to chronological age is an important task in the medical environment for the diagnosis of pediatric endocrinology, orthodontics and orthopedic disorders, and legal environment in what concerns if an individual is a minor or not when there is a lack of documents. Being a time-consuming activity that can be prone to inter- and intra-rater variability, the use of methods which can automate it, like Machine Learning techniques, is of value.

Objective

The goal of this paper is to present the state of the art evidence, trends and gaps in the research related to bone age assessment studies that make use of Machine Learning techniques.

Method

A systematic literature review was carried out, starting with the writing of the protocol, followed by searches on three databases: Pubmed, Scopus and Web of Science to identify the relevant evidence related to bone age assessment using Machine Learning techniques. One round of backward snowballing was performed to find additional studies. A quality assessment was performed on the selected studies to check for bias and low quality studies, which were removed. Data was extracted from the included studies to build summary tables. Lastly, a meta-analysis was performed on the performances of the selected studies.

Results

26 studies constituted the final set of included studies. Most of them proposed automatic systems for bone age assessment and investigated methods for bone age assessment

based on hand and wrist radiographs. The samples used in the studies were mostly comprehensive or bordered the age of 18, and the data origin was in most of cases from United States and West Europe. Few studies explored ethnic differences.

Conclusions

There is a clear focus of the research on bone age assessment methods based on radiographs whilst other types of medical imaging without radiation exposure (e.g. magnetic resonance imaging) are not much explored in the literature. Also, socioeconomic and other aspects that could influence in bone age were not addressed in the literature. Finally, studies that make use of more than one region of interest for bone age assessment are scarce.

Introduction

Motivation

Bone development is the process that drives changes in bones' size, shape and degree of mineralization. This happens in primary and secondary ossification centers, namely diaphysis and epiphysis, respectively, where cartilage gradually turns into bone tissue. This process persists as long as cartilage remains in the growth plate (or epiphyseal plate). At the end of the bone development process the epiphyseal plate is ossified, indicating that the diaphysis and epiphysis are fused [1].

Other important concepts that relate to such topic are skeletal maturity, bone age and chronological age. Skeletal maturity refers to the stage of development in which the bone is currently in [1]. Bone age is a closely related concept to skeletal maturity, and relates to the estimation of an age based on an individual's skeletal maturity [2], whereas the chronological age is calculated based solely on an individual's date of birth.

The importance of assessing an individual's skeletal maturity or bone age and its comparison with their chronological age arises from two main angles: First, from a medical angle, the assessment of bone age is useful for the diagnosis and treatment of pediatric endocrinology, orthodontics and pediatric orthopedic disorders, in addition to also being considered in estimations of an individual's final height [3]. Second, from a legal standpoint, bone age assessment is important in order to identify whether an individual is a minor in a situation where verified documents are lacking.

This latter aspect is extremely important given the growth in number of asylum seekers in Europe, where unaccompanied individuals under the age of 18 are granted special rights by the United Nations (UN) Convention on the Rights of the Child [4], which relate to reception arrangements, access to health care, education etc [5].

A report from the United Nations Refugee Agency (UNHCR) stated that individuals under the age of 18 contributed to 52% of the refugees population in 2017; and also in 2017 around 173,800 refugees children were unaccompanied or separated from their parents, in what the UN considers a "conservative estimate" [6]. Given such circumstances, the assessment of bone age and skeletal maturity constitute an important research topic nowadays.

Current scenario of bone age assessment

Currently, the two most commonly used procedures for bone age assessment are the Greulich-Pyle (GP) and Tanner-Whitehouse (TW) methods [1]. Both are based on hand radiographs

via the analysis of the epiphyses' and diaphyses' appearance. The GP method [7] is an atlas that contains reference images from hand radiographs collected from 1931 to 1942, from upper-middle class Caucasian children in Ohio, United States [1]. The attribution of bone age is done by comparing an individual's hand radiograph with the reference images in the atlas [3]. The TW method [8] evaluates maturity scores for the radius, ulna, carpals and 13 hand short bones. Some of these bones are evaluated and categorized into stages ranging from A to I, then a total score is calculated, which is later converted into the bone's age. The TW method was developed with data collected from 1950 to 1960, from children of average socioeconomic class in the United Kingdom. This method was further updated in 2001 based on new consistent patterns of development (secular trends) [3].

Besides their popularity, both GP and TW methods are criticized for many reasons. In a practical sense, since they are manually done by radiologists, the whole process can be time-consuming, which is aggravated by the increased demand for this activity due to the increase of individuals seeking refuge. Further, they can be prone to inter- and intra-rater variability, which raises ethical and legal issues, especially when considering that these assessments are done in relation to an individual being a minor or not [9].

In light of this scenario, a way to tackle these problems is to make use of automated methods, and a technology that is valuable in this matter is Machine Learning. This technology is already employed in many areas of medical research, ranging from genomics to diagnosis and prognosis of many disorders, aiming to find patterns in data and to provide useful estimates to improve decision making in the health area [10]. Machine Learning comprises a group of technologies that operates in the following way: firstly, the technique learns from a set of examples on how to perform a task, creating a model which encapsulates the knowledge to perform the task. Then, when new data is imputed, the model is able to correctly perform the learned task within an acceptable accuracy [11].

Systematic review of the literature on bone age assessment

Given the motivations and scenario abovementioned, this paper contributes to the literature on age assessment as it describes a systematic literature review (SLR) that presents the state of the art on the use of Machine Learning techniques in the context of bone age and skeletal maturity assessment, a theme not previously addressed by SLRs. Hereby answering the research question: "How Machine Learning techniques are being employed in studies concerning youth age assessment (10 to 30 years)?".

This paper is organized as follows: the Method section details the approach used to conduct the review, which followed the recommendations of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) Statement [12]; the Results section aggregates and synthetizes the data from the studies included in the review; the Discussion section argues about the results and provides considerations related to threats to validity; and finally the Conclusion section presents the final statements and comments on future work.

Materials and methods

A systematic literature review (SLR) is an approach in which a significant and representative sample of the literature regarding a certain topic is identified, evaluated and interpreted. This is done by executing a comprehensive search following a pre-defined method that specifies the research questions the SLR aims to answer, the criteria used to include and exclude studies, how to assess their quality and how to extract and make the synthesis of the data [13–14]. The common motivations for executing an SLR are: to summarize next to all the evidence about a

topic of interest; to find research gaps; to provide a grounding for new research; and to investigate how the research that is currently being done supports a certain hypothesis [13].

The SLR presented herein was conducted by four participants with different expertise (health, machine learning and health informatics). Throughout the text, references to the authors will use a notation, in which A1 refers to the first author; A2 refers to the second author, and so forth.

The main question this SLR aims to answer is: "How machine learning techniques are being employed in studies concerning youth age assessment (10 to 30 years)?". This main question is further decomposed into the following five research questions.

- RQ1: Which machine learning techniques are being used in the age assessment studies?
- RQ2: What data characteristics (database's origin, data collection mechanism, and ages) are being considered in the age assessment studies?
- RQ3: What type of medical imaging are being used in the studies?
- RQ4: What are the regions of interest being explored for the age assessment studies (hand, wrist, knee, etc) and what were the methods used to assess them?
- RQ5: What are the performances of the proposed methods?

For the RQ5, besides the SLR steps to summarize the information regarding the performances of the studies, a meta-analysis was also conducted. A meta-analysis is the application of statistical operations in order to synthetize the results of individual studies, so they can be understood in the context of all selected studies [15].

The protocol that guided this SLR can be found at: http://tiny.cc/4wuw8y.

Search strategy

The search strategy used to find primary studies employed a search string built based upon the PICO framework, in which the main question is re-written in terms of four elements: Population, Intervention, Comparison and Outcome [14]. The "comparison" component was not used because the goal of SLR detailed herein was to characterize existing evidence. The components used for the automated searches are defined below.

- Population: Studies involving age assessment in youth.
- Intervention: Use of medical imaging.
- Outcome: Machine learning models for age assessment.

The resulting search string used to conduct the automated searches is shown in Table 1. The searches were performed in Pubmed, Scopus and Web of Science databases (with the necessary adaptations).

Study selection

Two sets of searches were executed in this SLR. The first was carried out on March 21st 2018, and the second on February 6th 2019, aiming to search for additional *evid*ence since the first search was conducted. The procedure for each search is detailed next.

After the removal of duplicates, the first search screened 148 studies, assessing titles and abstracts, guided by the inclusion and exclusion criteria (see Table 2). Four participants took part in this procedure, with A1 evaluating all of the papers and A2, A4 and A6 one third each. The Cohen's Kappa was calculated as a measure of consensus between A1 and the other

Table 1. Search String used in the Pubmed database.

Search Dates	21/03/2018 and 06/02/2019
(((“age assessment” OR “age appraisal” OR “age diagnostics” OR “age estimate” OR “age estimation” OR “age determination” OR “age prediction” OR “age testing”) AND (“bone age measurement” OR “bone age assessment” OR “bone maturity” OR “bone development” OR “bone age testing” OR “bone age tests” OR “skeletal maturity” “skeletal maturation” OR “bone examination” OR “skeletal development” OR “developmental assessment” OR “bone age” OR “skeletal age” OR “growth zone”)) AND (“Magnetic Resonance imaging” OR “MRI” OR “x ray” OR “x-ray” OR “xray” OR “Radiography” OR “computed tomography” OR “CT” OR “ultrasound” OR “ultrasonography” OR “medical imaging”) AND (“machine learning” OR “unsupervised Machine Learning” OR “supervised Machine Learning” OR”Classification” OR”Regression” OR”Kernel” OR”Support vector machines” OR”Gaussian process” OR”Neural networks” OR”Logical learning” OR”relational learning” OR”Inductive logic” OR”Statistical relational” OR”probabilistic graphical model” OR”Maximum likelihood” OR”Maximum entropy” OR”Maximum a posteriori” OR”Mixture model” OR”Latent variable model” OR”Bayesian network” OR”linear model” OR”Perceptron algorithm” OR”Factorization” OR”Factor analysis” OR”Principal component analysis” OR”Canonical correlation” OR”Latent Dirichlet allocation” OR”Rule learning” OR”Instance-based” OR”Markov” OR”Stochastic game” OR”Learning latent representation” OR”Deep belief network” OR”Bio-inspired approach” OR”Artificial life” OR”Evolvable hardware” OR”Genetic algorithm” OR”Genetic programming” OR”Evolutionary robotic” OR”Generative and developmental approaches”))	

https://doi.org/10.1371/journal.pone.0220242.t001

authors, resulting in strong agreements with all of them (0.73 with A4, 0.76 with A2 and 0.70 with A6). A total of 65 studies were selected to be fully read and assessed for eligibility.

One round of backward snowballing was performed on the 65 selected studies' references, searching for additional evidence, and following an analogous process to the previous assessment. After screening 2124 items, 2 additional studies were selected. This step was performed by A1, and the remaining participants were consulted when necessary, based on their expertise.

In total, 67 studies were selected for eligibility assessment. These were fully read and re-assessed as per the criteria shown in Table 2. A quality assessment was also performed to minimize the chance of selecting studies with bias evidence. The quality questionnaire was adapted from Kitchenham's guidelines [13] for performing SLRs, and is detailed in the SLR protocol. If a paper fell below the threshold of 8 points (out of 13) it would be removed from SLR due to quality reasons. The threshold was defined in meeting with all of the participants. A considerable amount of papers was removed from the SLR in this phase, because during the selection based on title and abstracts many abstracts did not contain all the information necessary to judge the inclusion and exclusion criteria, so they were included to be re-assessed when fully

Table 2. Inclusion and exclusion criteria for assessing the retrieved papers.

Inclusion Criteria	Exclusion Criteria:
• Be a primary study in English; AND • Have been published in the last 10 years; AND • Address the research on age assessment using medical imaging; AND • Be a study concerning the building of models for the purpose of age assessment using at least one machine learning technique; AND • Be a study in which the age assessment is analyzed through growth zones in joints; AND • Be a study regarding age assessment in living individuals.	• Be a secondary or tertiary study; OR • Have been published before 2007; OR • Be written on another language other than English; OR • Do not address research on age assessment using medical imaging; OR • Do not address the building of models for the purpose of age assessment; OR • Be a study concerning height prediction, or a specific syndrome/disease that affects normal growth; OR • Be a study in which the age assessment is not analyzed through growth zones in joints; OR • Be a study in which the study population is out of the range of 10 to 30 years old by 3 years or less; OR • Be a study in which the study population is post-mortem material.

https://doi.org/10.1371/journal.pone.0220242.t002

read. The responsible for fully reading the papers was A1, but A2, A4 and A6 were also consulted when necessary. A total of 22 studies were selected by the first search.

On the second search, 29 studies were screened analogously to the first search, resulting in 19 studies to be fully read and assessed for eligibility. The same quality assessment questionnaire was applied to the studies afterwards. A1 performed the selection process of the second search and A2, A4 and A6 were consulted when pertinent. In total, 4 studies were selected on the second search.

All of the selected studies were also assessed for the risk of cumulative evidence bias. This means that studies from the same group or same age assessment system were checked. Validation studies of a system that was already included in the set of selected studies were considered duplicates and not included in the final set, but will be further referenced when necessary. The final set of selected studies consisted of 26 papers.

Data collection

Table 3 lists the data extracted from the studies. Besides these, other basic information was also extracted (i.e. authors, journal/source, year and type of publication).

Data analysis

From the data collected, summary tables were built to summarize the results for the SLR and to answer the research questions. The meta-analysis of BAA performances was done through the average of the performances weighted by the sample sizes. The software used to perform the data analyses was Excel.

Results

The PRISMA flow chart that describes the process of selection of the articles that were included in this SLR is shown in Fig 1.

The study selection resulted in the assessment of eligibility of a total of 86 studies (67 from the first search and 19 from the second), from which 26 were included in the final set (22 from the first search and 4 from the second). The most common reasons for ineligibility were: the case where the age assessment method was not performed on growth zones in joints (e.g. dental assessment), totaling 30 studies; cases where no ML technique was employed, totaling 7 studies; cases where the study population was mostly out of the range of 10 to 30 years old; and

Table 3. Data extracted from the selected studies.

Variables	Definition
Aim	The goal of the built model or the proposed study.
Age range of the subjects	The age range which the model for age assessment is concerned with.
Origin of the subjects	Characteristics related to the country/ethnicity of the subjects which the model is built upon.
Type of Image	Radiography, Magnetic Resonance Imaging (MRI), Ultrasound etc.
Regions of Interest for the images	Body part which the model analyses for the age assessment purposes.
Model Building Technique	The ML model building technique used to build an age assessment method.
Method used for age assessment	Method for age assessment that the model was built upon, if any (i.e. TW, GP).
Dataset size	The sample size utilized by the study.
Performance	Performance achieved by the best proposed model for bone age assessment.

https://doi.org/10.1371/journal.pone.0220242.t003

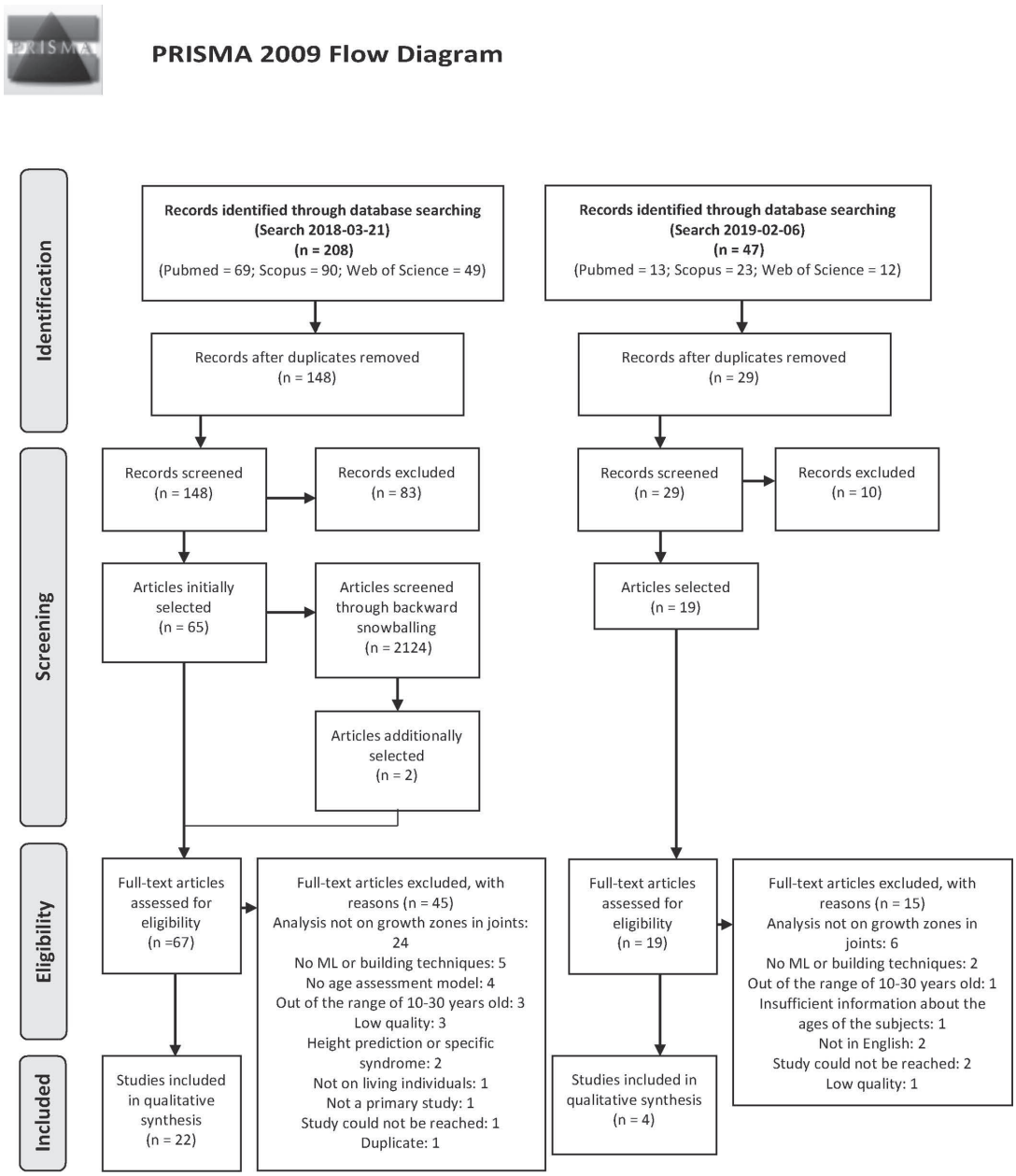


Fig 1. PRISMA flow chart.

https://doi.org/10.1371/journal.pone.0220242.g001

cases where no age assessment model was proposed. The complete list of reasons for exclusions with its respective numbers is shown on Fig 1.

In the list of included papers, two studies were found to be from the same research group [16–17], and used the same data e related techniques. After being assessed for cumulative bias, it was decided that only the most recent [16] would be kept.

The complete list of included studies is presented in S1 Table, as supporting information (see S2 Table for the filled PRISMA checklist).

Overview of the included Studies

Table 4 shows the summarized data arranged by the documented aims of the included studies. Eleven studies aimed to propose a new approach for an automatic Bone Age Assessment (BAA). Note that "automatic BAA system" means a fully automated approach without human intervention. These systems use a medical image as input and automatically detect the relevant region of interest (ROI), followed by the bone assessment age through an algorithm.

Studies that aimed to propose a "non-automatic BAA system" have some degree of automation by the use of Machine Learning techniques, but are still dependent of some kind of human input. Human interventions in these systems are usually related to the manual description of the information contained in the medical image to be inputted in the system. The manual location of regions of interest (e.g. epiphyses and metaphyses) and the assessment of the image by radiologists as to specific stages of ossification or tissue analyses are examples of interventions performed by humans in these systems. The "non-automatic BAA systems" were present in 7 studies in the SLR.

Three studies, besides proposing a new BAA approach, compared its results to the traditional TW method, obtaining better results in terms of performance or time spent. The study by Fan et al. [37] proposed models to estimate age from the ossification stages of the knee using MRI and Radiograph images, yielding better results for the MRI images.

In addition to these, there were other studies that did not propose either automatic or non-automatic BAA, as follows: the study by Soegiharto et al. [38] compare the skeletal maturation

Table 4. Aims of the studies included in the SLR.

Aim	Number of Studies	Featured Studies
Proposed an automatic BAA system.	10	[16, 18–26]
Proposes a non-automatic BAA system.	7	[27–33]
Proposes a non-automatic BAA system. Also, test how reliable is the TW method in the western Australian population.	1	[34]
Proposes a non-automatic BAA system and compares to TW3	1	[35]
Proposed an automatic BAA system and compared with the manual BAA	1	[36]
Comparison between age assessment with MRI and Radiograph	1	[37]
Predicts the Skeletal Maturity Index (by Fisherman) from the chronologic age and compares the results of the American sample and Indonesian sample	1	[38]
Proposes a simplified version of the TW3 method	1	[39]
Examines sex-specific differences in the maturation time of the bones in the hand and wrist in two cohorts of children subjects to investigate secular trends.	1	[40]
Investigated the effect of the African-American ancestry, linear growth, body composition, and pubertal maturation on the skeletal maturation in a cohort of non-obese children and adolescents.	1	[41]
Investigates the persistence of the epiphyseal scar of the distal radius and its relationship with chronological age and sex.	1	[42]

https://doi.org/10.1371/journal.pone.0220242.t004

index and the cervical vertebrae maturation in two cohorts, one composed of Indonesian subjects and other by Caucasian subjects. For both methods, the Indonesian children attained maturation stages from 0.5 to 1 year later in comparison with the Caucasian subjects. This study also proposed a model to predict the skeletal maturity index from chronological age that achieved good accuracy results in both cohorts. The study by Hsieh et al. [39] aimed to build on the TW method in an effort to simplify the RUS (Radius, Ulna and short bones) system assessment, in a way that only 9 out of the original 13 bones are assessed, reducing the time and effort needed for the BAA. An important study by Duren et al. [40] investigates the changes in bone maturation in two cohorts, one with subjects born between 1930 and 1964 and other with birth years between 1965 to 2001. The results showed that in the most recent cohort, the skeletal maturity was more advanced than in the earlier cohort for boys between ages of 0 to 8 years and 10 to 18 years, and girls between the ages of 4 to 17 years. McCormack et al. [41] conducted a longitudinal study with duration of 7 years, that performed annual assessments in children and adolescents from the African American and non-African American ancestry to investigate the effect of ancestry, height, BMI and pubertal maturation on the skeletal maturation. The results yielded that the subjects with African American ancestry had more advanced skeletal maturation, even when accounting for age, body composition and pubertal maturation. Lastly, the Davies et al. study [42] examine the presence of the epiphyseal scars in subjects after the bone growth ended. An epiphyseal scar is a thin layer of cartilage that remains between the diaphysis and epiphysis after they are completely fused. It is known that they remain for some time after the fusion, and the study in question investigates the bounds in which this occurs, concluding that subjects of the age of 50 still had visible epiphyseal scars.

Employed machine learning methods

This section aims to answer the research question RQ1 "Which machine learning techniques are being used in the age assessment studies?".

As shown in Table 5, the most frequently used techniques in the papers included in SLR were Regression-based methods (13 studies), followed by Artificial Neural Networks (8 studies) and Support Vector Machines (5 studies). Other less frequent techniques featured in the studies are: Bayesian Networks (2 studies), Decision Tress (1 study) and K-Nearest Neighbors (1 study). These results will be further detailed next, where we also provide a brief explanation of each of the techniques. Note that some studies used more than one technique.

Table 5. Summarized data about the ML techniques featured in the studies.

Machine Learning Techniques		Number of Studies
Regression-based methods	Linear regression [28–29, 29], Rule-based regression [28], General Linear Model [42], Mixed Effects Model [40–41], Logistic regression [40], Multiple Regression [38], Multivariate linear Stepwise regression [33], Polynomial Regression [34], Linear Discriminant analysis [31], Transition Analysis [30], Regression [18, 32, 37]	13
Artificial Neural Networks	Artificial Neural Networks [28, 35–36], Fuzzy Neural Networks [19],	4
Convolutional Neural Networks	Convolutional Neural Networks [22–25]	4
Support Vector Machines	Support Vector Machines [16, 21, 26, 28]	4
Bayesian Networks	Bayesian Networks [27], Gaussian process [28]	2
Decision Trees	Random Forest [20]	1
K-Nearest Neighbors	K-Nearest Neighbors [26]	1

https://doi.org/10.1371/journal.pone.0220242.t005

The **Regression-based me**thods aim in finding the effect of a set of independent variables in a dependent variable of interest. This is done by identifying a non-deterministic function (prone to random errors) that models the influence of the independent variables towards the mean of the dependent variable [43]. Despite being simple, Regression methods require a pre-determined model for data fitting.

These were the techniques employed the most in the included studies, being typically applied in the studies proposing non-automatic BAA systems, that is, systems that will require some sort of human intervention (commonly the assessment of epiphyses or other aspect of the medical image). Such studies aimed for a degree of automation, but also investigated the influence of multiple Regions of Interest (ROI) in the BAA. The only exception was the study by Thodberg et al. [18], which describes the BoneXpert, an automatic BAA system, which locates the ROI in hand radiographs automatically, but computes the Bone Age itself using a series of regressions.

An **Artificial Neural Network** is a technique that performs multifactorial analyses. It consists of a multi-layer structure that contains nodes connected by weighted edges that establish an input layer, one or more hidden layers and an output layer. With known input and output values, the network is trained by adjusting the weights in an incremental way until the network's output approaches the known output [44]. This technique is a powerful predictor that pairs well with image interpretation, but contrary to the regression case, this is a black box technique in which the results are not able to be interpreted in an intuitive way [45].

A specialization of the Artificial Neural Networks technique, which was frequently used in this SLR studies, is the **Convolutional Neural Network**. This technique enabled major advances in computer vision problems, where a computer needs to understand an input image to reconstruct its properties (i.e. color distributions, shapes, illumination etc) and perform a task such as localization, segmentation or detection of certain image elements [46]. One major difference from the traditional Artificial Neural Networks lays in its architecture, which is commonly composed by an input layer that receives an image as input, several convolutional and pooling layers, followed by several fully connected hidden layers, and finally the output layer [47]. The convolutional layers perform an image's feature extraction, and are organized into feature maps; the pooling layers reduce the feature maps to a point of spatial invariance (not affected by distortions and translations of the input image); and the fully connected layers are responsible for the interpretation of the abstract feature representations learned by the previous layers [47]. In what concerns the training of such type of network, it is analogous to the traditional Artificial Neural Network; however, it requires greater computational power and larger amounts of data.

Six out of the 8 studies that featured Artificial Neural Networks and Convolutional Neural Networks aimed to propose automatic BAA systems, which take as input a medical image and return the bone age.

The **Support Vector Machines** technique was conceived to solve classification problems and it is one of the most recently proposed ML techniques. The classification process happens in the following way: given a set of labeled data (data points with known classes), the algorithm maps the data points into a feature space (that does not include the outcome variable), then it finds in this feature space the location that separates the classes in the with the maximum margin of separation [48]. This technique does not require a pre-defined model and works in scenarios where there is a high number of variables in comparison to the number of data points; however, it still requires some algorithm considerations to be made beforehand (i.e. the choice of a kernel function) [49].

In the studies included in this SLR, the support vector machines were employed mostly proposing automatic BAA systems, in three out of the four cases.

The **Bayesian Network** technique estimates the posterior probability of a data point being of a certain class, given a set of features. The learning process happens in two phases: first it learns the structure of the network, which is a direct acyclic graph composed of nodes (representing the features) and edges (representing the probabilistic dependencies), then it learns the conditional probability distribution of each node [50]. The advantages of Bayesian Networks are that they are able to encode the knowledge of domain experts in the graph structure and they can work with smaller amounts of data, in comparison to other ML techniques [51]. A great disadvantage is that Bayesian Networks may become impractical in problems with a large number of variables.

Both studies that employed this ML technique proposed non-automatic BAA systems; one of these studies (Cunha et. al [28]) compared various techniques, and the Bayesian Network was not the best performing one.

Decision Trees aim in building a tree structure to represent knowledge that can be easily translated into if-then rules. It follows a recursive learning process that begins with each feature being tested on how well it alone can classify the data points into a certain class. The best one is set as the root node and the descendant nodes are set as the possible values (or ratios) of the selected feature. Then, the data points are sorted according to this rule. This process repeats until there are no possible splits [11]. The advantages of the Decision trees are that they are easy to interpret, to understand the interactions between features, and they do not depend on a pre-defined model [52]. The disadvantages are that they can be unstable and prone to overfitting [53].

The Decision Trees technique was employed in only one study by Urschler et al. [20], but in its ensemble form: Random Forest, which aims to address the abovementioned disadvantages of decision trees. That study proposed as a non-automatic BAA system.

The **K-Nearest Neighbors (KNN)** algorithm maps a data point of unknown class to a feature space. Then it assigns it to the class pertaining its k nearest neighbors. In this case, the concept of near is usually related to the Euclidean distance, but other measures can apply [54]. This technique is very simple and easy to implement, but it also considers all the features as equal in terms of importance, what can be a problem if superfluous features are inserted in the feature space [55].

The K-Nearest Neighbors technique was employed in the study by Harmsen et al. [26], which proposed a non-automatic BAA system. This study compared the KNN to SVM, and KNN presented worse performance.

Sample data characteristics

This section presents the results regarding the research question RQ2 "What data characteristics (database's origin, data collection mechanism, and ages) are being considered in the age assessment studies?"

Table 6 shows the summary regarding the data's origin in the samples of the included studies, grouped by region. Most of the studies utilized subjects from diverse countries of west Europe (8 studies), followed by North America (7 studies), Asia (6 studies) and finally Australia (2 studies). Two studies made use of samples from two regions, one used to build a model, and the other to validate the same model. Also, one of the studies reported no data origin, but mentioned the ethnicity of the sample to be Caucasian.

Regarding the studies' data collection, 20 out of 26 made use of private databases of medical images, or queries in hospital databases, but six studies made use of public databases for their experiments. These were:

Table 6. Summarized data regarding the origin of the subjects in the featured studies.

Region	Data's Origin	Featured Studies	Number of Studies	
Europe	Italy	[31–32]	2	**8**
	Austria	[20]	1	
	Denmark	[18]	1	
	Ireland	[29]	1	
	IRMA Database	[26]	1	
	Portugal	[28]	1	
	Scotland	[42]	1	
North America	United States of America	[22, 40–41]	3	**7**
	Digital Hand Atlas of the USC	[16, 21]	2	
	RSNA Pediatric Bone Age Challenge 2017 Database	[23–24]	2	
Asia	China	[33, 35–37]	4	**6**
	Taiwan	[19, 39]	2	
Australia	Australia	[30, 34]	2	**2**
Mixed	RSNA Pediatric Bone Age Challenge 2017 Database and China	[25]	1	**2**
	Unite States of America and Indonesia	[38]	1	
Non-Specified	Caucasian	[27]	1	**1**

https://doi.org/10.1371/journal.pone.0220242.t006

- Digital Hand Atlas of the University of Southern California (USC), which contains 1097 radiograph left hand images of subjects ranging from 0+ to 18 years of age, of both male and female genders and of different ethnicities, denoted as: African American, Asian, Caucasian and Hispanic.
- Radiological Society of North America (RSNA) Pediatric Bone Age Challenge 2017 Database, which was an initiative that brought 37 teams to develop algorithms for BAA. This database was developed by Stanford University and consisted of 12611 hand radiographs from male and female subjects, ranging from 0+ to 19 years of age. This database was only publicly available during the Challenge 2017 period.
- Image retrieval in Medical Applications (IRMA), which is a radiologic database brought together by the Aachen University of Technology in Germany, where one of its goals is to enable research in diverse automated classification problems involving radiographs. It comprises images from many body regions.

Note that few studies reported the ways used to verify the origin or ethnicity of the subjects in the sample; several reported that this information was not available, which may be the case when using private datasets. This problem is less significant in the case of the Asian studies, which employed data on subjects from China and Taiwan, since the Han ethnicity is very prominent in these countries reaching more than 90% of the total population and being the largest ethnical group in the world [56–57]. Other factors such as socioeconomic aspects were also not explored in the studies.

Regarding the subjects' age ranges for the samples used in studies detailed in this SLR, these were divided into 4 categories:

- Comprehensive Sample: sample contains data on subjects younger than 18 up to older than 18
- Bordering the age of 18: The maximum age of the sample is exactly 18 years' old

- Younger Subjects: the entire sample contains data on subjects younger than 18
- Older Subjects: the entire sample contains data on subjects older than 18

Table 7 shows that most of the studies, 17 out of 26, are employing a 'comprehensive' or a 'bordering 18 years old' sample, what may suggest that BAA is focused in assessing this specific age. In fact, all of the North American (7 studies) and half of the European studies (4 studies) focused their research on these types of sample. Eight studies were concerned with age assessment of younger subjects. There is only one case of the category Older Subjects, which is the study by Davies et al. [42], which aims to investigate the persistence of epiphyseal scars after the fusion of the diaphysis and epiphysis. This study warns about methods that consider the complete obliteration of the epiphyseal scars, since there are cases in which they remain not entirely fused until the fifth decade of life.

Age assessment methods

Table 8 summarizes the data regarding the BAA methods, which is the focus of RQ3: "What type of medical imaging are being used in the studies?". Evidence shows that to date the research being conducted in BAA favors the use of Radiographs over other types of medical imaging, with 22 out of 27 studies utilizing this method. Two other studies used radiographs in conjunction with MRI. The study by Hillewig et al. [27] made use of radiographs for the regions of hand and wrist, and MRI for the clavicle. Fan et al. [37] utilized these two methods in the knee region, but comparing the performance of both for the BAA, which yielded better results for the MRI. The studies that only featured MRI as the method of choice [20, 35] investigated the hand and wrist regions and features related to the appearance of osseous structures in terms of signal intensity provided by the MRI. The only case where a computer tomography (CT) assessed the Bone Age in the region of the clavicle was the study by Franklin and Flavel [30].

Regions of interest

This section presents the summarized data used to answer RQ4: "What are the regions of interest being explored for the age assessment studies (hand, wrist, knee, etc) and what were the methods used to assess them?".

As shown in Table 9, the research is very much focused on the hand region for estimating Bone Age, accounting to 22 out of the 27 studies in this SLR; however, note that only in seven cases the hand alone was employed in the models. Next, the wrist was the second mostly investigated region, present in 17 out the 27 studies; however, in the same way as the wrist, it was employed exclusively in only two studies. In most of cases, data on both hand and wrist are used, what can be explained by the easiness of getting both regions in the same image. Other less frequent ROIs present in the BAA studies are the knee (2 studies) and clavicle (2 studies). It is important to notice that in the case of the teeth and cervical, even if they are ROI, this SLR focused in the BAA analyzed through growth zones in joints, so they were classified as "Other variables".

Table 7. Summarized data regarding the age ranges of the subjects in the features studies.

Sample's Characteristics	Featured Studies	Number of Studies
Comprehensive Sample	[20, 23–24, 27–30, 34, 36–37, 40]	11
Younger Subjects	[18–19, 31–33, 35, 38–39]	8
Bordering the age of 18	[16, 21–22, 25–26, 41]	6
Older Subjects	[42]	1

https://doi.org/10.1371/journal.pone.0220242.t007

Table 8. Methods used for BAA.

Type of Image	Number	Featured Studies
Radiograph	21	[16, 18–19, 21–26, 28–29, 31–34, 36, 38, 39–42]
MRI	2	[20, 35]
MRI, Radiograph	2	[27, 37]
CT	1	[30]

https://doi.org/10.1371/journal.pone.0220242.t008

Other aspect that can be evidenced is that only 12 out of the 26 studies explored only one ROI. Furthermore, only five studies employed additional variables not related to the assessment of growth zones: dental assessment, cervical assessment, weight, height, DNA Methylation, Tanner scale, fat mass, lean mass and BMI.

With regard to BAA techniques, the research deviated from the classic BAA of growth zones—the TW and GP methods, which were only present in five studies each. Furthermore, computer vision techniques, which do not depend on any pre-existent guide, still have a subtle presence in the BAA research, accounting for 8 out of 26 studies. This is enabled by the advances in computing power and the existence of a large amount of digital labeled data [45].

Other techniques for BAA that were employed in specific ROIs are summarized in the Table 10. Following the same trend as TW and GP, these are either based on maturity scores or atlases.

Performance of the techniques in the included studies

The information regarding the performance of the techniques employed in the studies that proposed systems for BAA was extracted and it is summarized in this section. Results show that a

Table 9. Summarized data regarding ROI, types of medical image, additional variables and techniques for BAA in the studies of the SLR.

Regions of Interest	Type of Image	Other Variables	Techniques for BAA	Featured Studies
Hand, Wrist	MRI	None	Computer Vision	[20]
	Radiograph	None	Computer Vision	[22, 24–25]
	Radiograph	None	GP, TW, Computer Vision	[18]
	Radiograph	None	TW	[34, 36, 39]
	Radiograph	None	Fels Method [58]	[40]
	Radiograph	Cervical Assessment	Skeleton Maturation Index by Fishman [59]	[38]
	Radiograph	Dental Assessment	GP, TW	[31]
	Radiograph	DNA Methylation, Dental Assessment	GP, TW, DNA Methylation	[33]
	Radiograph	BMI, Height, Tanner scale, Fat Mass, Lean Mass	GP	[41]
	MRI	Weight, Height	TW	[35]
Hand	Radiograph	None	Computer Vision	[16, 23, 26]
	Radiograph	None	Gilsanz and Ratib [60]	[21]
	Radiograph	None	Own Method	[19]
	Radiograph	None	TW	[28]
Wrist	Radiograph	None	Cameriere et al. [61]	[32]
			Own method	[42]
Knee	Radiograph	None	O'Connor et al. [62]	[29]
	Radiograph and MRI	None	Kramer et al. [63]	[37]
Clavicle	CT	None	Schmeling et al. [64]	[30]
Hand, Wrist, Clavicle	Radiograph and MRI	None	Schmeling et al. [64], Kreitner et al. [65], GP	[27]

https://doi.org/10.1371/journal.pone.0220242.t009

Table 10. Techniques for BAA by ROI.

ROI	Techniques for BAA
Hand, Wrist	**Fels Method:** This method was proposed as a less subjective way to assess skeletal age. It considers ossification, radiopaque densities, bony projection, shape changes and fusion. It comprises 98 indicators of bone maturity, where 85 are categorical and are 13 continuous (epiphyseal and metaphyseal fusion ratios) [58].
	Skeleton Maturation Indicators by Fishman: It assesses the skeletal maturity based on the following indicators: width of the epiphysis compared to the diaphysis (third and fifth fingers), gapping of epiphysis (third and fifth fingers), fusion of epiphysis and diaphysis (third, fifth fingers and radius) and ossification of adductor sesamoid of the thumb [59].
Hand	**Gilsanz and Ratib:** This method consists of a digital hand atlas with reference images of 29 classes from the ages of 0 to 18 with an additional class to represent subjects older than 18. [60]
Wrist	**Cameriere et al.**: This is a quantitative method that proposes a mathematical formula based on the ratio of the carpal area and the total area of carpal bones and epiphyses of the radius and ulna [61].
Knee	**O'Connor et al.**: A method that proposes five stages of epiphyseal fusion for the femur, tibia and fibula bones. It uses the frontal and lateral radiograph image of the knee [62].
	Krammer et al.: A method that proposes five stages of epiphyseal fusion (from 1 to 5 of which stages 2 and 3 have 3 sub stages each) of the distal femur [63].
Clavicle	**Schmeling et al.:** This method defines 5 stages of ossification of the medial clavicular epiphysis [64].
	Kreitner et al.: This method is similar to the one proposed by Schmeling et al. (63), but instead of 5 stages there are 4, and the fourth stage is comparable to a combination of the stages 4 and 5 from the Schmeling classification [65].

https://doi.org/10.1371/journal.pone.0220242.t010

wide variety of different metrics were used to measure a technique's performance what makes comparisons between studies difficult. The most commonly used metric was the mean average error (MAE), which is the average of all the absolute errors, used by seven studies. Table 11 shows the studies that used this metric, and their performance information. The reported performances were aggregated by means of the average of the performances weighted by the sample sizes to perform the meta-analysis, so to understand the general performance of BAA systems.

The result of the weighted average was 9.96 MAE (months), which could indicate that BAA systems could reliably predict the bone age of a subject from zero to 19 years old, but this results should be regarded with caution since this statistic was based on only 7 out of the 20 studies that proposed BAA systems. The studies that employed the MAE metric had relatively similar age ranges, but still varied in two of the studies which did not consider infant and children subjects. Also, not all of the studies had a uniform sample in terms of age distribution. In the RSNA challenge database, although the sample was large enough, only 0.1% of the sample was composed of subjects of 18 years old and older, which is a very important bone age to be predicted for legal reasons.

Table 11. Performance of the comparable studies in terms of the mean absolute error (MAE) in months.

Proposed Method	Dataset size	Age Range	Performance in MAE (months)	Commentary
Ren et al. (2018) [25]	12480	0–18	5.2	2017 RSNA Pediatric Bone Age challenge entry, but the method is tested in a different sample of age range 0–18.
Kashif et al. (2016) [21]	1101	0–18	7.26	
Iglovikov et al. (2018) [24]	11600	0–19	7.52	2017 RSNA Pediatric Bone Age challenge entry
Zhao et al. (2018) [23]	12611	0–19	7.66	2017 RSNA Pediatric Bone Age challenge entry
Harmsen et al. (2013) [26]	1097	0–19	9.96	
Urschler et al. (2015) [20]	102	13–20	10.2	
Cunha et al. (2014) [28]	887	7–19	10.16	

https://doi.org/10.1371/journal.pone.0220242.t011

The remaining studies that proposed BAA systems, besides using a wide variety of different performance metrics, also were more heterogeneous in terms of age ranges, making the comparison between them not viable. The results for the performances of the employed techniques by the studies are shown in Table 12.

Discussion

Discussion of the current evidence

This SLR's main findings can be summarized as follows: (i) Most studies aimed to propose an automatic BAA system; (ii) The BAA research has focused on hand and wrist radiographs; (iii) Most studies made use of samples from either the United the States or from West Europe; (iv) Studies that considered ethnical differences were scarce and socioeconomic aspects were inexistent; (v) The estimations on Bone Age were using samples where subjects' age ranged from below to above 18 years of age, or bordering 18 years of age; (vi) The average performance weighted by sample size of the compared studies resulted in a MAE of 9.96 months, but there is still high heterogeneity in the studies what makes the comparing them a challenge.

The evidence gathered in this SLR suggests a clear trend towards automating the age identification within the context of BAA research. Most studies aimed at proposing automatic systems that would not require human intervention; however, a considerable amount of other studies proposed systems that do not automate age identification but reduce the dependability upon human input. We believe that either solution can be motivated by the following issues: i) to reduce the subjectivity of the traditional BAA methods, which depend upon radiologists' judgment and experience, and, as a consequence, can lead to inter-rater and intra-rater

Table 12. Performance of the non-comparable studies.

Proposed Method	Dataset size	Age Range	Performance	Commentary
Shi et al. (2017) [33]	124	6–15	0.47 (male) and 0.33 (female) MAE (years)	
Haak et al. (2013) [16]	1097	0–18	0.73 RMS	The RMS metric is the root mean squared error is the square root of the mean square error
Thodberg et al. (2009) [18]	1559	2–17	0.42 (GP), 0.80 (TW) MSE	The MSE is the mean squared error and measures the average squared differences between the estimated and observed values.
Lin et al. (2012) [19]	600	0–14	0.26 MSE	Predicts a bone age cluster instead of bone age.
Lee et al. (2017) [22]	8325	5–18	61.40% (male) and 57.32% (female) accuracy	
Maggio et al. (2016) [34]	360	0–24	1.31 (male) and 2.37 (female) SEE (years)	SEE is the standard error of estimate and measures the variation from the regression line.
De Luca et al. (2016) [32]	332	1–16	0.38 median of the absolute values of residuals	
O'Connor et al. (2014) [29]	221	9–19	-2.0 to +2.9 (male) and -2.3 to +2.4 (female) range residuals	
Pinchi et al. (2016) [31]	274	6–17	80.4% (male) and 83.3% (female) accuracy	The TW method performed better than the proposed method for male subjects (negative results).
Tang et al. (2018) [35]	79	12–17	0.13 (male) and 0.08 (female) mean disparity (years)	The mean disparity is a metric that compares the mean chronological age of all subjects to the mean estimated age for all subjects.
Hsieh et al. (2011) [39]	534	2–15	96.2% (male) and 95% (female) relative accuracy	Measures the relative accuracy between the proposed method and the TW method.
Franklin, D.; Flavel, A. (2015) [30]	388	10–35	NA	Create stages of ossification for the clavicle and compares to the bone age.
Hillewig et al. (2013) [27]	220	16–26	NA	Calculates the probability of being of bone age younger or older than 18 instead of the actual bone age.

https://doi.org/10.1371/journal.pone.0220242.t012

variability [66–67]; ii) the traditional BAA methods are a time consuming activity, which demand is increasing [66]; and iii) more automatic solutions could reduce assessment costs as they would require radiologists to spend less time in such this activity [9].

Note that in relation to using an automated solution to chronological age identification, the use of ML technologies appears to be a significant enabler of automatic BAA solutions; this is observed in particular when using Convolutional Neural Networks for computer vision tasks, which contrast with the use of regression-based methods techniques by systems that still require some human intervention. As a remark, the GP atlas, which is being used nowadays as a BAA, was not created to determine chronological age, but to compare the skeletal development of children and adolescents to their chronological age [5]; nevertheless some studies still used this method as the base for their investigations on age assessment.

Further, this SLR also showed that the research on BAA systems has focused upon methods that employ hand and wrist radiographs. However, such choice of medical imaging on healthy children and adolescents—the use of radiographs, hence radiation, without therapeutic purposes could raise ethical issues [5]. When comparing radiographs with other forms of medical imaging, such as Magnetic Resonance Imaging (MRI), there is the argument that the latter is more expensive; however, on the other hand, the MRI technology offers better contrast resolution, which in turn offers a more accurate analysis of the growth plate, especially considering 3.0-T MRI [68]. The small presence of the MRI technology on the research on BAA could be considered a gap in the research.

With regard to the predominant use in this SLR studies of hand and wrist as the focused ROI, one can argue that this is a small area that contains a large concentration of epiphyseal plates, thus making it easier to gather images from this region without much effort. Also, very few studies investigated the BAA with more than one ROI whenever they were not using hand and wrist, suggesting a gap in the current research.

This SLR also evidenced that two regions seem to show predominant interest in such research—the United States and West Europe; such attention resulted in the creation of databases of medical imaging (USC, IRMA and RSNA), and the Pediatric Bone Age Challenge 2017, which can be viewed as an effort to have standardization as basis of comparison of different studies on this topic.

In what concerns a sample's origin, most studies did not report on mechanisms used to document it. In addition, the ethnicity aspect was not much explored in the studies and the socioeconomic element was not investigated, what can be viewed as a gap in the current research on BAA. Only two studies (Soegiharto et al. [38] and McCormack et al. [41]) approached the effects of ethnicity in skeletal maturation. Besides the contradictory evidence in the literature about its influence [5], in this SLR both studies reported on differences in skeletal maturation on Indonesians and African Americans, in comparison to Caucasians, in the way that the first mature later and the second matures earlier. In contrast to that, the study by Thodberg et al. [18], which proposed a BAA system that is currently in commercial use—the BoneXpert, investigated the BAA issue using samples of Japanese [69], Dutch [70] and American subjects of four ethnicities (African American, Asian, Caucasian and Hispanic) [71].

Furthermore, regarding the subjects' age in the samples used by this SLR's included studies, most age ranges either included or bordered the age of 18 years. The interest for this particular age is justified by the legal systems in many countries, where younger than 18 individuals are classified as minors.

In regards to the meta-analysis of performances, it was evidenced a high heterogeneity in terms of age-ranges, dataset sizes and performance metrics that make the comparison between studies a challenge. The seven comparable studies had quite similar age-ranges and resulted in

a weighted average of 9.96 MAE (months), but caution should be made as the age distribution was not uniform in all of the studies.

Limitations

With regard to this SLR's limitations, they include whether a suitable large representative sample of relevant included studies were selected, and also the non-medical expertise of A1. The first issue was mitigated via using a more inclusive selection strategy, i.e., whenever a paper's abstract did not present all the information needed for its inclusion or exclusion, it would be included in the first phase to be fully read later. As for the second issue, A1's lack of medical expertise of A1 was mitigated by consulting A3 whenever necessary.

Future perspectives

The results of this SLR presented trends and gaps in the current research on age assessment that should be addressed, such as other common factors that could influence delay or acceleration of skeletal maturity and the further investigation of other ROIs for BAA.

Conclusion

This paper detailed an SLR on the topic of age assessment in growth zones research with the use of ML techniques, which resulted in the selection of a final set of 27 studies. These studies were summarized in terms of ML techniques applied, sample data, age assessment methods and regions of interest.

Our findings indicate the focus of the research on investigating the hand and wrist ROIs with radiographs, with most of the samples from the United States or West Europe. It has also pointed out gaps in the research, such as few studies on different ethnicities, no studies considering socioeconomic differences, and few studies considering more ROIs other than hand and wrist.

Supporting information

S1 Table. Included studies in the systematic literature review.
(PDF)

S2 Table. PRISMA checklist.
(PDF)

Author Contributions

Conceptualization: Ana Luiza Dallora, Peter Anderberg, Ola Kvist, Emilia Mendes, Sandra Diaz Ruiz, Johan Sanmartin Berglund.

Data curation: Ana Luiza Dallora.

Formal analysis: Ana Luiza Dallora.

Investigation: Ana Luiza Dallora.

Methodology: Ana Luiza Dallora, Peter Anderberg, Emilia Mendes, Johan Sanmartin Berglund.

Project administration: Peter Anderberg, Johan Sanmartin Berglund.

Supervision: Peter Anderberg, Emilia Mendes, Johan Sanmartin Berglund.

Validation: Ana Luiza Dallora, Peter Anderberg, Emilia Mendes, Johan Sanmartin Berglund.

Visualization: Ana Luiza Dallora.

Writing – original draft: Ana Luiza Dallora.

Writing – review & editing: Peter Anderberg, Ola Kvist, Emilia Mendes, Sandra Diaz Ruiz, Johan Sanmartin Berglund.

References

1. Gilsanz V, Ratib O. Hand Bone Age: A Digital Atlas Of Skeletal Maturity. Springer Science & Business Media; 2005. 106 p.
2. Manzoor Mughal A, Hassan N, Ahmed A. Bone Age Assessment Methods: A Critical Review. Pak J Med Sci. 2014; 30(1):211–5. https://doi.org/10.12669/pjms.301.4295 PMID: 24639863
3. Satoh M. Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol. 2015; 24 (4):143–52. https://doi.org/10.1297/cpe.24.143 PMID: 26568655
4. UN Convention on the Rights of the Child (UNCRC). Unicef UK. [cited 2019 Mar 20]. Available from: https://www.unicef.org.uk/what-we-do/un-convention-child-rights/
5. Hjern A, Brendler-Lindqvist M, Norredam M. Age assessment of young asylum seekers. Acta Paediatr. 2012; 101(1):4–7. https://doi.org/10.1111/j.1651-2227.2011.02476.x PMID: 21950617
6. UNHCR The UN Refugee Agency. Global Trends FORCED DISPLACEMENT IN 2017. 2017 [cited 2019 Feb 9]. Available from: https://www.unhcr.org/5b27be547.pdf
7. William Walter Greulich Sarah Idell Pyle, Thomas Wingate Todd. Radiographic atlas of skeletal development of the hand and wrist. Vol. 2. Stanford: Stanford university press; 1959.
8. Tanner James Mourilyan, Healy Michael J. R., Cameron N., Goldstein H. Assessment of skeletal maturity and prediction of adult height (TW3 method). London: WB Saunders; 2001.
9. Mansourvar M, Ismail MA, Herawan T, Gopal Raj R, Abdul Kareem S, Nasaruddin FH. Automated Bone Age Assessment: Motivation, Taxonomies, and Challenges [Internet]. Computational and Mathematical Methods in Medicine. 2013
10. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Brief Bioinform. 2006 Mar 1; 7(1):86–112. https://doi.org/10.1093/bib/bbk007 PMID: 16761367
11. Mitchell TM. Machine Learning and Data Mining. Commun ACM. 1999 Nov; 42(11):30–36.
12. Moher D, Liberati A, Tetzlaff J, Altman DG, Group TP. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Med. 2009 Jul 21; 6(7):e1000097. https://doi.org/10.1371/journal.pmed.1000097 PMID: 19621072
13. Kitchenham B, Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. 2007.
14. Pai M, McCulloch M, Gorman JD, Pai N, Enanoria W, Kennedy G, et al. Systematic reviews and meta-analyses: an illustrated, step-by-step guide. Natl Med J India. 2004; 17(2):86–95. PMID: 15141602
15. Glass GV. Primary, Secondary, and Meta-Analysis of Research. Educational Researcher, 5(10), 3–8. 1976.
16. Haak D, Yu J, Simon H, Schramm H, Seidl T, Deserno TM. Bone age assessment using support vector regression with smart class mapping. In: Novak CL, Aylward S, editors. Lake Buena Vista (Orlando Area), Florida, USA; 2013. p. 86700A.
17. Haak D, Simon H, Yu J, Harmsen M, Deserno TM. Bone Age Assessment Using Support Vector Machine Regression. In: Meinzer H-P, Deserno TM, Handels H, Tolxdorff T, editors. Bildverarbeitung für die Medizin 2013. Springer Berlin Heidelberg; 2013. p. 164–9. (Informatik aktuell).
18. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging. 2009 Jan; 28(1):52–66. https://doi.org/10.1109/TMI.2008.926067 PMID: 19116188
19. Lin H-H, Shu S-G, Lin Y-H, Yu S-S. Bone age cluster assessment and feature clustering analysis based on phalangeal image rough segmentation. Pattern Recognit. 2012 Jan 1; 45(1):322–32.
20. Urschler M, Grassegger S, Štern D. What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents. Ann Hum Biol. 2015; 42(4):358–67. https://doi.org/10.3109/03014460.2015.1043945 PMID: 26313328

21. Kashif M, Deserno TM, Haak D, Jonas S. Feature description with SIFT, SURF, BRIEF, BRISK, or FREAK? A general question answered for bone age assessment. Comput Biol Med. 2016 Jan 1; 68:67–75. https://doi.org/10.1016/j.compbiomed.2015.11.006 PMID: 26623943

22. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully Automated Deep Learning System for Bone Age Assessment. J Digit Imaging. 2017 Aug 1; 30(4):427–41. https://doi.org/10.1007/s10278-017-9955-8 PMID: 28275919

23. Zhao C, Han J, Jia Y, Fan L, Gou F. Versatile Framework for Medical Image Processing and Analysis with Application to Automatic Bone Age Assessment, Journal of Electrical and Computer Engineering, vol. 2018, Article ID 2187247, 13 pages, 2018.

24. Iglovikov VI, Rakhlin A, Kalinin AA, Shvets AA. Paediatric Bone Age Assessment Using Deep Convolutional Neural Networks. In: Stoyanov D, Taylor Z, Carneiro G, Syeda-Mahmood T, Martel A, Maier-Hein L, et al., editors. Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support. Springer International Publishing; 2018. p. 300–8. (Lecture Notes in Computer Science).

25. Ren X, Li T, Yang X, Wang S, Ahmad S, Xiang L, et al. Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment from Hand Radiograph. IEEE J Biomed Health Inform. 2018 Oct 19;

26. Harmsen M, Fischer B, Schramm H, Seidl T, Deserno TM. Support Vector Machine Classification Based on Correlation Prototypes Applied to Bone Age Assessment. IEEE J Biomed Health Inform. 2013 Jan; 17(1):190–7. https://doi.org/10.1109/TITB.2012.2228211 PMID: 23192601

27. Hillewig E, Degroote J, Van der Paelt T, Visscher A, Vandemaele P, Lutin B, et al. Magnetic resonance imaging of the sternal extremity of the clavicle in forensic age estimation: towards more sound age estimates. Int J Legal Med. 2013 May; 127(3):677–89. https://doi.org/10.1007/s00414-012-0798-z PMID: 23224029

28. Cunha P, Moura DC, Guevara López MA, Guerra C, Pinto D, Ramos I. Impact of Ensemble Learning in the Assessment of Skeletal Maturity. J Med Syst. 2014 Jul 11; 38(9):87. https://doi.org/10.1007/s10916-014-0087-0 PMID: 25012476

29. O'Connor JE, Coyle J, Bogue C, Spence LD, Last J. Age prediction formulae from radiographic assessment of skeletal maturation at the knee in an Irish population. Forensic Sci Int. 2014 Jan; 234:188.e1-8.

30. Franklin D, Flavel A. CT evaluation of timing for ossification of the medial clavicular epiphysis in a contemporary Western Australian population. Int J Legal Med. 2015 May; 129(3):583–94. https://doi.org/10.1007/s00414-014-1116-8 PMID: 25398635

31. Pinchi V, De Luca F, Focardi M, Pradella F, Vitale G, Ricciardi F, et al. Combining dental and skeletal evidence in age classification: Pilot study in a sample of Italian sub-adults. Leg Med Tokyo Jpn. 2016 May; 20:75–9.

32. De Luca S, Mangiulli T, Merelli V, Conforti F, Velandia Palacio LA, Agostini S, et al. A new formula for assessing skeletal age in growing infants and children by measuring carpals and epiphyses of radio and ulna. J Forensic Leg Med. 2016 Apr; 39:109–16. https://doi.org/10.1016/j.jflm.2016.01.030 PMID: 26874435

33. Shi L, Jiang F, Ouyang F, Zhang J, Wang Z, Shen X. DNA methylation markers in combination with skeletal and dental ages to improve age estimation in children. Forensic Sci Int Genet. 2018; 33:1–9. https://doi.org/10.1016/j.fsigen.2017.11.005 PMID: 29172065

34. Maggio A, Flavel A, Hart R, Franklin D. Skeletal age estimation in a contemporary Western Australian population using the Tanner-Whitehouse method. Forensic Sci Int. 2016; 263:e1–8. https://doi.org/10.1016/j.forsciint.2016.03.042 PMID: 27080619

35. Tang FH, Chan JLC, Chan BKL. Accurate Age Determination for Adolescents Using Magnetic Resonance Imaging of the Hand and Wrist with an Artificial Neural Network-Based Approach. J Digit Imaging. 2018 Oct 15;

36. Liu J, Qi J, Liu Z, Ning Q, Luo X. Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method. Comput Med Imaging Graph. 2008 Dec 1; 32(8):678–84. https://doi.org/10.1016/j.compmedimag.2008.08.005 PMID: 18835130

37. Fan F, Zhang K, Peng Z, Cui J-H, Hu N, Deng Z-H. Forensic age estimation of living persons from the knee: Comparison of MRI with radiographs. Forensic Sci Int. 2016 Nov; 268:145–50. https://doi.org/10.1016/j.forsciint.2016.10.002 PMID: 27770721

38. Soegiharto BM, Cunningham SJ, Moles DR. Skeletal maturation in Indonesian and white children assessed with hand-wrist and cervical vertebrae methods. Am J Orthod Dentofac Orthop Off Publ Am Assoc Orthod Its Const Soc Am Board Orthod. 2008 Aug; 134(2):217–26.

39. Hsieh C-W, Liu T-C, Wang J-K, Jong T-L, Tiu C-M. Simplified radius, ulna, and short bone-age assessment procedure using grouped-Tanner-Whitehouse method. Pediatr Int Off J Jpn Pediatr Soc. 2011 Aug; 53(4):567–75.

40. Duren DL, Nahhas RW, Sherwood RJ. Do Secular Trends in Skeletal Maturity Occur Equally in Both Sexes? Clin Orthop. 2015 Aug; 473(8):2559–67. https://doi.org/10.1007/s11999-015-4213-1 PMID: 25716212

41. McCormack SE, Chesi A, Mitchell JA, Roy SM, Cousminer DL, Kalkwarf HJ, et al. Relative Skeletal Maturation and Population Ancestry in Nonobese Children and Adolescents. J Bone Miner Res Off J Am Soc Bone Miner Res. 2017; 32(1):115–24.

42. Davies C, Hackman L, Black S. The persistence of epiphyseal scars in the distal radius in adult individuals. Int J Legal Med. 2016 Jan; 130(1):199–206. https://doi.org/10.1007/s00414-015-1192-4 PMID: 25904079

43. Fahrmeir L, Kneib T, Lang S, Marx B. Regression: Models, Methods and Applications. Springer Science & Business Media; 2013. 619 p.

44. Dayhoff JE, DeLeo JM. Artificial neural networks. Cancer. 2001 Apr 15; 91(S8):1615–35.

45. Dybowski R, Gant V. Clinical Applications of Artificial Neural Networks. Cambridge University Press; 2001. 382 p.

46. Szeliski R. Computer Vision: Algorithms and Applications. Springer Science & Business Media; 2010. 824 p.

47. Rawat W and Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Computation 2017 29:9, 2352–2449 https://doi.org/10.1162/NECO_a_00990 PMID: 28599112

48. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995 Sep 1; 20(3):273–97.

49. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. Data Min Knowl Discov. 1998 Jun 1; 2(2):121–67.

50. Cheng J, Greiner R. Comparing Bayesian Network Classifiers. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999

51. van Gerven MAJ, Taal BG, Lucas PJF. Dynamic Bayesian networks as prognostic models for clinical patient management. J Biomed Inform. 2008 Aug; 41(4):515–29. https://doi.org/10.1016/j.jbi.2008.01.006 PMID: 18337188

52. Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform. 2006 Jan 1; 2:117693510600200030.

53. Li R-H, Belford GG. Instability of Decision Tree Classification Algorithms. In: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM; 2002

54. Keller JM, Gray MR, Givens JA. A fuzzy K-nearest neighbor algorithm. IEEE Trans Syst Man Cybern. 1985 Jul; SMC-15(4):580–5.

55. Phyu TN. Survey of classification techniques in data mining. In: Proceedings of the International MultiConference of Engineers and Computer Scientists. 2009. p. 18–20.

56. Dincer OC, Wang F. Ethnic diversity and economic growth in China. J Econ Policy Reform. 2011; 14 (1):1–10.

57. CIA Central Intelligence Agency Guide to Country Profiles—The World Factbook—Central Intelligence Agency. [cited 2019 Mar 2]. Available from: https://www.cia.gov/library/publications/the-world-factbook/docs/profileguide.html

58. Jacobson A. Assessing the skeletal maturity of the hand-wrist: FELS method: Alex F. Roche, William Cameron Chumlea, and David Thissen, Springfield, Illinois: Charles C. Thomas, 1988. 339 pages, $57.50. Am J Orthod Dentofacial Orthop. 1989 May 1; 95(5):449.

59. Fishman LS. Radiographic Evaluation of Skeletal Maturation. Angle Orthod. 1982 Apr 1; 52(2):88–112. https://doi.org/10.1043/0003-3219(1982)052<0088:REOSM>2.0.CO;2 PMID: 6980608

60. Gilsanz V, Ratib O. Hand Bone Age: A Digital Atlas of Skeletal Maturity [Internet]. 2nd ed. Berlin Heidelberg: Springer-Verlag; 2012 [cited 2019 Mar 6]. Available from: https://www.springer.com/gp/book/9783642237614

61. Cameriere R, Ferrante L, Mirtella D, Cingolani M. Carpals and epiphyses of radius and ulna as age indicators. Int J Legal Med. 2006 May 1; 120(3):143–6. https://doi.org/10.1007/s00414-005-0040-3 PMID: 16211419

62. O'Connor JE, Bogue C, Spence LD, Last J. A method to establish the relationship between chronological age and stage of union from radiographic assessment of epiphyseal fusion at the knee: an Irish population study. J Anat. 2008 Feb; 212(2):198–209. https://doi.org/10.1111/j.1469-7580.2007.00847.x PMID: 18179475

63. Krämer JA, Schmidt S, Jürgens K-U, Lentschig M, Schmeling A, Vieth V. Forensic age estimation in living individuals using 3.0T MRI of the distal femur. Int J Legal Med. 2014 May 1; 128(3):509–14. https://doi.org/10.1007/s00414-014-0967-3 PMID: 24504560

64. Schmeling A, Schulz R, Reisinger W, Mühler M, Wernecke K-D, Geserick G. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med. 2004 Feb; 118(1):5–8. https://doi.org/10.1007/s00414-003-0404-5 PMID: 14534796

65. Kreitner K-F, Schweden FJ, Riepert T, Nafe B, Thelen M. Bone age determination based on the study of the medial extremity of the clavicle. Eur Radiol. 1998 Sep 1; 8(7):1116–22. https://doi.org/10.1007/s003300050518 PMID: 9724422

66. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a Deep-Learning Neural Network Model in Assessing Skeletal Maturity on Pediatric Hand Radiographs. Radiology. 2017 Nov 2; 287(1):313–22. https://doi.org/10.1148/radiol.2017170236 PMID: 29095675

67. Roche AF, Rohmann CG, French NY, Dávila GH. Effect of training on replicability of assessments of skeletal maturity (greulich-pyle). Am J Roentgenol. 1970 Mar 1; 108(3):511–5.

68. Margalit A, Cottrill E, Nhan D, Yu L, Tang X, Fritz J, et al. The Spatial Order of Physeal Maturation in the Normal Human Knee Using Magnetic Resonance Imaging. J Pediatr Orthop. 2019 Apr; 39(4):e318–22. https://doi.org/10.1097/BPO.0000000000001298 PMID: 30451813

69. Martin DD, Sato K, Sato M, Thodberg HH, Tanaka T. Validation of a New Method for Automated Determination of Bone Age in Japanese Children. Horm Res Paediatr. 2010; 73(5):398–404. https://doi.org/10.1159/000308174 PMID: 20389112

70. van Rijn RR, Lequin MH, Thodberg HH. Automatic determination of Greulich and Pyle bone age in healthy Dutch children. Pediatr Radiol. 2009 Jun 1; 39(6):591–7. https://doi.org/10.1007/s00247-008-1090-8 PMID: 19125243

71. Thodberg HH, Sävendahl L. Validation and Reference Values of Automated Bone Age Determination for Four Ethnicities. Acad Radiol. 2010 Nov 1; 17(11):1425–32. https://doi.org/10.1016/j.acra.2010.06.007 PMID: 20691616

S1 Table. Included studies in the systematic literature review

	Title	Author	Publication Title	Publication Year	Automatic or Non-Automatic BAA	ROI	Type of image
1	A new formula for assessing skeletal age in growing infants and children by measuring carpals and epiphyses of radio and ulna [31]	De Luca, S. et al.	Journal of Forensic and Legal Medicine	2016	Non-Automatic	Wrist	Radiograph
2	Accurate Age Determination for Adolescents Using Magnetic Resonance Imaging of the Hand and Wrist with an Artificial Neural Network-Based Approach [34]	Tang, F. H. et al.	Journal of Digital Imaging	2018	Non-Automatic	Hand, wrist	MRI
3	Age prediction formulae from radiographic assessment of skeletal maturation at the knee in an Irish population [28]	O'Connor, J. E. et al.	Forensic Science International	2014	Non-Automatic	Knee	Radiograph
4	Automatic bone age assessment based on intelligent algorithms and comparison with TW3 method [35]	Liu, J. et al.	Computerized Medical Imaging and Graphics: The Official Journal of the Computerized Medical Imaging Society	2008	Automatic	Hand, wrist	Radiograph
5	Bone age assessment using support vector regression with smart class mapping [15]	Haak, D. et al.	Proceedings of SPIE - The International Society for Optical Engineering	2013	Automatic	Hand	Radiograph
6	Bone age cluster assessment and feature clustering analysis based on phalangeal image rough segmentation [18]	Lin, H. H. et al.	Pattern Recognition	2012	Automatic	Hand	Radiograph
7	Combining dental and skeletal evidence in age classification: Pilot study in a sample of Italian sub-adults [30]	Pinchi, V. et al.	Legal Medicine (Tokyo, Japan)	2016	Non-Automatic	Hand, wrist	Radiograph
8	CT evaluation of timing for ossification of the medial clavicular epiphysis in a contemporary Western Australian population [29]	Franklin, D.; Flavel, A.	International Journal of Legal Medicine	2015	Non-Automatic	Clavicle	Computer Tomography

9	DNA methylation markers in combination with skeletal and dental ages to improve age estimation in children [32]	Shi, L. et al.	Forensic Science International. Genetics	2017	Non-Automatic	Hand, wrist	Radiograph
10	Do Secular Trends in Skeletal Maturity Occur Equally in Both Sexes? [39]	Duren, D. L. et al.	Clinical Orthopaedics and Related Research	2015	Non-Automatic	Hand, wrist	Radiograph
11	Feature description with SIFT, SURF, BRIEF, BRISK, or FREAK? A general question answered for bone age assessment [20]	Kashif, M. et al.	Computers in Biology and Medicine	2016	Automatic	Hand	Radiograph
12	Forensic age estimation of living persons from the knee: Comparison of MRI with radiographs [36]	Fan, F. et al.	Forensic Science International	2016	Non-Automatic	Knee	Radiograph, MRI
13	Fully Automated Deep Learning System for Bone Age Assessment [21]	Lee, H. et al.	Journal of Digital Imaging	2017	Automatic	Hand, wrist	Radiograph
14	Impact of ensemble learning in the assessment of skeletal maturity [27]	Cunha, P. et al.	Journal of Medical Systems	2014	Non-Automatic	Hand	Radiograph
15	Magnetic resonance imaging of the sternal extremity of the clavicle in forensic age estimation: towards more sound age estimates [26]	Hillewig, E. et al.	International Journal of Legal Medicine	2013	Non-Automatic	Hand, wrist, clavicule	MRI (clavicle), Radiograph (hand, wrist)
16	Paediatric bone age assessment using deep convolutional neural networks [23]	Iglovikov, V.I. et al.	Lecture Notes in Computer Science	2018	Automatic	Hand, wrist	Radiograph
17	Regression Convolutional Neural Network for Automated Pediatric Bone Age Assessment from Hand Radiograph [24]	Ren, X. et al.	IEEE journal of biomedical and health informatics	2018	Automatic	Hand, wrist	Radiograph
18	Relative Skeletal Maturation and Population Ancestry in Nonobese Children and Adolescents [40]	McCormack, S.E. et al.	Journal of Bone and Mineral Research	2017	Non-Automatic	Hand, wrist	Radiograph
19	Simplified radius, ulna, and short bone-age assessment procedure using grouped-Tanner-Whitehouse method [38]	Hsieh, C. et al.	Pediatrics International: Official Journal of the Japan Pediatric Society	2011	Non-Automatic	Hand, wrist	Radiograph
20	Skeletal age estimation in a contemporary Western Australian population using the Tanner-Whitehouse method [33]	Maggio, A. et al.	Forensic Science International	2016	Non-Automatic	Hand, wrist	Radiograph

21	Skeletal maturation in Indonesian and white children assessed with hand-wrist and cervical vertebrae methods [37]	Soegiharto, B.M. et al.	American Journal of Orthodontics and Dentofacial Orthopedics	2008	Non-Automatic	Hand, wrist	Radiograph
22	Support vector machine classification based on correlation prototypes applied to bone age assessment [25]	Harmsen, M. et al.	IEEE journal of biomedical and health informatics	2013	Automatic	Hand	Radiograph
23	The BoneXpert method for automated determination of skeletal maturity [17]	Thodberg, H.H. et al.	IEEE transactions on medical imaging	2009	Automatic	Hand, wrist	Radiograph
24	The persistence of epiphyseal scars in the distal radius in adult individuals [41]	Davies, C. et al.	International Journal of Legal Medicine	2016	Non-Automatic	Wrist	Radiograph
25	Versatile Framework for Medical Image Processing and Analysis with Application to Automatic Bone Age Assessment [22]	Zhao, C. et al.	Journal of Electrical and Computer Engineering	2018	Automatic	Hand	Radiograph
26	What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents [19]	Urschler, M. et al.	Annals of Human Biology	2015	Automatic	Hand, wrist	MRI

BAA, Bone Age Assessment; ROI, Region of Interest; MRI, Magnetic Resonance Imaging.

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-5
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	5
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	7
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	7,9
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	8
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	7,8
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	8-10
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	10
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	10
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	10
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	10,11
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	11

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	10
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	11
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	11
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	12-26
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	12
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	12-26
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	12-26
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	12
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	24-26
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	27
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	29
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	27-29
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	30

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: www.prisma-statement.org.

Study II

14 Predicting chronological age in living individuals with machine learning methods using bone age assessment staging and non-radiological aspects in a multifactorial approach

Original paper

Predicting chronological age in living individuals with machine learning methods using bone age assessment staging and non-radiological aspects in a multifactorial approach

Ana Luiza Dallora , Ola Kvist, Johan Sanmartin Berglund, Sandra Diaz Ruiz, Martin Boldt, Carl-Erik Flodmark, Peter Anderberg

Abstract

Background: Bone age assessment (BAA) is used in numerous pediatric clinical settings, as well as in legal settings when entities need an estimate of chronological age (CA) when valid documents are lacking. The latter case presents itself as critical since the law is harsher for adults and granted rights along with imputability changes drastically if the individual is a minor. Traditional BAA methods suffer from drawbacks such as exposure of minors to radiation, do not consider factors that might affect the bone age and they mostly focus on a single region. Given the critical scenarios in which BAA can affect the lives of young individuals it is important to focus on the drawbacks of the traditional methods and investigate the potential of estimating CA through BAA.

Objective: This paper aims to investigate CA estimation through BAA in young individuals of 14 to 21 years with machine learning methods, addressing the drawbacks in the research using magnetic resonance imaging (MRI), assessment of multiple ROIs and other factors that may affect the bone age.

Methods: MRI examinations of the radius, distal tibia, proximal tibia, distal femur and calcaneus were carried out on 465 males and 473 females subjects (14-21 years). Measures of weight and height were taken from the subjects and a questionnaire was given for additional information (self-assessed Tanner Scale, physical activity level, parents' origin, type of residence during upbringing). Two pediatric radiologists assessed, independently, the MRI images as to their stage of bone development (blinded to age, gender and each other). All the gathered information was used in training machine learning models for chronological age estimation and minor versus adults classification (threshold of 18 years). Different machine learning methods were investigated.

Results: The minor versus adults classification produced accuracies of 90% and 84%, for male and female subjects, respectively, with high recalls for the classification of minors. The chronological age estimation for the eight age groups (14-21 years) achieved mean absolute errors of 0.95 years and 1.24 years for male and female subjects, respectively. However, for the latter lower error occurred only for the ages of 14 and 15.

Conclusions: This paper proposed to investigate the CA estimation through BAA using machine learning methods in two ways: minor versus adults classification and CA estimation in eight age groups (14-21 years), while addressing the drawbacks in the research on BAA. The first achieved good results, however, for the second case BAA showed not precise enough for the classification.

Keywords: chronological age assessment; bone age; skeletal maturity; machine learning; magnetic resonance imaging; radius; distal tibia; proximal tibia; distal femur; calcaneus

1. Introduction

1.1 Background

Skeletal maturity is a radiological concept that refers to the stage of bone development of an individual [1]. This maturation process happens gradually in the growth plates and it is measured by the degree of mineralization of the bone along with its size and shape [1]. Bone age (BA) is a closely related concept in which an age is estimated based on the degree of skeletal maturity of an individual [2].

The estimation of the BA of an individual, or bone age assessment (BAA), is performed in numerous clinical settings involving diagnosis and time of treatment of orthopedics, orthodontics, endocrinology, growth disorders, and estimations of final height [3]. In these cases, the BA of an individual is assessed by medical professionals and compared to their chronological age (CA). If found to be relatively advanced or retarded, appropriate actions are taken by the medical professionals.

BAA is also performed outside the clinical settings when legal entities need an estimation of the CA of an individual for judicial decisions when valid documents are lacking. This refers to cases regarding adoption, criminal proceedings, and pedopornography judicial issues, as well as in determining age fraud in youth sports competitions [4–7]. Furthermore, with the upsurge of immigration due to the rise of worldwide conflicts, another critical scenario in which BAA is applied concerns the determination of an individual being minor in the absence of valid or trustworthy

documents. This is the case of numerous young asylum seekers, who are given special rights granted by the United Nations Convention on the Rights of the Child, regarding reception, healthcare, education, etc [8,9].

From all of these examples, it is possible to assume that especially regarding legal standpoints, BAA is a crucial tool for making high stake decisions that have the potential to greatly affect individuals' lives.

1.2 Traditional bone age assessment

The traditional methods for BAA are based on the appearance of growth plates through the analysis of diaphyses (primary ossification centers) and epiphysis (secondary ossification centers), where cartilage tissue gradually turns into bone tissue during the process of bone development. A process that ceases when the diaphysis and epiphysis are fused indicating that the growth plate is ossified [1].

The most common procedures for BAA are the Greulich-Pyle (GP) and Tanner-Whitehouse (TW) methods. Both these methods assess radiograph images of the hand and wrist areas since these are regions of interest (ROI) with a large number of ossification centers aggregated in a small area that is easy to have images taken from.

The GP method [10] attributes a bone age by comparing the radiograph image of the individual being assessed to the nearest reference image in a hand and wrist atlas in terms of bone development. The TW method [11] is a scoring system that evaluates the ulna, radius, carpals and 13 short bones of the hand. Scores are attributed to these regions based on the stage of development of the bones, which ranges from A to I. Then, the scores are aggregated in a total score that is converted into the bone age.

Having been developed in the 30s and 50s, the GP and TW methods conveyed groundbreaking developments in numerous clinical settings and are still heavily employed for BAA purposes to this day.

1.3 Other proposed bone age assessment methods

The field of BAA evolved since the GP and TW methods were proposed, exploring new ROI with different ossification timings. This section summarizes proposed studies regarding BAA in various ROIs.

Newer hand and wrist studies on BAA include the Gilsanz and Ratib [1] digital hand atlas and the FELS method [12]. The first is composed of artificially created reference images that represent the average development of 29 classes of subjects from 0 to 18 years old. The FELS method [12] is a statistical method that provides a relative

measure of the BA and standard error that takes in consideration the distribution of chronological ages in the study's sample with BA similar to the individual being assessed and it is based on 98 indicators of bone maturity (ossification, radiopaque densities, bony projection, shape changes and ossification of epiphyses).

Clavicle staging systems observe one or both sides of the medial clavicular epiphysis. The method proposed by Kreitner et al. [13] presents four stages of ossification of the medial clavicular epiphysis, in which the last stage may have the epiphyseal scar visible. Schmeling et al. [9] propose five stages of ossification, but the last stage is only achieved when the epiphyseal scar is not apparent. Kellinghaus et al. [14] build on the Schmeling et al. staging [9] by applying sub-classifications for the second and third stages. These studies report complete ossification of this growth plate around the ages of 26-27 years.

Knee studies propose staging systems that also vary on subscales on specific stages and the appearance of the epiphyseal scar on the last stage. O'Connor et al. [15] propose five stages of ossification of the distal femur, proximal tibia and proximal fibula epiphyses (the epiphyseal scar may be visible in the last stage). Dedouit et al. [16] propose five stages of ossification of the distal femur and proximal tibia epiphysis, assessing the appearance of the cartilage signal intensity with magnetic resonance imaging (MRI). Krammer et al. [17] propose five stages of ossification of the distal femur epiphysis, with sub-classifications on the second and third stages, with the last stage achieved only when the epiphyseal scar is no longer visible. This method also makes use of MRI images. Knee studies usually argue about a subject being younger or older than the age of 18 years old.

Studies on foot ROI are usually concerned with younger ages. Ekizoglu et al. [18] propose a staging system for the foot ROI that shows complete ossification in ages between 12 and 16.

Not very much explored in the literature, the arm ROI is studied in the proximal humerus epiphysis by Ekizoglu et al. [19] employing a scoring system based on Schmeling et al. [9] and Kellinghaus et al. [14] on MRI images. This study points out the earliest ages for the last stage of ossification at 17 and 18 years.

1.4 Drawbacks in assessing chronological age using bone age assessment methods

In the lack of valid or trustworthy documents, BAA is employed nowadays as a valuable tool for legal entities to evaluate CA in regards to important legal ages. Nevertheless, it is possible to identify several drawbacks about the largely employed

GP and TW methods, as well as recently proposed methods, regarding the use of BAA for CA determination:

- they almost exclusively employ medical imaging techniques that expose the individual to ionizing radiation, such as radiographs, which raises grave ethical issues especially in regards to exposing minors to radiation for non-therapeutic purposes;
- they only focus on the physical appearance of the growth plates, not including other information that might possibly affect the bone development [20];
- they mostly focus on a single ROI, which in the vast majority of cases is the hand area [20].

The first drawback can be addressed by the employment of the MRI technology, which is already present in some of the mentioned knee and arm studies. Besides being a radiation-free modality of medical imaging, it also allows the manipulation of contrast in order to highlight different tissue types [21]. The epiphyseal plate consists of cartilage tissue, which is mainly composed of collagen fiber protein. Collagen has a 3D structure of fibers that, in MRI images is shown as zones of different intensities, giving it a multilaminar appearance. It is known that the structure of cartilage changes in terms of the number of laminae and thickness in the course of bone development [22]. Hence, contrary to radiographs that highlight the bone, the MRI technology might have the potential to offer better visualization of growth plates, thus being an interesting radiation-free modality of medical imaging for BAA.

To address the second drawback, the methods for assessing BA should investigate factors that may play a role in the process of bone development and ossification of growth plates, i.e. BMI [20,23], pubertal growth [24], physical activity [25], ethnicity [8,20,26] and socioeconomic factors [8], that are often overlooked [20].

Addressing the third drawback could be done by the employment of multiple ROI. When it comes to estimations of CA, most of the BAA studies, especially methods that propose stages of maturity for set ROIs, follow an approach of identifying the minimum age in which the ossification of the growth plate is completed for a particular ROI. These studies usually focus on a single age of legal importance, which varies a lot between countries, comprehending ages from 14 to 21 years [13]. Using multiple ROI may provide more information about more ages.

One additional drawback that is specific about the GP and TW methods is that they are based on data collected from subjects of average and upper socioeconomic classes in the 30s and 50s, respectively. Hence, these methods may not reflect on secular trends that nowadays point to higher height and earlier puberty [27], which could affect the accuracy of the methods. For the TW method, an update released in 2001

(TW3) revised the calculation of the BA from the attributed scores in order to address this problem [28].

1.5 Machine learning for bone age assessment

From the presented drawbacks, it is noticeable that the BAA research could benefit from methods that are able to aggregate multiple information (i.e. multiple ROI and factors) in a systematic way. A technology that is able to work in this setting is machine learning (ML). ML consists of various types of algorithms that are able to learn how to perform a task from a set of examples while improving its performance based on its experience in carrying out the particular task. It builds a model that encapsulates the knowledge to perform the task, then in light of new data, the model is able to correctly perform the learned task within an acceptable measure of performance [29].

ML algorithms have already been employed in various models for assessing the BA of an individual. A recent systematic literature review on BAA with ML methods [20] showed that the research is heavily focused on models that make use of a single ROI, the hand in most of the cases, having radiographs as the choice of imaging technology, and do not usually consider other factors that could play a role in bone development [20]. The most notable commercially available ML BAA system is the BoneXpert [30], which performs an automatic radiograph analysis based on the GP and TW methods. However, it covers the age range of 2 to 17 years and leaves out important legal ages.

1.6 Objectives of the study

Given the importance of the assessment of CA through BAA in numerous scenarios and its potential ways it could affect the lives of young individuals, it is important to focus on the drawbacks of the methods currently in use and investigate the potential of BAA in estimating CA. Thus, the objectives of this study are:

- to investigate to what extent ML models can aid in the CA estimation through BAA in young individuals of 14 to 21 years;
- to investigate whether ML models can aid in the determination of minors thought BAA, considering the threshold of 18 years, in young individuals of 14 to 21 years;
- to address the drawbacks in the research on CA estimation from BAA, in regards to using radiation-free medical imaging technology, the assessment of multiple ROIs and other factors that may play a role in bone development.

2. Materials and methods

2.1 Overview

In order to train the chronological age estimation ML models proposed in this paper, MRI images of the wrist, knee and foot were taken from volunteer subjects and assessed by radiologists as to their stage of bone development. The five growth zones considered in this study were: Calcaneus, Distal Tibia, Proximal Tibia, Distal Femur, and Radius. Each growth zone was assessed separately and blinded to gender and age.

Before the examination, the subjects had their height and weight measured for the BMI calculation and were asked to answer a questionnaire to gather information on their physical activity level, parents' origin, type of residence during upbringing, and a self-assessed Tanner Scale of pubertal growth [31,32].

All the radiological and non-radiological data gathered were used to train binary and multi-class classifiers. For the binary classifier, the individuals in the sample were divided into minors or adults, regarding the threshold of 18 years, and the classification followed into discriminating individuals into one of the two classes. The multi-class classifier aims to classify an individual into one of the eight classes defined by the age groups ranging from 14 to 21 years.

The remainder of this section details further the population, data used in the experiments, statistical analysis and procedures for model building in the experiments.

2.2 Recruitment

This study, prospectively, conducted MRI examinations on 938 healthy subjects (465 males and 473 females) with ages between 14 to 21 years old (inclusive), during 2017 and 2018. The participants of the study had images taken from the knee, foot, and wrist on the same examination session. Additionally, the weight and height of each participant were also collected to calculate the BMI.

The following criteria were used to determine participation in the study:

- Inclusion criteria: the participants should have been born in Sweden, where the study was conducted, and have a birth certificate verified by the Swedish national authorities.
- Exclusion criteria: a history of bilateral fractures or trauma near the regions of assessment, a history of chronic disease or the use of long-term medications, noncompliance during the examination, having resided outside

Sweden for more than six consecutive months, past or current pregnancy (all female subjects were tested).

2.3 Data privacy and study ethics

The study was conducted in accordance with the Declaration of Helsinki and was approved by the Central Ethical Review Board in Stockholm (diary numbers: 2017/4-31/4, 2017/1184-32, 2017/1773-32). Written informed consent was collected from all subjects and legal guardians (in the case of subjects younger than 18 years old).

All data was anonymized and stratified by age and gender.

2.4 Population

A total of 455 male and 467 female subjects constituted the final sample (Table 1). After the MRI examinations and assessment of images by radiologists, 10 male, and 6 female subjects were removed from the study's sample because they had the assessment of one or more ROI missing. The missing values for the assessment by the radiologists could be due to one of the following reasons: movement artifact, error in the sequence that made the image non-gradable, likely trauma in the region of assessment and missing MRI examination in one or more ROI.

Table 1. Demographics of the final sample

Age Group	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**Total**
Number of female subjects	59	58	57	60	59	57	57	60	**467**
Number of male subjects	58	56	60	58	53	58	53	59	**455**

2.5 Data and data collection procedures

The data used to train the classifiers were the radiologists' assessment of the Calcaneus, Distal Tibia, Proximal Tibia, Distal Femur, and Radius growth zones; the additional information gathered before the examination which was: physical activity level, parents' origin, type of residence during upbringing, and a self-assessed Tanner Scale of pubertal growth; and the BMI. The following section details the data and the procedures for collection.

2.5.1 Magnetic resonance imaging examinations

MRI examinations were carried out to take images of the Calcaneus, Distal Tibia, Proximal Tibia, Distal Femur and Radius growth plates of the subjects participating

in the study. All MRI examinations were conducted within six months of the subjects' birthday date on 1.5-T whole-body MRI scanners with dedicated hand, knee and ankle coils. The examinations were performed on the non-dominant side of the knee, hand, and foot, save when past fracture or trauma had taken place nearby the region. In these cases, the dominant side was imaged. The images of all ROIs were taken on the same examination session.

The examinations were carried out in two sites. Site 1 used Magnetom Avanto Fit (Siemens Healthcare Gmbh, Erlangen, Germany) and Achieva (Philips Healthcare, Amsterdam, The Netherlands) whole-body scanners; and Site 2 used a Signa (GE Healthcare, Milwaukee, Wisconsin) whole-body scanner. All examinations followed the same protocol which included a T2 sequence with cartilage dedicated exposure. The settings were: 256 x 256 pixel resolution and 160 by 160 mm field of view.

2.5.2 Assessment of the Magnetic resonance images

The assessment of the MRI images was done independently by two radiologists with 3 and 30 years of experience in pediatric radiology, who were blinded to the age and gender of the participants. A third radiologist with 13 years of experience in pediatric radiology assessed the images when the first two radiologists couldn't reach a final agreement about the stage.

The staging system used to assess the MRI images is a version of the staging methods proposed by Dedouit et al. [16] and Kellinghaus et al. [14] with minor modifications. This staging is defined as follows:

- **Stage 1:** Continuous, stripe-like, cartilage signal intensity present between the metaphysis and the epiphysis with a thickness greater than 1.5mm with a multilaminar appearance.
- **Stage 2:** Continuous cartilage signal intensity present between the metaphysis and the epiphysis with a thickness greater than 1.5 mm with increased signal intensity but without a multilaminar appearance.
- **Stage 3:** Continuous cartilage signal intensity present between the metaphysis and the epiphysis with a thickness of less than 1.5mm with increased signal intensity.
- **Stage 4a:** Non-continuous cartilage signal intensity. A hazy area involving one third or less of the growth plate is present between the metaphysis and the epiphysis, representing the epiphyseal-metaphyseal fusion
- **Stage 4b:** Non-continuous cartilage signal intensity. A hazy area involving between one third and two-thirds of the growth plate is present between the metaphysis and the epiphysis, representing epiphyseal-metaphyseal fusion.

- **Stage 4c:** Non-continuous cartilage signal intensity. A hazy area involving more than two-thirds of the growth plate is present between the metaphysis and the epiphysis, representing epiphyseal-metaphyseal fusion.
- **Stage 5:** The epiphyseal cartilage has fused completely with or without an epiphyseal scar, in all MRI slices.

2.5.3 Questionnaire Information

The additional information regarding pubertal growth, physical activity, parents' origin and socioeconomic factors from the participants was gathered by a questionnaire, given to the subject at the examination session. Table 2 shows the gathered information.

Table 2. Summary of the information gathered by the questionnaire

Variable	Description	Values
Residence	Type of residence the participant lives in (or lived during upbringing).	Rented; Owned
Physical Activity	The participants' daily level of activity.	Highly Inactive; Inactive; Little Active; Active; Highly Active
Parent Origin	Origin of the participants' parents, regarding if they were born outside Sweden or not.	no foreign-born parents; one foreign-born parent; both foreign-born parents
Tanner Scale	Self-assessed Tanner Scale for pubertal growth [31,32].	Stage 1; Stage 2; Stage 3; Stage 4; Stage 5

2.5.4 Body Mass Index

The Body Mass Index (BMI) was calculated with the measures of the participants' weight *w* and height *h* as in the formula [33]:

$BMI = \frac{w}{h^2}$ (1).

2.6 Statistical analyses

Cohen's kappa coefficient [34] was calculated to measure the inter-observer agreement between the pediatric radiologists in all investigated ROIs. Statistical analyses were performed in the IBM SPSS Statistics[3] software platform.

[3] IBM SPSS Statistics version 24 (release 24.0.0.0). URL: https://www.ibm.com/analytics/spss-statistics-software

2.7 Model building

2.7.1 Chronological age estimation models

In this study, various ML algorithms were investigated in order to build classifiers to discriminate subjects into minor or adults, and classifiers to classify subjects into one of eight age groups (14 to 21 years). Models for male and female subjects were built separately.

2.7.2 Data preparation

A K-nearest neighbor's (KNN) multiple imputations were applied to entries with missing data. This technique finds the K complete entries that are the closest to an incomplete entry (i.e. contains missing data) and fills its missing values with the mean (in case of numeric variables), or the most frequent one (in case of categorical variables) [35]. In this study, the number of nearest neighbors K for the KNN imputation was set to 1. The motivation for this choice is based on literature findings that advise limiting K as a way to preserve the original variability of the data, reducing the risk of entries having few neighbors that are too distant from each other. [36].There is also a risk of increasing the influence of noise in the data with a small K, but since in the dataset of this study, the highest rate of imputed instances was 1.9%, this influence was considered not very relevant. The number of imputed instances for each variable in both male and female subsets is shown in Table 3.

Table 3. Number of imputed instances and percentage over the male and female datasets.

Variable	Male dataset	Female dataset
Radiologists' assessments of the Radius, Distal Femur, Proximal Tibia, Distal Tibia, Calcaneus	0 (0%)	0 (0%)
Residence	1 (0.2%)	3 (0.6%)
Physical Activity	9 (1.9%)	6 (1.2%)
Tanner Scale	3 (0.6%)	1 (0.2%)
BMI	0 (0%)	0 (0%)
Parents Origin	0 (0%)	3 (0.6%)

2.7.3 Machine learning algorithms

The choice of the ML algorithms explored in this study was based on the summary of the evidence of a recently published systematic literature review (SLR) on the application of ML for BAA [20]. This SLR points out that the studies proposing BAA classifiers employ algorithms of the following categories: Artificial Neural Networks,

Support Vector Machines, Bayesian Networks, Decision Trees and K-Nearest Neighbors. An additional search was conducted in the literature (Scopus[4], PubMed[5] and Web of Science[6]), after the search date of the mentioned SLR [20] (February of 2019) to look for additional algorithms, but no new categories were found to be added to the list.

Another motivation for this choice of ML algorithms is that it also guarantees a diversified list of classifiers that make use of different types of learning techniques, such as rule-based, instance-based, bayesian inference, kernel and perceptron learners. We refer to the following book by Kuhn and Johnson [37] for the specific algorithms and implementations used in this study.

Therefore, the choice of ML algorithms for the experiments of the present study includes: Decision Tree, Random Forest, Multi-layer Perceptron, Support Vector Machines, Naïve Bayes, K-Nearest Neighbors.

2.7.4 Experimental setup

All experiments were performed using stratified nested cross-validation [38]. In this approach, in each iteration, one fold of the outer cross-validation is used for testing and the remaining four are used in an inner cross-validation for tuning the algorithms' hyperparameters. This was done in order to get a more reliable estimate of the error, since the test fold in each outer iteration is not being used to execute performance optimization [39]. Also, it is worth noting that the data splits were done in a stratified manner, which means that the classes' proportions in each split are kept the same as in the total sample. In the experiments of this study a 5-fold outer, 3-fold inner stratified nested cross validation was performed.

Additionally, before each inner cross validation, a grid search was performed in order to find suitable hyperparameters for each of the selected ML algorithms. The hyperparameters for each selected algorithm is shown in Table 4. The machine learning experiments were conducted in the R framework[7] with the *caret* package. The default version of the algorithms were used.

[4] Scopus, URL: https://www.scopus.com/
[5] Pubmed, URL: https://www.ncbi.nlm.nih.gov/pubmed/
[6] Web of Science, URL: https://www.webofknowledge.com/
[7] R version 3.4.1. URL: https://www.r-project.o

Table 4. Configuration of the R algorithms included in the experiment.

ML Algorithm	R Implementation	Tuning Parameters
Decision Tree	*rpart*	*cp*
Random Forest	*rf*	*mtry*
Multi-layer Perceptron	*mlp*	*size*
Support Vector Machines	*svmRadial*	*Sigma, C*
Naïve Bayes	*nb*	*fL, usekernel, adjust*
K-Nearest Neighbors	*knn*	*k*

2.8 Model evaluation metrics

The performance metrics used to evaluate the models were the following: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Accuracy, Precision, Recall and Area Under the Curve (AUC), as in Gaudette and Japkowicz (2009) [40]; and Sokolova and Lapalme (2009) [41] guidelines for ordinal multiclass classification. For the binary classification models all but MAE and RMSE are used. The standard deviations for each metric are also reported.

The mean absolute error (MAE), represents the mean of the absolute difference between the estimated age output of the classifier and the correct chronological age of the subject, over all examples. The RMSE gives more weight to larger errors compared to MAE, which tends to prefer fewer errors overall. The MAE and RMSE are calculated as follows in equations (2) and (3), respectively:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y - \hat{y}| \quad (2)$$

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(y - \hat{y})^2}{n}} \quad (3)$$

with n being the number of samples, $\hat{y}$ being the estimated age, and y being the chronological age of the subject.

For the remaining evaluation metrics, considering l the number of classes, we define:

- True positives (TP): Entries predicted to be in class C_l, actually in class C_l.
- False Positives (FP): Entries predicted to be in C_l, but are not actually in class C_l.
- True Negatives (TN): Entries not predicted to be in C_l and are not actually in class C_l.
- False Negatives (FN): Entries not predicted to be in C_l ,but are actually in class C_l.

The Accuracy, Precision, Recall and AUC, for binary classification, are calculated as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (4)$$

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (7)$$

In the case of the multi-class classification, these are calculated as the average, calculated for each class C_l [41].

General results are given for the ML algorithms in terms of the mean and standard deviations of each of the performance metrics for the outer cross-validation test sets. In depth results are given to the best performing models.

3. Results

3.1 Inter-observer agreement

The Kappa Cohen's coefficient was calculated to evaluate the agreement between the two observers' assessment of the MRI images. The results pointed to substantial agreement according to the general guidelines [42] for all of the assessed ROI: 0.77 for the Calcaneus, 0.65 for the Distal Femur, 0.72 for the Distal Tibia, 0.73 for the Proximal Tibia and 0.67 for the Radius.

3.2 Results for the growth plate assessments

The results from the assessments of the Calcaneus, Distal Tibia, Proximal Tibia, Distal Femur and Radius, for male and female subjects are shown in detail in the Multimedia appendix 1 and 2, respectively.

In all of the assessed growth plates, for both sexes, stages 1 and 2 were not evidenced. Few instances of stage 3 were evidenced on male subjects on the Calcaneus and Radius growth plates, accounting for 2 and 15 cases, respectively. On female subjects, stage 3 was evidenced in only two cases for the Radius growth plate.

The female subjects' results show that for all assessed growth plates, nearly all or most of the sample was already on the last stage of ossification (stage 5): 94.6% of Calcaneus, 90.8% of Distal Tibia, 81.6% of Proximal tibia, 74.5% of Distal femur,

and 65.5% of Radius cases. These numbers moderately change for male subject's, accounting for: 80.4% of Calcaneus, 70.1% of Distal tibia, 57.6% of Proximal Tibia, 54.9% of Distal Femur, and 47.4% of Radius cases.

Table 5 shows the proportion, within each age group, of subjects who have all of the growth plates considered in this study already in stage 5. This table shows that female subjects had all growth plates fused two years prior to the male subjects. For female subjects, from the age of 19, all subjects of the sample have already all of the growth plates fused, while for male subjects the same happens from the age of 21.

Table 5. Numbers and percentages (over each age group) of subjects with all of the growth plates in stage 5, for male and female subjects.

Age group	Female subjects	Male subjects
14	2 (3.3%)	0 (0%)
15	8 (13.7%)	0 (0%)
16	23 (40.3%)	3 (5%)
17	44 (73.3%)	13 (22.4%)
18	53 (89.8%)	31 (58.4%)
19	57 (100%)	50 (86.2%)
20	57 (100%)	50 (94.3%)
21	60 (100%)	59 (100%)
Total	**304 (65.1%)**	**206 (45.2%)**

3.3 Results for the classification of minors versus adults

The threshold of 18 years was used to determine adulthood in the classification of minors versus adults, which is the case in many European countries. MAE and RMSE were not used as performance metrics in this case because for classifications they only make sense in the context of an ordinal classification. The results for the male subjects binary classifiers in terms of the mean and standard deviation of the performance metrics on the outer cross-validation test sets is shown in table 6. The Decision Tree, Random Forest and Support vector Machines algorithms had very similar results in general. The Random Forest algorithm was chosen in terms of the best combination of Precision and Recall, but in practical settings there are no differences between these algorithms. Table 7 shows the Random Forest's results in each of outer cross-validation test sets. The average model was chosen in terms of the median accuracy which was 0.90. Between models 1 and 4, the Model 1 was chosen for better recall in classifying minors. The optimized hyperparameter given by the

grid search for Model 1 was *mtry = 2* (number of candidate variables at each tree split).

Table 6. Mean performance metrics and respective standard deviations, in years, for the classification of minor vs adults for the male subjects.

Types	Accuracy	AUC	Precision	Recall
Decision Tree	**0.90 ± 0.02**	**0.90 ± 0.02**	0.86 ± 0.04	0.96 ± 0.03
Random Forest	**0.90 ± 0.01**	**0.90 ± 0.01**	**0.87 ± 0.03**	0.94 ± 0.04
Support vector Machines	**0.90 ± 0.02**	**0.90 ± 0.02**	**0.87 ± 0.04**	0.93 ± 0.07
Multi-Layer Perceptron	0.82 ± 0.17	0.82 ± 0.16	0.79 ± 0.16	0.95 ± 0.04
K-Nearest Neighbors	0.87 ± 0.02	0.87 ± 0.02	0.84 ± 0.03	0.92 ± 0.03
Naïve Bayes	0.73 ± 0.04	0.74 ± 0.04	0.65 ± 0.03	**1.00 ± 0.00**

Table 7. Performance results for the Random Forest algorithm on each of the outer cross-validation test sets, for the male sample.

Model	Accuracy	AUC	Precision	Recall - Minors	Recall - Adults
1 (median)	**0.90**	**0.90**	**0.83**	**1.00**	**0.80**
2	0.89	0.89	0.87	0.91	0.87
3	0.89	0.89	0.87	0.96	0.83
4	0.90	0.90	0.90	0.89	0.91
5	0.92	0.92	0.89	0.96	0.89

Figure 1 presents the results achieved by the Model 1, per age group. It is important to notice that even with low Accuracy results for the age of 17 (41.7%), the model still minimizes the error of classifying minors as adults, achieving a Recall of 100% for this classification.

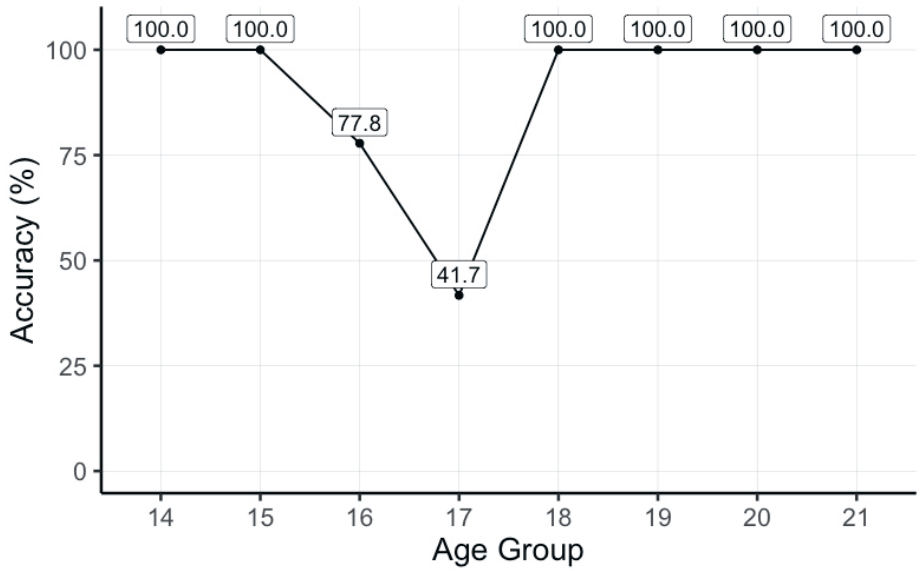


Figure 1. Accuracy per age group for the minor vs adults classification model, for male subjects.

In the female subjects case, the algorithm that achieved the best mean accuracy was also the Random Forest (see table 8). Table 9 shows the results for the Random Forest algorithm in each of the outer cross-validation test sets. Except for the model 1, essentially there was no relevant variation between models, and in practical settings they can be considered equal. Thus, Model 2 was chosen as the average model. The optimized hyperparameter given by the grid search for Model 1 was *mtry = 6*.

Table 8. Mean performance metrics and respective standard deviations, in years, for the classification of minor vs adults for the female subjects.

Types	Accuracy	AUC	Precision	Recall
Decision Tree	0.82 ± 0.02	0.82 ± 0.02	0.74 ± 0.02	0.97 ± 0.01
Random Forest	**0.83 ± 0.02**	**0.83 ± 0.01**	**0.76 ± 0.02**	0.97 ± 0.01
Support vector Machines	0.81 ± 0.04	0.81 ± 0.04	0.75 ± 0.04	0.92 ± 0.05
Multi-Layer Perceptron	0.82 ± 0.02	0.82 ± 0.02	0.75 ± 0.02	0.95 ± 0.04
K-Nearest Neighbors	0.78 ± 0.06	0.78 ± 0.06	0.73 ± 0.06	0.87 ± 0.08
Naïve Bayes	0.67 ± 0.03	0.67 ± 0.02	0.60 ± 0.02	**1.00 ± 0.00**

Table 9. Performance results for the Random Forest on each of the outer cross-validation test sets, for the female sample.

Model	Accuracy	AUC	Precision	Recall - Minors	Recall - Adults
1	0.81	0.81	0.73	0.96	0.66
2 (median)	**0.84**	**0.84**	**0.77**	**0.96**	**0.72**
3	0.84	0.84	0.77	0.98	0.70
4	0.84	0.84	0.78	0.96	0.72
5	0.84	0.84	0.77	0.98	0.70

The accuracies per age group are shown in the graph of figure 2. The model achieves lower accuracies for the ages of 16 and 17 (50.0% and 58.3%, respectively), but as in the male subjects case, the model minimizes the worst type of error that is the misclassification of minors, achieving a high recall of 96%.

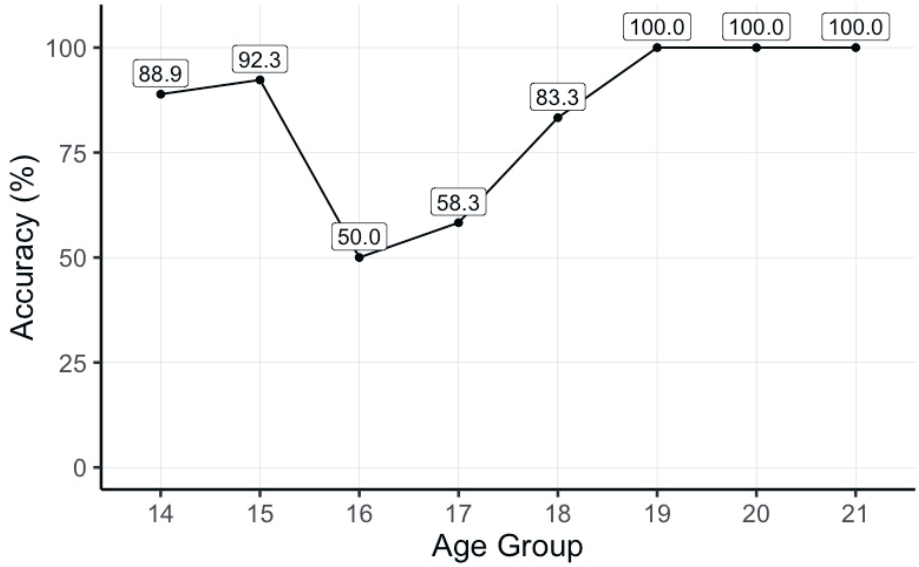


Figure 2. Accuracy per age group for the minor vs adults classification model, for female subjects.

3.4 Results for the chronological age estimation models

The chronological age estimation models are multi-class classifiers that aim at classifying subjects in one of the eight age groups (from 14 to 21 years). Table 10 shows the results for the male subjects models in terms of the mean and standard deviations of the performances on the outer cross-validation test sets. The best performing algorithm in the male case was the Multi-Layer Perceptron (MLP), which achieved the best mean MAE (0.98 years), mean RMSE (1.32 years) and mean precision (0.65 years) values, in addition to having the second best values of mean accuracy and mean AUC.

Table 10. Mean ± standard deviations of the performance metrics, for the male subjects' classification models.

Algorithm	MAE (years)	Accuracy	RMSE (years)	AUC	Precision	Recall
Decision Tree	1.28 ± 0.13	0.32 ± 0.03	1.78 ± 0.17	0.81 ± 0.02	0.49 ± 0.06	**0.81 ± 0.11**
Random Forest	1.04 ± 0.07	**0.34 ± 0.03**	1.44 ± 0.13	**0.85 ± 0.01**	0.57 ± 0.09	0.73 ± 0.14
Support Vector Machine	1.03 ± 0.09	**0.34 ± 0.03**	1.43 ± 0.08	**0.85 ± 0.01**	0.52 ± 0.09	0.67 ± 0.12
Multi-Layer Perceptron	**0.98 ± 0.08**	0.33 ± 0.02	**1.32 ± 0.13**	0.84 ± 0.01	**0.65 ± 0.27**	0.61 ± 0.31
K-Nearest Neighbor	1.16 ± 0.11	0.30 ± 0.04	1.57 ± 0.15	0.82 ± 0.03	0.59 ± 0.10	0.59 ± 0.10
Naïve Bayes	1.07 ± 0.10	0.29 ± 0.02	1.39 ± 0.19	0.81 ± 0.01	0.57 ± 0.06	0.58 ± 0.21

The performances for the MLP algorithm on each of the outer cross-validation test sets are shown on Table 11. The average model was chosen in terms of the median MAE which corresponds to the Model 1 with a value of 0.95 years. The optimized hyperparameter given by the grid search for the average MLP model was *size = 27* (number of units in the hidden layer). The average model was chosen in order to select an algorithm that would not be overly optimistic in its estimation.

Table 11. Performance results for the Multi-Layer Perceptron algorithm on each of the outer cross-validation test sets, for the male sample.

Model	MAE (years)	Accuracy	AUC	RMSE (years)	Recall	Precision
1 (median)	**0.95**	**0.33**	**0.83**	**1.29**	**0.91**	**0.48**
2	1.08	0.30	0.85	1.40	073	0.35
3	0.89	0.32	0.84	1.17	0.17	1.00
4	0.91	0.33	0.83	1.23	0.83	0.59
5	1.05	0.35	0.84	1.49	0.42	0.83

The results for the chosen model, discriminated by age groups, are shown in table 12 and the graph of figure 3. The model shows lower errors for the younger and older ages of the age spam considered in the study. Also, the model has a clear trend of overestimating the ages for the male subjects in general. Thus, even with a MAE of 0.95 years, the model is limited to its capacity of classifying individuals from the age of 16. From the age of 19, the model tends to classify all subjects as 20 years old since nearly all subjects of these ages have all growth plates on stage 5.

Table 12. Mean Absolute Error and standard deviation for the average male model

Age Group	14	15	16	17	18	19	20	21
MAE (years)	0.18 ± 0.60	0.82 ± 0.90	1.25 ± 1.44	1.91 ± 1.56	1.50 ± 1.45	1.00 ± 0.60	0.00 ± 0.00	0.92 ± 0.29

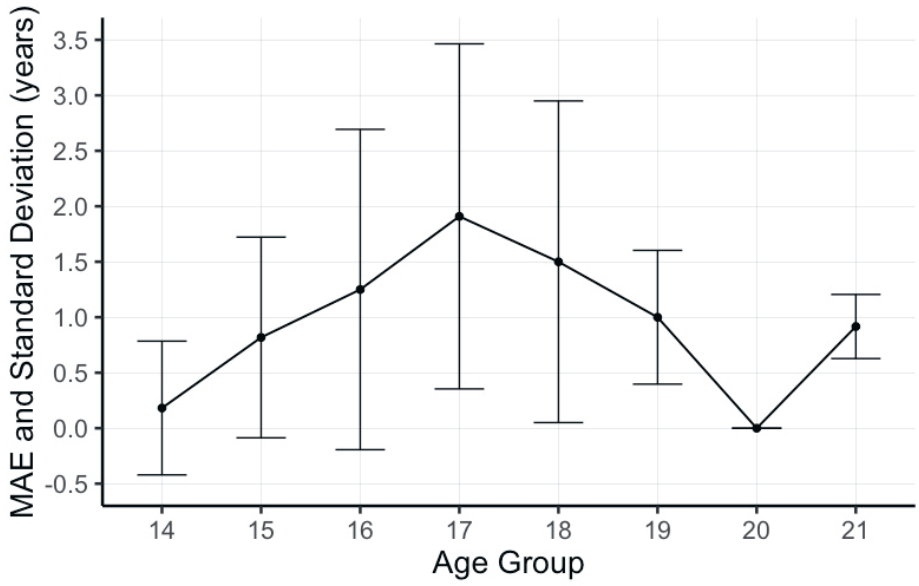


Figure 3. Mean absolute error (MAE) and standard deviation for the male Multi-Layer Perceptron model, in years.

Table 13 shows the results for the chronological age estimation models for female subjects in terms of the mean and standard deviations of the performances on the outer cross-validation test sets. In the female case, the best performing algorithm was the Support Vector Machine (SVM), which achieved the best mean MAE (1.21 years), mean Accuracy (0.32), mean RMSE (1.68 years) and mean AUC (0.80).

Table 14 shows the performance results for each of the outer cross-validation test sets for the SVM algorithm. For the female subjects case, the median resulted in a MAE of 1.24 years, which pertained to Models 1 and 2. Model 1 was chosen as the average model for presenting the best Accuracy between the two. The optimized parameter given by the grid search for the average SVM model was *sigma = 0.0421* (kernel parameter) *and C = 4* (penalty parameter).

Table 13. Mean ± standard deviations of the performance metrics, for the female subjects' classification models.

Algorithms	MAE (years)	Accuracy	RMSE (years)	AUC	Precision	Recall
Decision Tree	1.31 ± 0.09	0.28 ± 0.02	1.78 ± 0.13	**0.80 ± 0.02**	0.56 ± 0.05	**0.82 ± 0.09**
Random Forest	1.29 ± 0.10	0.30 ± 0.03	1.77 ± 0.10	0.79 ± 0.02	0.59 ± 0.13	0.74 ± 0.17
Support Vector Machine	**1.21 ± 0.06**	**0.32 ± 0.04**	**1.68 ± 0.06**	**0.80 ± 0.01**	0.55 ± 0.07	0.71 ± 0.11
Multi-Layer Perceptron	1.36 ± 0.24	0.30 ± 0.02	1.85 ± 0.37	0.77 ± 0.02	**0.60 ± 0.11**	0.63 ± 0.22
K-Nearest Neighbors	1.41 ± 0.12	0.30 ± 0.02	1.96 ± 0.12	0.76 ± 0.03	0.55 ± 0.07	0.61 ± 0.18
Naïve Bayes	1.74 ± 0.23	0.22 ± 0.02	2.23 ± 0.27	0.65 ± 0.03	0.58 ± 0.06	**0.82 ± 0.09**

Table 14. Performance results for the Support Vector Machine algorithm on each of the outer cross-validation test sets, for the female sample.

Model	MAE (years)	Accuracy (years)	AUC (years)	RMSE (years)	Recall (years)	Precision (years)
1 (median)	**1.24**	**0.37**	**0.79**	**1.75**	**0.75**	**0.56**
2	1.24	0.27	0.80	1.67	0.55	0.67
3	1.25	0.33	0.78	1.70	0.75	0.50
4	1.11	0.32	0.81	1.58	0.67	0.53
5	1.20	0.32	0.81	1.72	0.83	0.83

The MAE results per age group are shown in table 15 and in the graph of the figure 4. As in the male subjects case, the female model also overestimate the ages of female subjects in general, but with higher MAE and standard deviations.

Table 15. Mean Absolute Error and standard deviation for the male median model

Age Group	14	15	16	17	18	19	20	21
MAE (years)	0.42 ± 0.79	1.42 ± 1.93	1.17 ± 1.79	2.17 ± 2.49	1.75 ± 1.88	1.09 ± 1.43	0.90 ± 1.27	1.00 ± 1.34

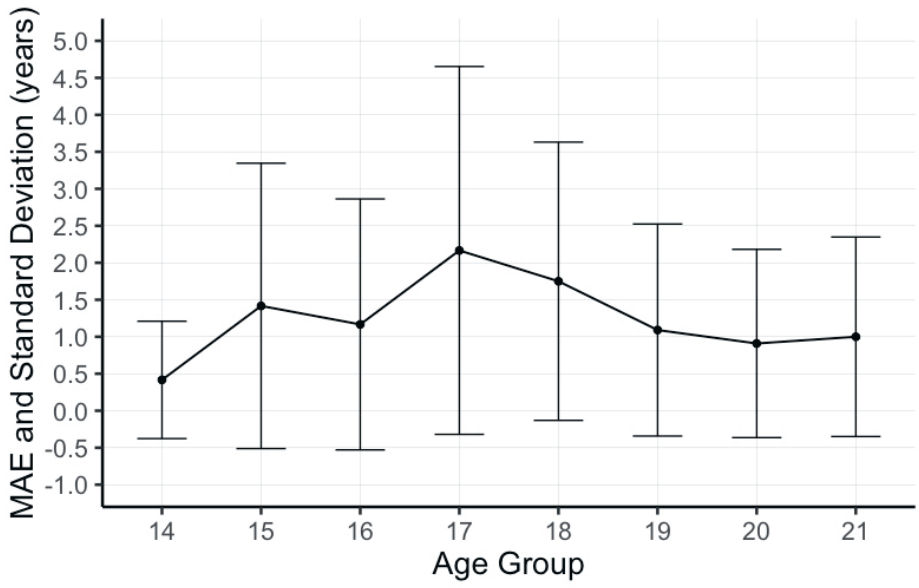


Figure 4. Mean absolute error (MAE) and standard deviation for the female Support Vector Machine model, in years.

4. Discussion

4.1 Principal Findings

This paper presented experiments with the estimation of chronological age and classification of minors versus adults (on the threshold of 18 years), of male and female subjects, using the ML algorithms. To build the models two radiologists assessed the stage of bone development of the Calcaneus, Distal Tibia, Proximal Tibia, Distal Femur and Radius growth plates of 455 male and 467 female volunteer subjects (922 subjects in total) from MRI images. Additional variables were also used to build the models: BMI, physical activity level, parents' origin, type of residence during upbringing, and self-assessed Tanner Scale of pubertal growth. The methodology adopted in the study aimed at addressing the drawbacks in the BAA methods that are employed in chronological age estimation for legal scenarios.

From the stage assessments of the MRI images we could infer that female subjects mature earlier than male subjects in regards to the bone development of the knee, wrist and foot, which is in line with prior studies [1,17,18,43,44]. In the present study,

the first age in which 100% of the sample had all fused growth plates (stage 5) was 19 years for female and 21 for male subjects.

Another important point to be discussed in regards to the stage assessments is that the female sample had cases which had all of the considered growth plates already fused since the age of 14, spamming throughout all ages considered in the study (14 to 21 years). Since the assessment of the stage 5, unlike the other stages, requires that all of the slices from the MRI examination to present a fused growth plate, even if there is a degree of misassessment, it would still mean that these cases would display a well advanced level of maturation in all of these ages, implying a high degree of biological variation in the female sample in regards to bone age. Additionally, in total 65.5% of the female sample consisted of cases in which the subjects presented all growth plates already in stage 5, which means that for classification purposes, for more than half of the sample the estimation of chronological age would depend only on the additional factors (self-assessed Tanner Scale, BMI, Residence Type, Physical Activity and Parents Origin) which were not enough to discriminate between age groups. This hindered classifiers' performance, especially the chronological age estimation models. The same phenomenon occurred for the male sample which also negatively affected the performance of the classifiers, but to a lesser degree, as 45.2% of the sample had all growth plates of stage 5, from the age of 16 to 21 years.

The minors versus adults classification achieved good accuracy results for both male (90%) and female subjects (84%). These models portrayed a drop in the performance for the ages of 16 and 17, however the recalls regarding the correct classification of minors were very high in both male and female cases (100% and 96% respectively). This is important because the problem of minors vs adults classification is asymmetric as the misclassification of minors for adults in a judicial scenario is much more problematic than the inverse. In most cases the application of the law is harsher for adults, and imputability along with granted rights can drastically change between these groups.

The chronological age estimation models achieved MAEs of 0.95 years and 1.24 years for male and female subjects, respectively. However, a look in depth of the models showed that for both male and female models, only the ages of 14 and 15 achieved acceptable MAE values. It could be argued that for the ages of 16 to 21 years, the estimation of a precise chronological age based on stages of bone development of the Calcaneus, Distal Tibia, Proximal Tibia, Distal Femur and Radius growth plates would be somewhat unfit for male individuals and very unfit for female individuals. Further, we could argue that staging may not offer a precise enough measure for the estimation of chronological age of individuals of the ages considered in this study.

Compared to dental age, height and age at menarche, bone age is still the most trustable biological indicator for assessing maturation in young individuals [45], but it may not be a strong predictor of CA. BAA was conceived to be used in conjunction with CA in order to evaluate the maturation of an individual that can be delayed or advanced due to various factors that may include hormonal disorders, chronic illnesses etc [8].

Regarding the agreement of the radiologists on the assessment of the growth plates' stage of development, substantial agreement was achieved, which is a satisfactory result as there is a lack of guidelines for BAA using MRI in the research. Also, the individuals employed in the assessment of the MRI images were specialized pediatric radiologists with experience in BAA.

From a methodological point of view, the present study employed a nested cross-validation approach which aims to avoid reporting overly optimistic results that could be derived from a 'lucky' test set.

4.2 Comparison with prior work

Most of the studies in the area of BAA that employ ML algorithms aim at building automatic approaches for estimating BA and evaluating on BA given by radiologists [20]. The biggest initiative for proposing automated approaches in this direction was the Radiological Society of North America (RSNA) 2018 Bone Age Challenge [46]. This challenge provided a database of circa 12,000 radiographs, of subjects from 0 to 19 years, labeled with the BA given by radiologists, following the GP method. While the first places achieved MAEs of 4.26, 4.35 and 4.38 months, no comparisons were done to the actual CA of the subjects [46].

For studies which employ BA concepts in order to predict the CA of subjects there are Dallora et al. [47] and Stern et al. [48]. Both employ MRI as the medical imaging of choice and most importantly, they are not based on traditional BAA to make their predictions of CA. They employ the deep learning technology, which is able to learn the important features in the images and then perform the task of regression or classification [49]. The reasoning behind using deep learning to interpret images and learn features is that it is difficult for humans to translate image features into descriptive means, and it is easy to lose information on the process. On the other hand, this problem has a reduced risk of occuring with algorithms able to analyse images pixel by pixel [50].

Dallora et al. [47] used knee MRI images and achieved a MAE of 0.793 years for male subjects in the range of 14-20 years, and 0.988 years for female subjects in the range of 14 to 19 years. Stern et al. [48] used MRI images of the hand and achieved a MAE of 0.82 years for male subjects in the range of 13 to 19 years.

4.3 Limitations

Regarding limitations of the study, it could be argued that due to the high number of classes in the multiclass classification, the sample size in each class would not be big enough to build a generalizable model. However, to address this issue we employed methods to ensure that the model would not overfit and also for not choosing the most overly optimistic choice given by the nested cross-validation. Also, during data collection we ensured a uniform number of subjects in each class to guarantee a balanced dataset.

The selected ROI for this work took into consideration the stress levels for the minors and young adults subjects in regards to the MRI examination. Hence, the clavicle and arm were not considered since it would require the subjects to go head in the MRI machine, which could cause discomfort and stress to the young subjects due to loud noises and small enclosed spaces. Also, the clavicle has a high risk of producing moving artifacts due to the breathing movements. On the practical side, the examination time was on average 15 minutes and the inclusion of these two regions would take approximately double the time.

4.4 Conclusions

This paper presented models for CA estimation and minors versus adults classification (on a threshold of 18 years), using ML algorithms. The models were trained with radiologists assessment of the Calcaneus, Distal Tibia, Proximal Tibia, Distal Femur and Radius; and the additional information regarding physical activity level, parents' origin, type of residence during upbringing, and a self-assessed Tanner Scale of pubertal growth. The models proposed for the classification of minor versus adults produced accuracies of 90% and 84%, for male and female subjects, respectively, with very high recalls for the classification of minors. However, for the chronological age estimation for the eight age groups, ranging from 14-21, the variables in the model did not turn out to be precise enough for estimating the exact CA, only showing acceptable values of MAE for the ages of 14 and 15 years.

Future research should be directed into applying the deep learning technology for the estimation of CA, using multiple ROI.

5. Acknowledgements

We would like to express our greatest appreciation to the participants and staff who took part in our study. This work was supported by the National Board of Health and Welfare of Sweden (Socialstyrelsen). The funding source had no involvement

regarding study design, data collection, analysis, interpretation, or reporting of this work.

6. Conflicts of Interest

None declared.

References

1. Gilsanz V, Ratib O. Hand Bone Age: A Digital Atlas Of Skeletal Maturity. Springer Science & Business Media; 2005.

2. Mansourvar M, Ismail MA, Herawan T, Raj RG, Kareem SA, Nasaruddin FH. Automated bone age assessment: motivation, taxonomies, and challenges. Comput Math Methods Med. 2013;2013: 391626.

3. Satoh M. Bone age: assessment methods and clinical applications. Clin Pediatr Endocrinol. 2015;24: 143–152.

4. Dvorak J, George J, Junge A, Hodler J. Application of MRI of the wrist for age determination in international U-17 soccer competitions. Br J Sports Med. 2007;41: 497–500.

5. Dvorak J, George J, Junge A, Hodler J. Age determination by magnetic resonance imaging of the wrist in adolescent male football players. Br J Sports Med. 2007;41: 45–52.

6. Schmidt S, Vieth V, Timme M, Dvorak J, Schmeling A. Examination of ossification of the distal radial epiphysis using magnetic resonance imaging. New insights for age estimation in young footballers in FIFA tournaments. Sci Justice. 2015;55: 139–144.

7. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, et al. The problem of aging human remains and living individuals: a review. Forensic Sci Int. 2009;193: 1–13.

8. Hjern A, Brendler-Lindqvist M, Norredam M. Age assessment of young asylum seekers. Acta Paediatr. 2012;101: 4–7.

9. Schmeling A, Schulz R, Reisinger W, Mühler M, Wernecke K-D, Geserick G. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med. 2004;118: 5–8.

10. Greulich W, Pyle SL. Radiographic atlas of skeletal development of the hand

and wrist. Am J Med Sci. 1959;238.

11. Tanner JM, Whitehouse RH, Cameron N, Marshall WA, Healy MJR, Goldstein H, et al. Assessment of skeletal maturity and prediction of adult height (TW2 method). Academic Press London; 1975.

12. Nahhas RW, Sherwood RJ, Chumlea WC, Duren DL. An update of the statistical methods underlying the FELS method of skeletal maturity assessment. Ann Hum Biol. 2013;40: 505–514.

13. Kreitner KF, Schweden FJ, Riepert T, Nafe B, Thelen M. Bone age determination based on the study of the medial extremity of the clavicle. Eur Radiol. 1998;8: 1116–1122.

14. Kellinghaus M, Schulz R, Vieth V, Schmidt S, Pfeiffer H, Schmeling A. Enhanced possibilities to make statements on the ossification status of the medial clavicular epiphysis using an amplified staging scheme in evaluating thin-slice CT scans. Int J Legal Med. 2010;124: 321–325.

15. O'Connor JE, Bogue C, Spence LD, Last J. A method to establish the relationship between chronological age and stage of union from radiographic assessment of epiphyseal fusion at the knee: an Irish population study. J Anat. 2008;212: 198–209.

16. Dedouit F, Auriol J, Rousseau H, Rougé D, Crubézy E, Telmon N. Age assessment by magnetic resonance imaging of the knee: a preliminary study. Forensic Sci Int. 2012;217: 232.e1–7.

17. Krämer JA, Schmidt S, Jürgens K-U, Lentschig M, Schmeling A, Vieth V. Forensic age estimation in living individuals using 3.0 T MRI of the distal femur. Int J Legal Med. 2014;128: 509–514.

18. Ekizoglu O, Hocaoglu E, Can IO, Inci E, Aksoy S, Bilgili MG. Magnetic resonance imaging of distal tibia and calcaneus for forensic age estimation in living individuals. Int J Legal Med. 2015;129: 825–831.

19. Ekizoglu O, Inci E, Ors S, Kacmaz IE, Basa CD, Can IO, et al. Applicability of T1-weighted MRI in the assessment of forensic age based on the epiphyseal closure of the humeral head. Int J Legal Med. 2019;133: 241–248.

20. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PLoS One. 2019;14: e0220242.

21. Crema MD, Roemer FW, Marra MD, Burstein D, Gold GE, Eckstein F, et al. Articular cartilage in the knee: current MR imaging techniques and applications in clinical practice and research. Radiographics. 2011;31: 37–61.

22. Gründer W. MRI assessment of cartilage ultrastructure. NMR Biomed. 2006;19: 855–876.

23. De Simone M, Farello G, Palumbo M, Gentile T, Ciuffreda M, Olioso P, et al. Growth charts, growth velocity and bone development in childhood obesity. Int J Obes Relat Metab Disord. 1995;19: 851–857.

24. Cutler GB Jr. The role of estrogen in bone growth and maturation during childhood and adolescence. J Steroid Biochem Mol Biol. 1997;61: 141–144.

25. Mirtz TA, Chandler JP, Eyers CM. The effects of physical activity on the epiphyseal growth plates: a review of the literature on normal physiology and clinical implications. J Clin Med Res. 2011;3: 1–7.

26. Ontell FK, Ivanovic M, Ablin DS, Barlow TW. Bone age in children of diverse ethnicity. AJR Am J Roentgenol. 1996;167: 1395–1398.

27. Karlberg J. Secular trends in pubertal development. Horm Res. 2002;57 Suppl 2: 19–30.

28. Tanner JM, Whitehouse RH, Cameron N. Assessment of skeletal maturity and prediction of adult height (TW3 Method). 3rd edn, 2001. WB Saunders: London;

29. Mitchell TM. Machine learning and data mining. Commun ACM. 1999;42. Available: http://www.ri.cmu.edu/pub_files/pub1/mitchell_tom_1999_1/mitchell_tom_1999_1.pdf

30. Thodberg HH, Kreiborg S, Juul A, Pedersen KD. The BoneXpert method for automated determination of skeletal maturity. IEEE Trans Med Imaging. 2009;28: 52–66.

31. Marshall WA, Tanner JM. Variations in pattern of pubertal changes in girls. Arch Dis Child. 1969;44: 291–303.

32. Marshall WA, Tanner JM. Variations in the pattern of pubertal changes in boys. Arch Dis Child. 1970;45: 13–23.

33. Keys A, Fidanza F, Karvonen MJ, Kimura N, Taylor HL. Indices of relative weight and obesity. International Journal of Epidemiology. 2014. pp. 655–665. doi:10.1093/ije/dyu058

34. Cohen J. A Coefficient of Agreement for Nominal Scales. Educ Psychol Meas. 1960;20: 37–46.

35. Zhang S. Nearest neighbor selection for iteratively kNN imputation. J Syst Softw. 2012;85: 2541–2552.

36. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak. 2016;16 Suppl 3: 74.

37. Kuhn M, Johnson K. Applied Predictive Modeling. Springer, New York, NY; 2013.

38. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6: 10.

39. Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1809.09446

40. Gaudette L, Japkowicz N. Evaluation Methods for Ordinal Classification. Advances in Artificial Intelligence. Springer Berlin Heidelberg; 2009. pp. 207–210.

41. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. Inf Process Manag. 2009;45: 427–437.

42. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33: 159–174.

43. Margalit A, Cottrill E, Nhan D, Yu L, Tang X, Fritz J, et al. The Spatial Order of Physeal Maturation in the Normal Human Knee Using Magnetic Resonance Imaging. Journal of Pediatric Orthopaedics. 2019;39: e318–e322.

44. O'Connor JE, Coyle J, Bogue C, Spence LD, Last J. Age prediction formulae from radiographic assessment of skeletal maturation at the knee in an Irish population. Forensic Sci Int. 2014;234: 188.e1–8.

45. Cox LA. The biology of bone maturation and ageing. Acta Paediatr Suppl. 1997;423: 107–108.

46. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology. 2019;290: 498–503.

47. Dallora AL, Berglund JS, Brogren M, Kvist O, Ruiz SD, Dübbel A, et al. Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach. JMIR Medical Informatics. 2019. p. e16291. doi:10.2196/16291

48. Štern D, Payer C, Urschler M. Automated age estimation from MRI volumes of the hand. Medical Image Analysis. 2019. p. 101538. doi:10.1016/j.media.2019.101538

49. Ker J, Wang L, Rao J, Lim T. Deep Learning Applications in Medical Image Analysis. IEEE Access. 2018;6: 9375–9389.

50. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep Learning for Health Informatics. IEEE J Biomed Health Inform. 2017;21: 4–21.

Abbreviations

AUC: Area Under the Curve

BA: Bone Age

BAA: Bone Age Assessment

CA: Chronological Age

GP: Greulich-Pyle

KNN: K-nearest Neighbors

MAE: Mean Absolute Error

MLP: Multi-Layer Perceptron

MRI: Magnetic Resonance Imaging

RMSE: Root Mean Squared Error

RSNA: Radiological Society of North America

SLR: Systematic Literature Review

SVM: Support Vector Machines

TW: Tanner-Whitehouse

Supplementary Table 1. results from the assessment of MRI images of male subjects

		Calcaneus assessment results - male subjects								
		Age								
		14	15	16	17	18	19	20	21	Total
CALCANEUS	Stage3	1	1	0	0	0	0	0	0	2
	Stage4a	23	8	2	0	0	0	0	0	33
	Stage4b	7	6	2	0	0	0	0	0	15
	Stage4c	16	11	10	1	0	1	0	0	39
	Stage5	11	30	46	57	53	57	53	59	366
Total		58	56	60	58	53	58	53	59	455
		Distal Tibia assessment results - male subjects								
		Age								
		14	15	16	17	18	19	20	21	Total
DISTAL TIBIA	Stage4a	20	10	2	0	0	0	0	0	32
	Stage4b	18	12	4	0	0	0	0	0	34
	Stage4c	17	23	19	6	1	2	2	0	70
	Stage5	3	11	35	52	52	56	51	59	319
Total		58	56	60	58	53	58	53	59	455
		Proximal Tibia assessment results - male subjects								
		Age								Total
		14	15	16	17	18	19	20	21	
PROXIMAL TIBIA	Stage4a	25	16	6	0	0	0	0	0	47
	Stage4b	24	16	7	2	0	0	0	0	49
	Stage4c	9	23	35	23	5	2	0	0	97
	Stage5	0	1	12	33	48	56	53	59	262
Total		58	56	60	58	53	58	53	59	455
		Distal Femur assessment results - male subjects								
		Age								
		14	15	16	17	18	19	20	21	Total
DISTAL FEMUR	Stage4a	23	21	11	0	0	0	0	0	55
	Stage4b	25	23	15	7	2	0	0	0	72
	Stage4c	10	12	24	25	6	1	0	0	78
	Stage5	0	0	10	26	45	57	53	59	250
Total		58	56	60	58	53	58	53	59	455
		Radius assessment results - male subjects								
		Age								Total
		14	15	16	17	18	19	20	21	
RADIUS	Stage3	11	3	1	0	0	0	0	0	15
	Stage4a	38	33	22	11	0	1	0	0	105
	Stage4b	7	15	12	12	4	0	0	0	50
	Stage4c	2	4	20	20	17	5	1	0	69
	Stage5	0	1	5	15	32	52	52	59	216
Total		58	56	60	58	53	58	53	59	455

Supplementary Table 2. results from the assessment of MRI images of female subjects

Calcaneus assessment results - female subjects										
		Age								Total
		14	15	16	17	18	19	20	21	
CALCANEUS	Stage4a	1	0	0	0	0	0	0	0	1
	Stage4b	3	0	0	0	0	0	0	0	3
	Stage4c	12	6	2	1	0	0	0	0	21
	Stage5	43	52	55	59	59	57	57	60	442
Total		59	58	57	60	59	57	57	60	467
Distal Tibia assessment results - female subjects										
		Age								Total
		14	15	16	17	18	19	20	21	
DISTAL TIBIA	Stage4a	2	0	0	0	0	0	0	0	2
	Stage4b	6	0	0	0	0	0	0	0	6
	Stage4c	21	9	4	1	0	0	0	0	35
	Stage5	30	49	53	59	59	57	57	60	424
Total		59	58	57	60	59	57	57	60	467
Proximal Tibia assessment results - female subjects										
		Age								Total
		14	15	16	17	18	19	20	21	
PROXIMAL TIBIA	Stage4a	10	3	0	0	0	0	0	0	13
	Stage4b	10	4	0	0	0	0	0	0	14
	Stage4c	28	22	7	2	0	0	0	0	59
	Stage5	11	29	50	58	59	57	57	60	381
Total		59	58	57	60	59	57	57	60	467
Distal Femur assessment results - female subjects										
		Age								Total
		14	15	16	17	18	19	20	21	
DISTALA FEMUR	Stage4a	19	7	0	0	0	0	0	0	26
	Stage4b	19	11	2	1	0	0	0	0	33
	Stage4c	17	18	16	8	1	0	0	0	60
	Stage5	4	22	39	51	58	57	57	60	348
Total		59	58	57	60	59	57	57	60	467
Radius assessment results - female subjects										
		Age								Total
		14	15	16	17	18	19	20	21	
RADIUS	Stage3	2	0	0	0	0	0	0	0	2
	Stage4a	26	13	3	1	0	0	0	0	43
	Stage4b	18	11	8	1	0	0	0	0	38
	Stage4c	11	26	22	13	6	0	0	0	78
	Stage5	2	8	24	45	53	57	57	60	306
Total		59	58	57	60	59	57	57	60	467

Study III

15 Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach

Original Paper

Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach

Ana Luiza Dallora[1], MSc; Johan Sanmartin Berglund[1], PhD; Martin Brogren[2], MSc; Ola Kvist[3], MD; Sandra Diaz Ruiz[3], MD, PhD; André Dübbel[2], MSc; Peter Anderberg[1], PhD

[1]Department of Health, Blekinge Institute of Technology, Karlskrona, Sweden

[2]Optriva AB, Stockholm, Sweden

[3]Department of Pediatric Radiology, Karolinska University Hospital, Stockholm, Sweden

Corresponding Author:
Peter Anderberg, PhD
Department of Health
Blekinge Institute of Technology
Valhallavägen 1
Karlskrona, 37141
Sweden
Phone: 46 0734223736
Email: pan@bth.se

Abstract

Background: Bone age assessment (BAA) is an important tool for diagnosis and in determining the time of treatment in a number of pediatric clinical scenarios, as well as in legal settings where it is used to estimate the chronological age of an individual where valid documents are lacking. Traditional methods for BAA suffer from drawbacks, such as exposing juveniles to radiation, intra- and interrater variability, and the time spent on the assessment. The employment of automated methods such as deep learning and the use of magnetic resonance imaging (MRI) can address these drawbacks and improve the assessment of age.

Objective: The aim of this paper is to propose an automated approach for age assessment of youth and young adults in the age range when the length growth ceases and growth zones are closed (14-21 years of age) by employing deep learning using MRI of the knee.

Methods: This study carried out MRI examinations of the knee of 402 volunteer subjects—221 males (55.0%) and 181 (45.0%) females—aged 14-21 years. The method comprised two convolutional neural network (CNN) models: the first one selected the most informative images of an MRI sequence, concerning age-assessment purposes; these were then used in the second module, which was responsible for the age estimation. Different CNN architectures were tested, both training from scratch and employing transfer learning.

Results: The CNN architecture that provided the best results was GoogLeNet pretrained on the ImageNet database. The proposed method was able to assess the age of male subjects in the range of 14-20.5 years, with a mean absolute error (MAE) of 0.793 years, and of female subjects in the range of 14-19.5 years, with an MAE of 0.988 years. Regarding the classification of minors—with the threshold of 18 years of age—an accuracy of 98.1% for male subjects and 95.0% for female subjects was achieved.

Conclusions: The proposed method was able to assess the age of youth and young adults from 14 to 20.5 years of age for male subjects and 14 to 19.5 years of age for female subjects in a fully automated manner, without the use of ionizing radiation, addressing the drawbacks of traditional methods.

(JMIR Med Inform 2019;7(4):e16291) doi: 10.2196/16291

KEYWORDS

age assessment; bone age; skeletal maturity; deep learning; convolutional neural networks; transfer learning; machine learning; magnetic resonance imaging; medical imaging; knee

Introduction

Background

Bone age and skeletal maturity are closely related concepts that measure the stage of bone development of an individual [1,2]. When compared to the chronological age, they aid in the diagnosis and in determining the time of treatment of many pediatric disorders related to orthodontics, orthopedics, and endocrinology. Further, they are also used in estimations about the final height of an individual [3].

From a legal standpoint, bone age assessment (BAA) also plays an important role in the estimation of chronological age. In this sense, the estimation of the bone age is employed when determining if an individual is a minor in the absence of valid documents, which is the case for numerous unaccompanied minors seeking asylum [2], as well as in adoption, imputability, and pedopornography judicial and civil issues [4]. The estimation of chronological age is also used in age-related sports competitions to guarantee fair play [5,6]. In all of these cases, BAA is an important tool that is used to make important legal decisions that can enormously affect an individual's life.

The traditional methods for performing BAA are the Greulich-Pyle (GP) atlas and the Tanner-Whitehouse (TW) scoring system. The GP atlas [7] comprises hand and wrist radiograph reference images of subjects from 0 to 19 years of age for males and 0 to 18 years of age for females. The process for determining bone age is done by comparing the nearest matching reference image in the atlas to the image of the individual being assessed [3]. The TW scoring system [8] first analyzes the hand and wrist radiograph of a subject and categorizes the skeletal maturity scores of the ossification centers of the radius, ulna, and 13 short bones of the hand and carpals into stages ranging from A to I. Then, all of the stages are aggregated into a numerical score that is converted to the bone age [2].

Drawbacks of the Traditional Age-Assessment Methods

The drawbacks of the GP and TW methods derive from the fact that they are done manually by radiologists; thus, they can be prone to inter- and intrarater variability, in addition to being time-consuming tasks [9,10].

Also, there is an important ethical issue related to submitting healthy subjects to ionizing radiation without therapeutic purposes, which is especially important in the case of assessing if an individual is a minor for legal purposes [10]. This scenario suggests that new approaches for the assessment of age should be explored by research in order to address these drawbacks.

The use of radiation-free medical imaging can be achieved by the employment of magnetic resonance imaging (MRI). An additional advantage of MRI technology is that it supports the manipulation of the image's contrast, granting the possibility of highlighting different tissue types and allowing better visualization of ossification centers [11,12]. Additionally, since MRI images are volumetric, more information can be extracted and analyzed when compared to 2D radiographs [13].

The issues related to rater variability and time spent in the assessment are big motivators for the use of more automated techniques like deep learning. Deep learning is a type of machine learning technique, which refers to algorithms that are able to learn a task from a set of training examples; in view of a new set of data, this task can be reproduced with an acceptable performance [14]. The use of machine learning for health applications is not new and is broadly employed for disease prediction and prognosis [14,15], genomics, proteomics, and microarrays [16]; it has also been used to predict health care utilization through Web search logs [17]. Contrary to many machine learning techniques, deep learning methods perform feature engineering: instead of having a domain expert specify important data characteristics, it learns the informative representations in the data and performs a task of classification or regression [18,19]. When working with medical images, this is especially advantageous since image features are difficult to translate into descriptive means [20]. That is the reason why the first applications of deep learning with health data were aimed at analyzing medical images, specifically MRI images of the brain for the prediction of Alzheimer disease and MRI images of the knee to estimate the risk of osteoarthritis [21]. In the specific area of BAA, most computerized approaches extract features following established procedures (eg, TW or GP), which can be limiting in terms of the information available in the image [22]. When using deep learning, the algorithm finds the important representations in the images without any constraint, which could allow more features in the image to be considered in the classification or regression task not previously known by the current methods [22].

Goal of This Study

Taking into account the numerous settings in which the estimation of chronological age is employed and their importance and potential effect on individuals' lives, it is important to address the drawbacks in the methods currently in use. Thus, this paper proposes an automated approach for age assessment of youth and young adults (14-21 years of age) employing deep learning methods with MRI images of the knee.

The knee region aggregates four ossification centers—femur, tibia, fibula, and patella—but it has not been explored very much by the research in BAA, which is mostly focused on the hand and wrist regions; this research makes use of radiograph images, due to the impact the GP method, which is still considered by many to be the gold standard for BAA [23]. The choice of the knee region in this study was motivated by findings in the research with MRI images that reported the presence of cartilage signal intensity at the knee ossification centers in male individuals from 17.8 to 30.0 years of age and female individuals from 16.6 to 29.6 years of age, which could imply later fusion of maturation centers [24]. Additionally, recent findings in the research of BAA with MRI images of the knee also reported a uniform spatial pattern of maturation of ossification centers in the knee in both male and female individuals [12].

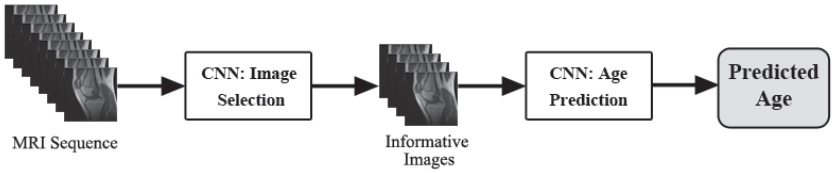
Methods

Overview

The fully automated age-assessment method proposed in this paper uses MRI images of the knee and the subjects' chronological ages to train deep learning models for continuous age estimation with convolutional neural networks (CNNs).

An overview of the method is shown in Figure 1. It comprises two CNN models: the first one is responsible for selecting the most informative images of an MRI sequence for age-assessment purposes; these are then fed to the age-prediction CNN, which outputs an estimated age. The remainder of this section further details the process of training, deploying, and evaluating the CNN models of the proposed method as well as the materials used in the experiments.

Figure 1. Overview of the proposed automated age-assessment method. CNN: convolutional neural network; MRI: magnetic resonance imaging.



Recruitment

This study prospectively acquired MRI images of the knee region of 402 volunteer subjects—221 males (55.0%) and 181 (45.0%) females—aged 14.0-21.5 years (see Table 1) between 2017 and 2018. It is important to note that throughout the text of this paper, the mention of an age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years). The criteria used for subject recruitment in the study were as follows:

1. Inclusion criteria: subjects (1) were born in Sweden and (2) have a birth certificate verified by national authorities.
2. Exclusion criteria: subjects (1) have a history of bilateral fractures or trauma near the growth plate, (2) have a history of chronic disease or long-term medication, (3) exhibit noncompliance during MRI examinations, (4) have resided outside Sweden for more than 6 consecutive months, and (5) experienced a past pregnancy or were pregnant at the time of recruitment: all female volunteer subjects were tested.

Table 1. Age distribution of the volunteer subjects[a] (N=402).

Gender	Subject age group[b], years, n (%)								Total, n (%)
	14	15	16	17	18	19	20	21	
Male (N=221)	22 (10.0)	26 (11.8)	31 (14.0)	25 (11.3)	24 (10.9)	25 (11.0)	35 (15.8)	33 (14.9)	221 (100)
Female (N=181)	22 (12.2)	21 (11.6)	30 (16.6)	27 (14.9)	20 (11.0)	12 (6.6)	25 (13.8)	24 (13.3)	181 (100)
Total (N=402)	44 (10.9)	47 (11.7)	61 (15.2)	52 (12.9)	44 (10.9)	37 (9.2)	60 (14.9)	57 (14.1)	402 (100)

[a]All data were acquired within a maximum of 6 months after the subjects' birth dates.

[b]Age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years).

Magnetic Resonance Imaging Examinations

The MRI examinations were performed on 1.5 Tesla whole-body MRI scanners with dedicated knee coils. The images were taken from the nondominant side of the knee; however, in the case of previous fracture or trauma near these regions, the dominant side was imaged.

The examinations were performed in two sites, with the same protocol, 256 x 256-pixel resolution, and 160 x 160 mm field of view. The following machinery was used:

1. Site 1: MAGNETOM Avanto Fit (Siemens Healthcare Gmbh) and Achieva (Philips Healthcare) whole-body scanners.
2. Site 2: SIGNA (GE Healthcare) whole-body scanner.

Data Privacy and Study Ethics

All acquired data were anonymized and stratified by age and gender. The study was approved by the local ethics committee and was conducted in accordance with the Declaration of Helsinki. Written informed consent was acquired from all subjects and legal guardians, in the case of minors.

Image Selection

Each MRI examination produced 17-35 images per subject, however, not all of them were equally informative in regard to the assessment of the age of an individual. To simplify the age estimation learning task, only the best images were considered for the *CNN: Age Prediction* model. To make the method fully automated without any need for human input, a CNN classifier was trained to be able to select the most informative images in an MRI sequence. An *informative* image in the context of the proposed method corresponds to the part of the bone that

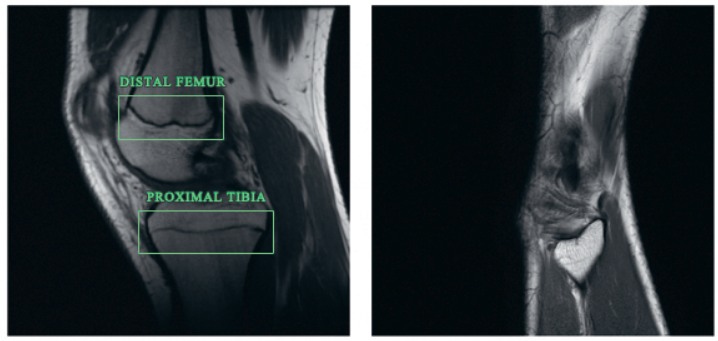
contains anatomical structures of interest, which include the growth plate, epiphysis, and metaphysis. This classifier corresponds to the *CNN: Image Selection* block in Figure 1.

The CNN architecture used was GoogLeNet [25], a model that has been shown to generalize well to a wide variety of image classification tasks, medical and otherwise [26].

To be able to train this classifier, one image from each MRI sequence that had growth zones clearly visible was annotated as *informative*. Also, one image from each MRI sequence in which the growth zones were occluded by other tissue types was selected and labelled as *noninformative*. Examples of informative and noninformative images are shown in Figure 2.

The output of the CNN model is the confidence levels of the two classes—informative and noninformative—for the given MRI image. The confidence level is a continuous value between 0 and 1, where 1 is the highest confidence level and the confidence levels of the two classes sum up to 1. In later steps, only images with a confidence level for the informative class above a threshold C on the test set were used.

Figure 2. Examples of informative and noninformative images from the same subject.



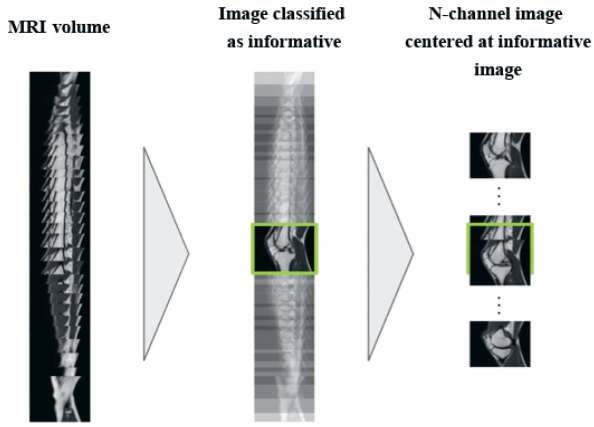
Age Prediction

For predicting the age of an individual from the MRI images, another CNN model was built. This model corresponds to the *CNN: Age Prediction* block in Figure 1. Seven different CNN architectures were considered; these were as follows: GoogLeNet [25], ResNet-50 [27], Inception-v3 [28], Visual Geometry Group (VGG) [29], AlexNet [30], DenseNet [31], and U-Net [32].

The final classification layer of these networks was replaced with a linear scalar output providing the age estimation. The only exception from this was U-Net, which is a fully connected model without classification layers in the end. Here, the linear scalar output was added after the last convolutional layer instead.

The age-prediction model takes an MRI image with N channels as input, then outputs the estimated chronological age of the subject. To create an image with N channels, a subset of the MRI volume, centered on an image classified as informative, is extracted (see Figure 3).

Figure 3. Example of how an N-channel image is created from one of the images in the magnetic resonance imaging (MRI) volume classified as informative.



Input images of 1-9 channels were tested. The idea was that the model might be able to use information from neighboring images to improve results and make the model more robust to mistakes in the image-selection process.

Training the Models

Training and Evaluation

The Convolutional Architecture for Fast Feature Embedding (Caffe) deep learning framework [33] was used to train the models. Training and evaluation were done on Amazon Web Services on an Elastic Compute Cloud (EC2) P3.2xlarge with a Tesla V100 Nvidia graphics processing unit.

Optimization

The Adam optimizer [34] was used to minimize the cross-entropy loss when training the classifier and the Euclidean loss when training the regressor. Cross-entropy loss for binary classification is calculated as follows:

$$-1/N \Sigma^{N}_{i=1} y_i \times log(p(y_i)) + (1-y_i) \times log(1-p(y_i)) \quad (1)$$

with N being the number of training samples per batch, y being a binary indicator (0 or 1) of the correctness of classification for an observation o being of class c, and p being the predicted probability of an observation o being of class c. Euclidean loss is calculated as follows:

$$1/2N \Sigma^{N}_{i=1} \ |x_i^1 - x_i^2|^{\ 2}_{\ 2} \quad (2)$$

with N being the number of training samples per batch, x^1 the estimated age, and x^2 the verified chronological age.

Cross-Validation

All experiments were performed using six-fold cross-validation, including the test set. The dataset was split into six equal-sized parts, with data stratified for age and gender. This data partition followed the procedure that all of the images from a subject were assigned to a single fold. Four parts were used for training, one part was used for validation during training, and one part was used to finally evaluate and measure the model's performance. This was done to be able to evaluate the models on the full dataset.

Before performing a full cross-validation, a sparse grid search was performed for each model to find good hyperparameters. This was done using the validation set of the first cross-validation split only. The hyperparameters tuned during the grid search were as follows: learning rate, weight decay, momentum, dropout ratio, and batch size.

Transfer Learning

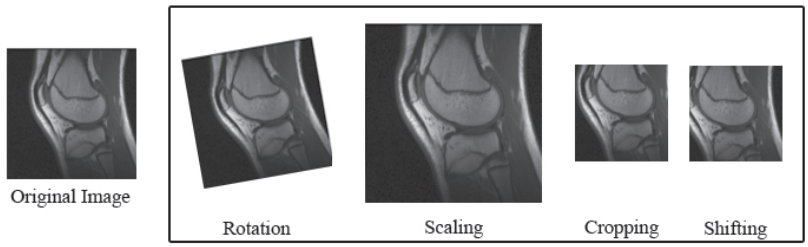
Both training from scratch and transfer learning were tested. Transfer learning is a technique that, instead of using randomly initialized weights, takes the weights from a CNN that has already been trained to perform well on a generic task as a starting point. The model is then adapted by carefully updating the weights using the task-specific training data. This makes it possible to leverage larger datasets to avoid overfitting when the task-specific dataset is small [35,36]. All pretrained models used in this paper were trained on ImageNet [37]. During the task-specific training, the weights of all layers were updated.

Data Augmentation

Data augmentation is a technique that aims to synthetically increase the size of the training set from existing data without additional labelling work, using geometric or photometric transformations, noise injections, and color jittering operations. It is used to prevent overfitting when training CNNs on small datasets [38,39].

In the proposed method, data augmentation was performed on all training samples to increase the dataset. The images were randomly cropped, shifted, rotated at a maximum of five degrees, and scaled up to 20%. Figure 4 shows examples of the applied data augmentation operations.

Figure 4. Examples of data augmentation operations applied in the proposed method.



Estimation

When estimating the age on the test set for each subject, all images with a confidence higher than threshold C of 0.95 for the informative class were used. Each of these test images were used to create a number of copies with different augmentations applied to each copy. All augmented test images were fed through the network to produce one result each. Finally, the results from the augmented versions of the images were used to estimate a final result. This technique has been shown to improve the performance of the predictions and is widely used within deep learning [25].

In this method, each image was augmented 15 times, using the same augmentations as during training, generating 15 new images. If none of the images for a subject had a confidence higher than the threshold, the image with the highest confidence was used instead. This was the case for two subjects only. The highest confidence value for these subjects were 0.91 and 0.81. If more than 10 images had a confidence level higher than the

threshold, only the 10 images with the highest confidence were used in order to set a maximum limit on the processing time.

Age was estimated for all augmented images and, finally, the median of all estimated ages for each subject was computed to get the final prediction. For example, if a subject had eight images with high-enough confidence, 120 augmented images were created and 120 ages were estimated, of which the median was used as the final estimated age.

Results

Overview

Hyperparameters and settings were tuned to optimize the models' performance. This was done through a sparse grid search on the first cross-validation split, as specified previously. The validation set was used for tuning in order to avoid tuning specifically toward the test set and thereby overestimating the models' performance on new data. The final results reported in this section were evaluated on the full dataset from the cross-validation test sets in terms of the mean absolute error (MAE), calculated as follows:

$$MAE = 1/n \sum^{n}_{i=1} |x_i - x| \quad (3)$$

with n being the number of samples, x_i being the estimated age, and x being the verified chronological age.

Conclusions From Experiments

Fine-tuning pretrained models showed significantly better results compared to training the models from scratch. The two architectures that showed best results were GoogLeNet and ResNet-50. Training on men and women subjects separately gave better results for both groups compared to single training using all data.

The best results were achieved using a confidence threshold C of 0.95 in the image selection data preprocessing stage for choosing the most informative MRI images. The results did not change much using different thresholds. MAE differed only by 0.004 years when using thresholds in the range of 0.5-0.99.

Results were very similar when using MRI images with one or three channels, but with more channels than three the performance dropped. This can be due to the increasing number of parameters in the models when using more channels, which might lead to overfitting. Using one channel gave a slightly better result, which is why we used this in our final models.

The hyperparameters that gave the best results were as follows:

1. Learning rate: 1e-4
2. Weight decay: 1e-2
3. Momentum: 0.83
4. Dropout ratio: 0.7 for GoogLeNet and 0.6 for ResNet-50
5. Batch size: 66 for GoogLeNet and 30 for ResNet-50

The best results were achieved when resizing the images to 256×256 pixels for both GoogLeNet and ResNet-50. Both these architectures use cropped images of size 224×224 pixels as input.

Results for the Best Models

The results for the experiments with the best-performing models, GoogLeNet and ResNet-50, in terms of the MAE and SD per age group is shown in Figure 5 and detailed in Table 2 below. The acquisition of the MRI images happened in a window within 6 months from the subjects' birthdays. The best overall results for male subjects were achieved by the GoogLeNet model using knee MRI images. When training the age-prediction model for women, only the architecture performing best on men was considered.

There is a clear trend on all of the experiments among male subjects in which the MAE increases substantially from the age of 21. The same phenomenon occurs for the model among women subjects but from the age of 20. These results lead us to believe that after the ages of 20.5 for men and 19.5 for women, no information regarding older ages can be extracted from the MRI image data, regarding the knee region. This is also supported by Figure 6 and Table 3, which show that the mean estimated age planes out around these ages for the respective genders. The models underestimated the age more and more the older the subjects got after these ages. In conclusion, the presented method is not able to estimate ages above 20.5 for men and above 19.5 for women. Therefore, these ages were removed in the results below, which focus on the applicable age ranges for the models: 14 to 20.5 years for men and 14 to 19.5 years for women.

Figure 5. Comparison of the best-performing models: GoogLeNet and ResNet-50. MAE: mean absolute error.

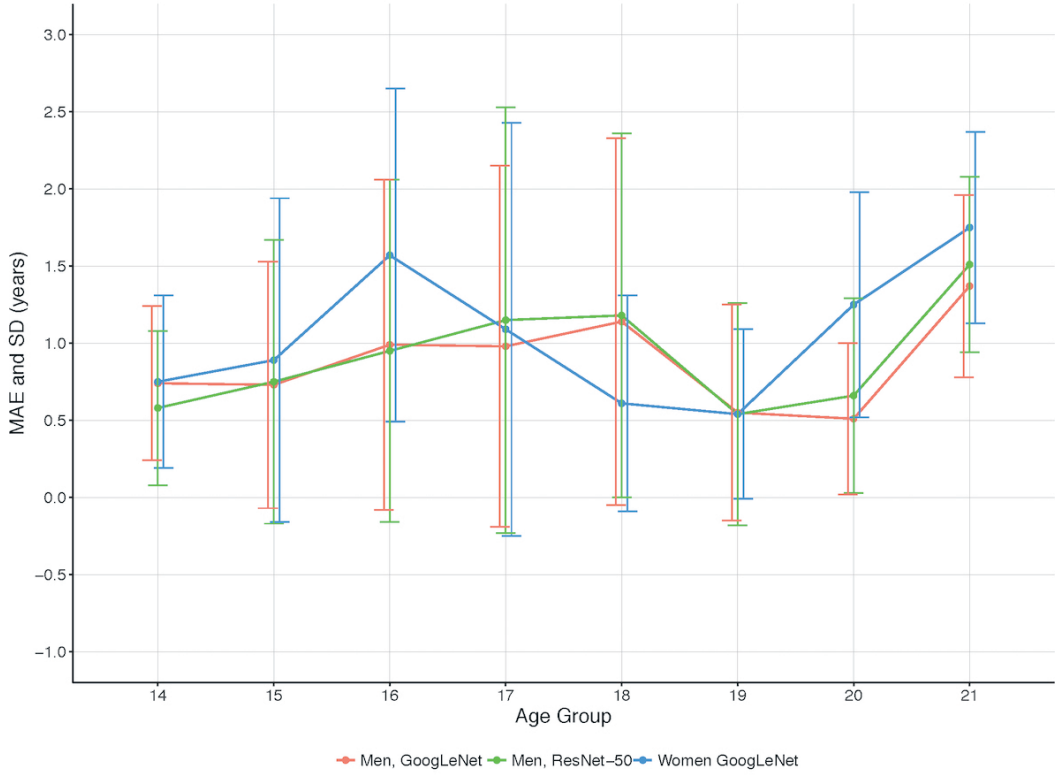


Table 2. Results from the experiments with the best-performing models: GoogLeNet and ResNet-50.

Gender, model	Subject age group[a] in years, MAE[b] (SD)							
	14	15	16	17	18	19	20	21
Men, GoogLeNet	0.74 (0.50)	0.73 (0.80)	0.99 (1.07)	0.98 (1.17)	1.14 (1.19)	0.55 (0.70)	0.51 (0.49)	1.37 (0.59)
Men, ResNet-50	0.58 (0.50)	0.75 (0.92)	0.95 (1.11)	1.15 (1.38)	1.18 (1.18)	0.54 (0.72)	0.66 (0.63)	1.51 (0.57)
Women, GoogLeNet	0.75 (0.56)	0.89 (1.05)	1.57 (1.08)	1.09 (1.34)	0.61 (0.70)	0.54 (0.55)	1.25 (0.73)	1.75 (0.62)

[a]Age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years).

[b]MAE: mean absolute error.

Figure 6. Mean age and mean estimated age per age group with the best-performing model, GoogLeNet, on male and female subjects.

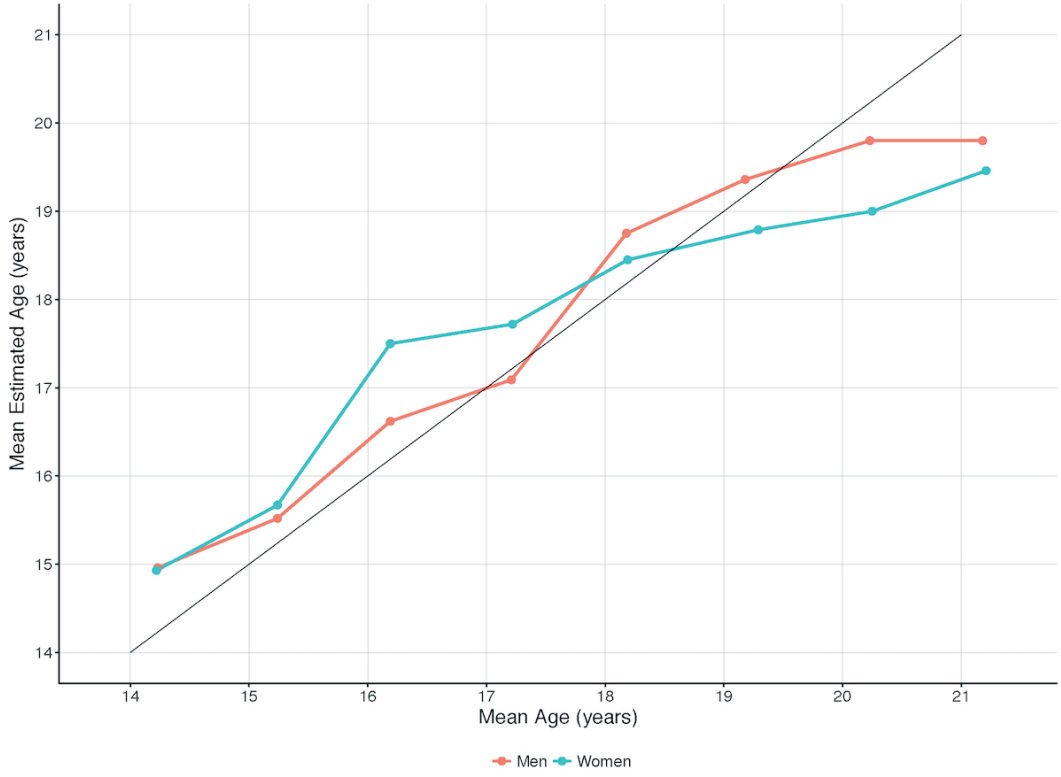


Table 3. Mean age and mean estimated age per age group by the best-performing model, GoogLeNet, on male and female subjects.

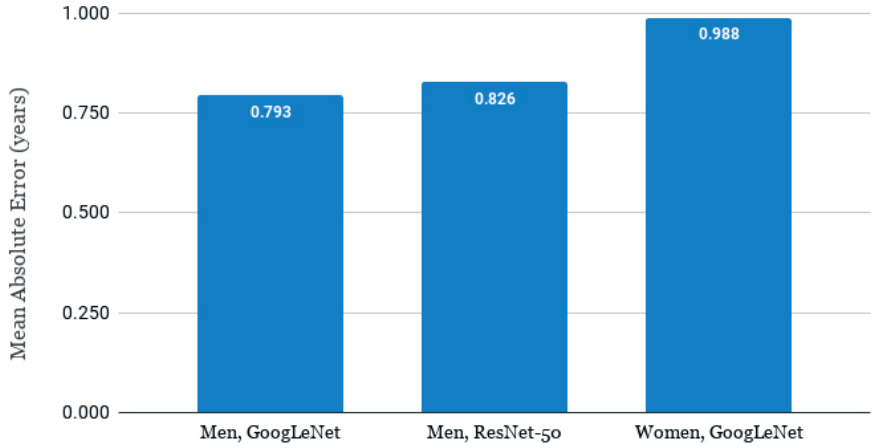
Gender	Subject age group[a], years							
	14	15	16	17	18	19	20	21
Men, mean age	14.23	15.24	16.19	17.21	18.18	19.18	20.23	21.18
Men, mean estimated age	14.96	15.52	16.62	17.09	18.75	19.36	19.80	19.80
Women, mean age	14.22	15.24	16.19	17.22	18.19	19.29	20.25	21.21
Women, mean estimated age	14.93	15.67	17.50	17.72	18.45	18.79	19.00	19.00

[a]Age group X refers to an age span from X to X.5 (eg, the age group 14 refers to an age span of 14 to 14.5 years).

Results for the Best Models in the Applicable Age Ranges

Figure 7 shows the MAE in years for the best models in their applicable ranges: 14-20.5 years for men and 14-19.5 years for women. The best achieved result for the age prediction of youth and young adult individuals in this study corresponds to an MAE of 0.793 years for men and 0.988 years for women, using the GoogleNet architecture.

Figure 7. Mean absolute error (MAE) of the best-performing models in the applicable age ranges.



Results for the GoogLeNet Model in the Applicable Age Ranges for Male and Female Subjects

Figures 8 and 9 show the MAE for the GoogLeNet model applied to male and female subjects, respectively, in the applicable age ranges. It is interesting to notice that the age range with the highest error occurs earlier for females (age group of 16) compared to men (age group of 18). This goes in line with previous knee studies where findings showed that women mature earlier than men [40].

Figure 8. Mean absolute error (MAE) for the GoogLeNet model for male subjects in the applicable age ranges.

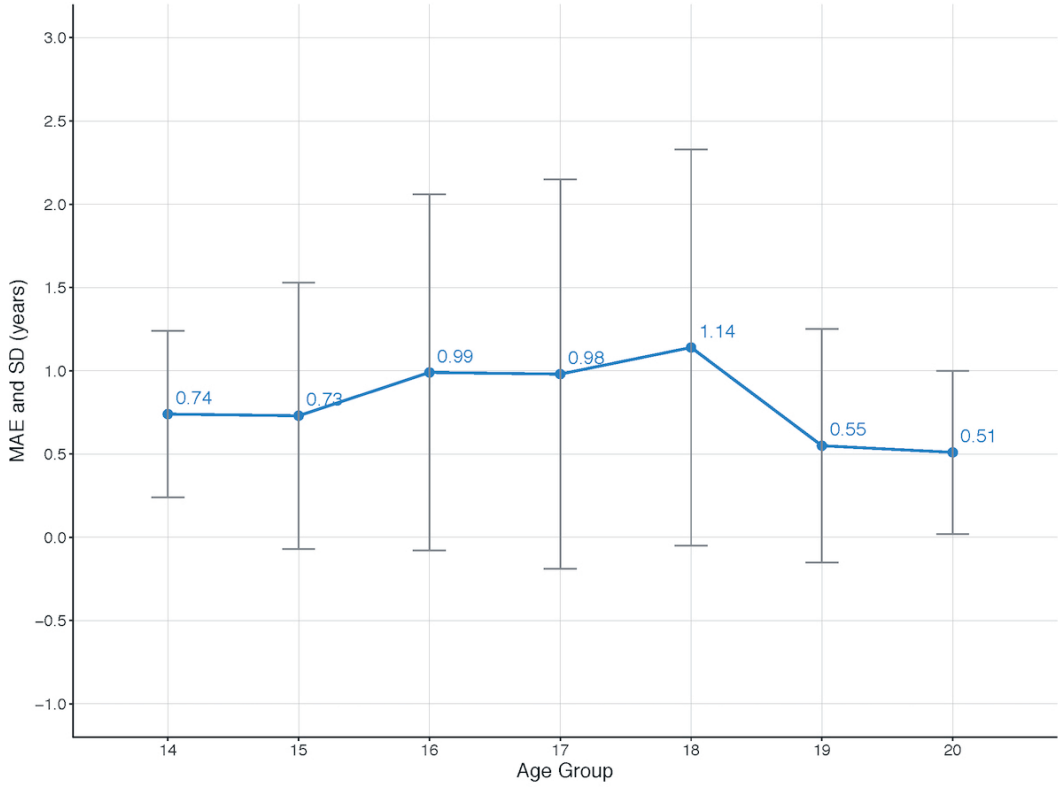
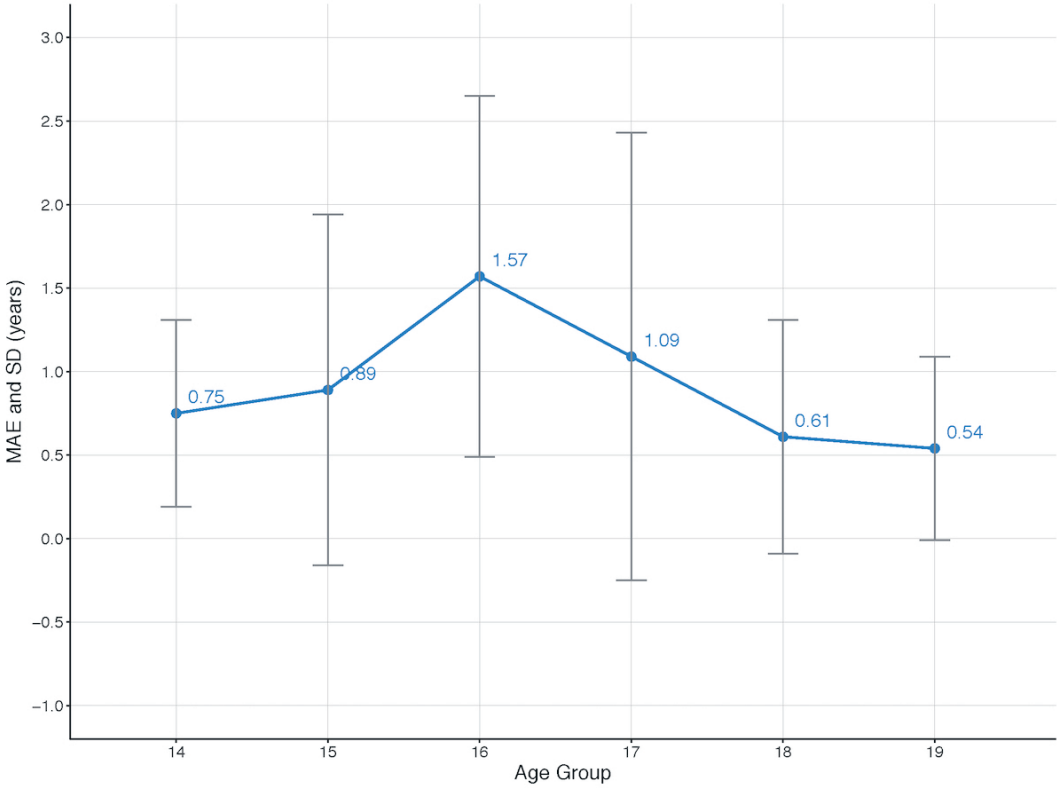


Figure 9. Mean absolute error (MAE) for the GoogLeNet model for female subjects in the applicable age ranges.



Classification Performance of Minors Versus Adults

Experiments were also performed for classification of subjects as being adults or minors, considering the age of 18 years old as the adulthood threshold. This classification is especially important in cases regarding the age assessment of minors from a legal standpoint.

No new training of models was performed. Instead, the classification of adults and minors was performed by applying a threshold to the estimated age from the best-performing models trained in the age-assessment experiments.

Three different strategies for setting the threshold were evaluated:

1. Setting the threshold to increase the accuracy for minors and sacrificing accuracy for adults.
2. Setting the threshold to get as equal accuracy as possible for adults and minors.
3. Using the threshold of 18 years of age without any modification.

The results for male subjects are shown in Figure 10 and Table 4. The same procedures and reasoning were also applied to the women's case and the results are shown in Figure 11 and Table 5.

Figure 10. Accuracies for minor versus adult classification of male subjects, using threshold to increase accuracy for minors.

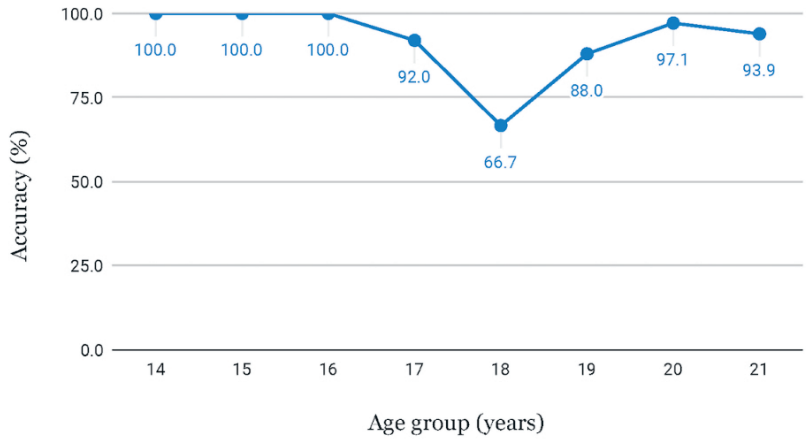


Table 4. Accuracies for minor versus adult classification of male subjects.

Strategy for setting the threshold	Threshold in years	Accuracy for minors, %	Accuracy for adults, %
Using the threshold to get lower errors for minors	18.73	98.1	88.0
Using the threshold to get as equal accuracy for adults and minors as possible	18.38	93.3	93.2
Using estimated age without modifying the threshold	18.00	90.4	95.7

Figure 11. Accuracies for minor versus adult classification of female subjects, using threshold to increase accuracy for minors.

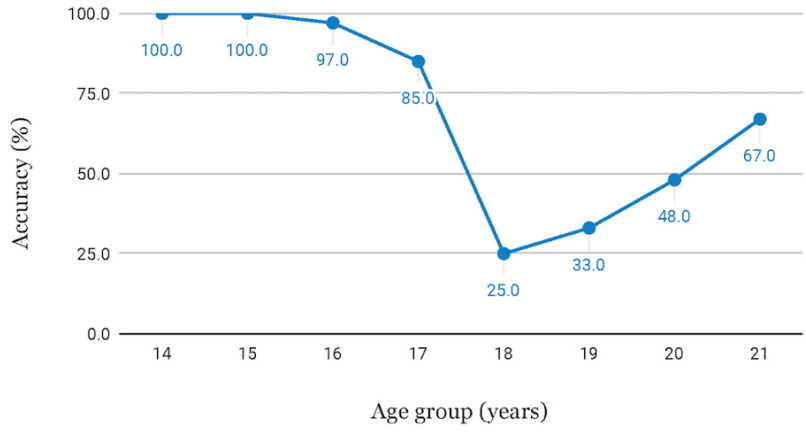


Table 5. Accuracies for minor versus adult classification of female subjects.

Strategy for setting the threshold	Threshold in years	Accuracy for minors, %	Accuracy for adults, %
Using threshold to get lower errors for minors	19.11	95.0	45.7
Using threshold to get as equal accuracy for adults and minors as possible	18.20	85.0	85.2
Using estimated age without modifying the threshold	18.00	77.0	88.9

Discussion

Principal Findings

This paper proposed a fully automated method, free from ionizing radiation, for age assessment based on MRI images of the knee using CNNs. The method was able to assess the age of male subjects in the range of 14-20.5 years of age, with an MAE of 0.793 years, and of female subjects in the range of 14-19.5 years of age, with an MAE of 0.988 years.

The method developed in this paper addresses and proposes solutions to the drawbacks in age-assessment research, which currently deals with the following:

1. Ethical issues of submitting healthy individuals to ionizing radiation for nontherapeutic purposes [10], since most of the established methods (ie, GP and TW) and recently published methods make use, mostly, of radiographs as the analysis input [23]. This paper showed that it is possible to achieve a good estimation of age by employing MRI images instead.
2. Lowering the risk of intra- and interrater variability, which can be very high when general radiologists are employed in the assessment of age instead of high-expertise pediatric radiologists [41,42]. Also, there is limited evidence that contrasts with the findings of manual raters and automatic systems regarding chronological age assessment, since most of the published material is directed to predict bone age [23]. However, a novel study reports a higher rate of false positives in classifying adults—with a threshold of 18 years—from hand images for manual raters compared to a deep learning system [43].
3. Time spent on assessment [9] addressed by the automation of the proposed method, which is able to perform evaluations in real time.

It is also important to mention that the proposed method in this paper provides the estimation of chronological age based on MRI images of the knee, contrary to most previous research, which aimed at estimating bone age and evaluating the methods using bone age and not chronological age. While the concept of bone age is certainly useful and important in many clinical settings, it was not conceived as a method to determine the chronological age of an individual. It was used to examine the developmental status of children and adolescents in comparison to their known chronological age, which can be advanced or delayed due to a multitude of factors that include chronic illnesses, hormonal disorders, etc [7,10]. The widespread use of BAA as an estimation of chronological age sometimes confuses these concepts and they are erroneously used interchangeably, as in many studies to justify the execution of BAA to judicial and civil issues. Also, it can be argued that the bone age attributed to an individual may be subjective and there is no objective way to obtain a confirmation of the exact number. In a clinical setting this may not be a problem since doctors can work with secure thresholds, but if the estimation is done for legal purposes it can become problematic, since decisions based on this estimation, especially regarding the ages of adulthood, can greatly affect the life of the individual in question.

Regarding our experiments, it is shown that for the male subjects, after the age of 20.5 the model could not identify any more information in the MRI images to discriminate the age of individuals. The same phenomenon occurred at the age of 19.5 for female subjects, which could indicate that the transformations that occur in the knee area related to the maturation process occur earlier in women than in men. This is in line with prior research on the knee region [12,24,44].

We also had satisfactory results for the problem regarding the classification of minors versus adults, considering the threshold of 18 years of age, which can be especially important in civil and judicial scenarios. Misclassification of minors as adults can often be viewed as much more problematic than the inverse, since the imputability for the application of laws, as well as guaranteed rights, may be different for these groups of individuals and usually harsher for adults. Our method can reduce that problem by distributing the errors depending on the application, using a modifiable threshold applied to the estimated age. Our method achieved an accuracy of 98.1% for male subjects and 95.0% for female subjects when it came to correctly classifying minors from the MRI images, when using a threshold that increased the accuracy for minors and sacrificed accuracy for adults.

From an operational point of view, the CNN technology employed with transfer learning can be seen as an enabler in performing research with medical images. The high cost for medical imaging can result in smaller datasets for many studies, but this caveat can be partially addressed when using the transfer learning technology on pretrained CNNs that have learned features from generic images. In this study, even if the features changed during training they were not changed much in our case. Generic features seem to work in a satisfactory way for MRI images; it is just detecting edges, corners, and blobs, which are relevant in MRI images as well as in generic images. Therefore, there is a possibility of applying automated methods even for smaller datasets. The study by Spampinato et al reported similar conclusions, but for radiographs of the hand [36].

Comparison With Prior Work

We propose a fully automated and radiation-free method for chronological age assessment based on MRI images of the knee region, employing deep learning techniques. We could not find prior published work with the same attributes in the literature, as not much work has been done in estimating chronological age per se.

A recent study by Stern et al [43] employed MRI volumes of the hand with CNNs in order to predict chronological age of male subjects from 13 up to 19 years of age. They reported an MAE of 0.82 years for subjects under 18 years of age. They also reported results on majority age classification for male subjects between the ages of 13 and 25 years. An error of 5% for minors gave an error of 27.5% for adults, and an error of 1% for minors gave an error of 67.2% for adults. This can be compared to our results where an error of 1.9% for minors gave an error of 12% for adults on male subjects between the ages of 14 and 22 years. In an earlier study by Stern et al [45], they proposed a multi-factorial age estimation method using MRI

volumes of the hand, clavicle, and teeth with CNNs. With this approach, they managed to predict chronological age of male subjects from 13 up to 25 years of age with an MAE of 1.01 years. They also reported results on majority age classification, where an error of 0.5% for minors gave an error of 25.0% for adults, and an error of 3% for minors gave an error of 18.1% for adults. This can be compared to our results, where an error of 1.9% for minors gave an error of 12% for adults on male subjects between 14 and 22 years of age. The results on majority age classification in these two papers by Stern et al [43,45] are the best published results so far, using one or multiple body parts. However, our results are significantly better even compared to their method using MRI data from three different body parts.

The study by Tang et al [46] proposed an artificial neural network model for estimating the chronological age of subjects (12-17 years old) using MRI images of the hand and wrist and other skeletal maturity factors of 79 subjects. In this study, the authors chose as the performance metric the comparison between the mean chronological age for all subjects and the mean estimated age for all subjects (ie, mean disparity), not calculating the error per subject, which could be misleading. The mean disparity measures whether there is a constant offset in the estimations, not the performance of the model on a per-subject level, like MAE does. A model can, therefore, have large errors in age estimation for all subjects and high MAE but can still have a small mean disparity; the MAE was not reported in this paper. Additionally, the reported results were on the validation set, probably due to the small sample size. In this fashion, the authors reported a mean disparity of 0.1 years between the estimated skeletal age and the chronological age.

Prior published methods for BAA that employed automated methods still focused mostly on the hand and wrist regions for the age assessment and made heavy use of radiographs as the input for their systems, as reported by a recent systematic literature review (SLR) and meta-analysis on BAA systems [23].

In this SLR, only two studies were reported to have made assessments based on the knee. The study by O'Connor et al [44] proposed a scoring system based on the assessment of knee radiographs as to the stage of epiphyseal fusion of the femur, tibia, and fibula on subjects from 9 to 19 years of age, employing regression model-building techniques. This study reported residuals of more than 2 bone-age years for both male and female individuals. The study by Fan et al [24] aimed to compare the age assessment based on the knee region from radiographs and MRI images on subjects from 11 to 25 years of age. They built regression models for bone age based on the scoring system by Krämer et al [47] for both image modalities, yielding better results for the MRI images, achieving R^2 values (eg, the variance in the dependent variable that is predicted from the independent variables in regression models) of 0.634 and 0.654 for female and male subjects, respectively.

On the choice of medical imaging, the referred SLR reported only three studies that built systems for BAA based on MRI images; one of these was the study by Tang et al [46], mentioned previously. The study by Urchsler et al [13] designed a system with the deep learning technology to automatically locate the ossification centers on MRI images of the hand and wrist to assess the bone age of individuals, 13-20 years of age, with random forests. This study obtained an MAE of 0.850 bone-age years. The study by Hillewig et al [48] obtained MRI images from the clavicle and radiograph images from the hand and wrist of 220 subjects, 16-26 years of age, and evaluated these regions according to the Schmeling et al [49] and Kreitner et al [50] scoring systems for the clavicle and the hand and wrist, respectively. The study concluded that the assessment of the clavicle alone was not sufficient to discriminate individuals as younger or older than 18 years of age, thus requiring the information from the hand and wrist for the assessment.

Another noninvasive and radiation-free medical imaging method for the estimation of age that is reported in the literature is the assessment of retinal images, which is an approach that provides diagnostic evidence about important diseases, such as cardiovascular disease and diabetes. Retinal images were assessed with deep learning in the study by Poplin et al [51] in predicting a variety of cardiovascular risk factors, including age, which achieved an MAE of 3.26 years. Retinal images were also assessed by Ting et al [52] in estimating the prevalence and systematic risk factors for diabetic retinopathy, which included young age.

In regard to approaches that make use of deep learning methods in the field of BAA, the biggest initiative posed in recent years was done so by the Radiological Society of North America (RSNA) for the prediction of bone age: the RSNA 2018 Pediatric Bone Age Challenge [53]. This challenge aimed to encourage participants to develop algorithms that could most-accurately determine the bone age of subjects from 0 to 19 years of age, providing a database of around 12,000 radiograph images of the hand and wrist, labeled as to their bone age [53]. The participants proposed CNN models, like the ones by Iglovikov et al [54], Zhao et al [55], and Ren et al [22], which achieved MAEs of 7.52, 7.66, and 5.2 months. However good the obtained results were, they were not comparable to our results, since our aim was to predict the chronological age of a subject, and the RSNA project's goal was to predict the bone age. It is also important to note that although these studies made use of large-enough sample sizes, the data were not uniformly distributed, as only 0.1% of the dataset was composed of individuals of 18 and 19 years of age. Additionally, Dallora et al [23] provided a meta-analysis on the performances based on seven studies, which contained all three deep learning studies mentioned previously, where the age ranges were mostly within 0-19 years of age and the performance metrics were given in MAE (bone-age months). The weighted average by the dataset size resulted in 9.96 MAE (bone-age months), which is higher than the results presented in this paper.

Limitations

Regarding the limitations of this study, it could be argued that the sample size would not be big enough to be generalizable; therefore, we employed methods to ensure that the models did not overfit by using test sets separated from the training and validation sets. The results showed that the model was able to generalize to new data in the test sets. Additionally, further

work will be directed to the collection of more data, which may improve the precision and MAE of our models.

Also, we aimed at having a uniform number of subjects for each age group, which was achieved by the data acquisition process; an exception was for the 19-year-old female subjects, who accounted for only 12 subjects, which could be seen as a caveat to the female model.

Additionally, the acquisition of ages for the first half year from each age group may interfere with the estimation accuracy of the minor versus adult classification. The largest impact occurs for the ages closest to 18 years. The missing data for those 17.5-17.99 years of age is important and we plan to collect new data to complement those ages in future work. Concerning the MAE numbers, these missing ages do not have as much impact as for the accuracy numbers.

Finally, the method was built upon data from healthy youth and young adult subjects and the effect of disorders that can affect growth was not explored.

Conclusions

This paper proposed a model for the estimation of chronological age in youth and young adults using MRI images of the knee. Our method demonstrated good results and addressed the biggest drawbacks in the traditional age-estimation procedures that are still currently in use. Our results on majority age classification were significantly better than the best results previously published.

Acknowledgments

We would like to express our greatest appreciation to the participants and staff who took part in our study. This work was supported by the National Board of Health and Welfare of Sweden (Socialstyrelsen). The funding source had no involvement regarding study design, data collection, analysis, interpretation, or reporting of this work.

Conflicts of Interest

None declared.

References

1. Gilsanz V, Ratib O. Hand Bone Age: A Digital Atlas of Skeletal Maturity. Berlin, Germany: Springer-Verlag; 2005.
2. Manzoor Mughal A, Hassan N, Ahmed A. Bone age assessment methods: A critical review. Pak J Med Sci 2014 Jan;30(1):211-215 [FREE Full text] [doi: 10.12669/pjms.301.4295] [Medline: 24639863]
3. Satoh M. Bone age: Assessment methods and clinical applications. Clin Pediatr Endocrinol 2015 Oct;24(4):143-152 [FREE Full text] [doi: 10.1297/cpe.24.143] [Medline: 26568655]
4. Cunha E, Baccino E, Martrille L, Ramsthaler F, Prieto J, Schuliar Y, et al. The problem of aging human remains and living individuals: A review. Forensic Sci Int 2009 Dec 15;193(1-3):1-13. [doi: 10.1016/j.forsciint.2009.09.008] [Medline: 19879075]
5. Fatehi M, Nateghi R, Pourakpour F. Automatic bone age determination using wrist MRI based on FIFA grading system for athletes: Deep learning approach. In: Proceedings of the 26th Annual Scientific Meeting of the European Society of Musculoskeletal Radiology (ESSR). 2019 Presented at: 26th Annual Scientific Meeting of the European Society of Musculoskeletal Radiology (ESSR); June 26-29, 2019; Lisbon, Portugal. [doi: 10.1055/s-0039-1692580]
6. Dvorak J, George J, Junge A, Hodler J. Application of MRI of the wrist for age determination in international U-17 soccer competitions. Br J Sports Med 2007 Aug;41(8):497-500 [FREE Full text] [doi: 10.1136/bjsm.2006.033431] [Medline: 17347314]
7. Greulich WW, Pyle SI. Radiographic atlas of skeletal development of the hand and wrist. Am J Med Sci 1959;238(3):393. [doi: 10.1097/00000441-195909000-00030]
8. Tanner JM, Whitehouse RH, Cameron N, Marshall WA, Healy MJR, Goldstein H. Assessment of Skeletal Maturity and Prediction of Adult Height (TW2 Method). 2nd edition. London, UK: Academic Press; 1975.
9. Mansourvar M, Ismail M, Herawan T, Raj R, Kareem S, Nasaruddin F. Automated bone age assessment: Motivation, taxonomies, and challenges. Comput Math Methods Med 2013;2013:391626 [FREE Full text] [doi: 10.1155/2013/391626] [Medline: 24454534]
10. Hjern A, Brendler-Lindqvist M, Norredam M. Age assessment of young asylum seekers. Acta Paediatr 2012 Jan;101(1):4-7. [doi: 10.1111/j.1651-2227.2011.02476.x] [Medline: 21950617]
11. Crema MD, Roemer FW, Marra MD, Burstein D, Gold GE, Eckstein F, et al. Articular cartilage in the knee: Current MR imaging techniques and applications in clinical practice and research. Radiographics 2011;31(1):37-61. [doi: 10.1148/rg.311105084] [Medline: 21257932]
12. Margalit A, Cottrill E, Nhan D, Yu L, Tang X, Fritz J, et al. The spatial order of physeal maturation in the normal human knee using magnetic resonance imaging. J Pediatr Orthop 2019;39(4):e318-e322. [doi: 10.1097/bpo.0000000000001298]
13. Urschler M, Grassegger S, Štern D. What automated age estimation of hand and wrist MRI data tells us about skeletal maturation in male adolescents. Ann Hum Biol 2015;42(4):358-367. [doi: 10.3109/03014460.2015.1043945] [Medline: 26313328]

14. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8-17 [FREE Full text] [doi: 10.1016/j.csbj.2014.11.005] [Medline: 25750696]
15. Chen M, Hao Y, Hwang K, Wang L, Wang L. Disease prediction by machine learning over big data from healthcare communities. IEEE Access 2017;5:8869-8879. [doi: 10.1109/access.2017.2694446]
16. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Brief Bioinform 2006 Mar;7(1):86-112. [doi: 10.1093/bib/bbk007] [Medline: 16761367]
17. Agarwal V, Zhang L, Zhu J, Fang S, Cheng T, Hong C, et al. Impact of predicting health care utilization via Web search behavior: A data-driven analysis. J Med Internet Res 2016 Sep 21;18(9):e251 [FREE Full text] [doi: 10.2196/jmir.6240] [Medline: 27655225]
18. van Hartskamp M, Consoli S, Verhaegh W, Petkovic M, van de Stolpe A. Artificial intelligence in clinical health care applications: Viewpoint. Interact J Med Res 2019 Apr 05;8(2):e12100 [FREE Full text] [doi: 10.2196/12100] [Medline: 30950806]
19. Shen D, Wu G, Suk H. Deep learning in medical image analysis. Annu Rev Biomed Eng 2017 Jun 21;19(1):221-248 [FREE Full text] [doi: 10.1146/annurev-bioeng-071516-044442] [Medline: 28301734]
20. Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. IEEE J Biomed Health Inform 2017 Jan;21(1):4-21. [doi: 10.1109/jbhi.2016.2636665]
21. Miotto R, Wang F, Wang S, Jiang X, Dudley J. Deep learning for healthcare: Review, opportunities and challenges. Brief Bioinform 2018 Nov 27;19(6):1236-1246 [FREE Full text] [doi: 10.1093/bib/bbx044] [Medline: 28481991]
22. Ren X, Li T, Yang X, Wang S, Ahmad S, Xiang L, et al. Regression convolutional neural network for automated pediatric bone age assessment from hand radiograph. IEEE J Biomed Health Inform 2019 Sep;23(5):2030-2038. [doi: 10.1109/JBHI.2018.2876916] [Medline: 30346295]
23. Dallora AL, Anderberg P, Kvist O, Mendes E, Diaz Ruiz S, Sanmartin Berglund J. Bone age assessment with various machine learning techniques: A systematic literature review and meta-analysis. PLoS One 2019 Jul 25;14(7):e0220242 [FREE Full text] [doi: 10.1371/journal.pone.0220242] [Medline: 31344143]
24. Fan F, Zhang K, Peng Z, Cui J, Hu N, Deng Z. Forensic age estimation of living persons from the knee: Comparison of MRI with radiographs. Forensic Sci Int 2016 Nov;268:145-150. [doi: 10.1016/j.forsciint.2016.10.002] [Medline: 27770721]
25. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 7-12, 2015; Boston, MA. [doi: 10.1109/cvpr.2015.7298594]
26. Ker J, Wang L, Rao J, Lim T. Deep learning applications in medical image analysis. IEEE Access 2018;6:9375-9389. [doi: 10.1109/access.2017.2788044]
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30 , 2016; Las Vegas, NV URL: http://toc.proceedings.com/32592webtoc.pdf [doi: 10.1109/cvpr.2016.90]
28. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 27-30, 2016; Las Vegas, NV. [doi: 10.1109/cvpr.2016.308]
29. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd International Conference on Learning Representations. 2015 Presented at: 3rd International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: http://arxiv.org/abs/1409.1556
30. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM 2017 May 24;60(6):84-90. [doi: 10.1145/3065386]
31. Huang G, Liu Z, Van DML, Weinberger K. Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; July 21-26 , 2017; Honolulu, HI. [doi: 10.1109/cvpr.2017.243]
32. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention. 2015 Presented at: 18th International Conference on Medical Image Computing and Computer-Assisted Intervention; October 5-9, 2015; Munich, Germany. [doi: 10.1007/978-3-319-24574-4_28]
33. Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R. Caffe: Convolutional Architecture for Fast Feature Embedding. In: Proceedings of the 22nd ACM International Conference on Multimedia. 2014 Presented at: 22nd ACM International Conference on Multimedia; November 3-7, 2014; Orlando, FL. [doi: 10.1145/2647868.2654889]
34. Kingma D, Ba J. Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations. 2015 Presented at: International Conference on Learning Representations; May 7-9, 2015; San Diego, CA URL: http://arxiv.org/abs/1412.6980
35. Kumar A, Kim J, Lyndon D, Fulham M, Feng D. An ensemble of fine-tuned convolutional neural networks for medical image classification. IEEE J Biomed Health Inform 2017 Jan;21(1):31-40. [doi: 10.1109/jbhi.2016.2635663]

36. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal 2017 Feb;36:41-51. [doi: 10.1016/j.media.2016.10.010] [Medline: 27816861]
37. Deng J, Dong W, Socher R, Li L, Li K, Fei-Fei L. ImageNet: A large-scale hierarchical image database. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2009 Presented at: IEEE Conference on Computer Vision and Pattern Recognition; June 20-25, 2009; Miami, FL. [doi: 10.1109/cvprw.2009.5206848]
38. Zhong Z, Zheng L, Kang G, Li S, Yang Y. arXiv. 2017. Random erasing data augmentation URL: https://arxiv.org/pdf/1708.04896.pdf [accessed 2019-11-20]
39. Lee H, Tajmir S, Lee J, Zissen M, Yeshiwas BA, Alkasab TK, et al. Fully automated deep learning system for bone age assessment. J Digit Imaging 2017 Aug 8;30(4):427-441 [FREE Full text] [doi: 10.1007/s10278-017-9955-8] [Medline: 28275919]
40. Dedouit F, Auriol J, Rousseau H, Rougé D, Crubézy E, Telmon N. Age assessment by magnetic resonance imaging of the knee: A preliminary study. Forensic Sci Int 2012 Apr 10;217(1-3):232.e1-232.e7. [doi: 10.1016/j.forsciint.2011.11.013] [Medline: 22153621]
41. Kaplowitz P, Srinivasan S, He J, McCarter R, Hayeri MR, Sze R. Comparison of bone age readings by pediatric endocrinologists and pediatric radiologists using two bone age atlases. Pediatr Radiol 2011 Jun 16;41(6):690-693. [doi: 10.1007/s00247-010-1915-0] [Medline: 21161206]
42. Shen J, Zhang CJP, Jiang B, Chen J, Song J, Liu Z, et al. Artificial intelligence versus clinicians in disease diagnosis: Systematic review. JMIR Med Inform 2019 Aug 16;7(3):e10010 [FREE Full text] [doi: 10.2196/10010] [Medline: 31420959]
43. Štern D, Payer C, Urschler M. Automated age estimation from MRI volumes of the hand. Med Image Anal 2019 Dec;58:101538 [FREE Full text] [doi: 10.1016/j.media.2019.101538] [Medline: 31400620]
44. O'Connor JE, Coyle J, Bogue C, Spence LD, Last J. Age prediction formulae from radiographic assessment of skeletal maturation at the knee in an Irish population. Forensic Sci Int 2014 Jan;234:188.e1-188.e8. [doi: 10.1016/j.forsciint.2013.10.032] [Medline: 24262807]
45. Stern D, Payer C, Giuliani N, Urschler M. Automatic age estimation and majority age classification from multi-factorial MRI data. IEEE J Biomed Health Inform 2019 Jul;23(4):1392-1403. [doi: 10.1109/jbhi.2018.2869606]
46. Tang FH, Chan JL, Chan BK. Accurate age determination for adolescents using magnetic resonance imaging of the hand and wrist with an artificial neural network-based approach. J Digit Imaging 2019 Apr 15;32(2):283-289. [doi: 10.1007/s10278-018-0135-2] [Medline: 30324428]
47. Krämer JA, Schmidt S, Jürgens KU, Lentschig M, Schmeling A, Vieth V. Forensic age estimation in living individuals using 3.0 T MRI of the distal femur. Int J Legal Med 2014 May 7;128(3):509-514. [doi: 10.1007/s00414-014-0967-3] [Medline: 24504560]
48. Hillewig E, Degroote J, Van der Paelt T, Visscher A, Vandemaele P, Lutin B, et al. Magnetic resonance imaging of the sternal extremity of the clavicle in forensic age estimation: Towards more sound age estimates. Int J Legal Med 2013 May 9;127(3):677-689. [doi: 10.1007/s00414-012-0798-z] [Medline: 23224029]
49. Schmeling A, Schulz R, Reisinger W, Mühler M, Wernecke K, Geserick G. Studies on the time frame for ossification of the medial clavicular epiphyseal cartilage in conventional radiography. Int J Legal Med 2004 Feb 1;118(1):5-8. [doi: 10.1007/s00414-003-0404-5] [Medline: 14534796]
50. Kreitner K, Schweden FJ, Riepert T, Nafe B, Thelen M. Bone age determination based on the study of the medial extremity of the clavicle. Eur Radiol 1998 Sep 2;8(7):1116-1122. [doi: 10.1007/s003300050518] [Medline: 9724422]
51. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2018 Mar 19;2(3):158-164. [doi: 10.1038/s41551-018-0195-0] [Medline: 31015713]
52. Ting DSW, Cheung CY, Nguyen Q, Sabanayagam C, Lim G, Lim ZW, et al. Deep learning in estimating prevalence and systemic risk factors for diabetic retinopathy: A multi-ethnic study. NPJ Digit Med 2019 Apr 10;2(1):24 [FREE Full text] [doi: 10.1038/s41746-019-0097-x] [Medline: 31304371]
53. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, et al. The RSNA Pediatric Bone Age Machine Learning Challenge. Radiology 2019 Feb;290(2):498-503. [doi: 10.1148/radiol.2018180736] [Medline: 30480490]
54. Iglovikov V, Rakhlin A, Kalinin A, Shvets A. Paediatric bone age assessment using deep convolutional neural networks. In: Proceedings of the 4th International Workshop on Deep Learning in Medical Image Analysis. 2018 Presented at: 4th International Workshop on Deep Learning in Medical Image Analysis; September 20, 2018; Granada, Spain. [doi: 10.1101/234120]
55. Zhao C, Han J, Jia Y, Fan L, Gou F. Versatile framework for medical image processing and analysis with application to automatic bone age assessment. J Electr Comput Eng 2018 Dec 31;2018:1-13. [doi: 10.1155/2018/2187247]

Abbreviations

BAA: bone age assessment
Caffe: Convolutional Architecture for Fast Feature Embedding
CNN: convolutional neural network

EC2: Elastic Compute Cloud
GP: Greulich-Pyle
MAE: mean absolute error
MRI: magnetic resonance imaging
RSNA: Radiological Society of North America
SLR: systematic literature review
TW: Tanner-Whitehouse
VGG: Visual Geometry Group

Edited by G Eysenbach; submitted 18.09.19; peer-reviewed by A Korchi, L Zhang, G Lim; comments to author 08.10.19; revised version received 31.10.19; accepted 13.11.19; published 05.12.19

Please cite as:
Dallora AL, Berglund JS, Brogren M, Kvist O, Diaz Ruiz S, Dübbel A, Anderberg P
Age Assessment of Youth and Young Adults Using Magnetic Resonance Imaging of the Knee: A Deep Learning Approach
JMIR Med Inform 2019;7(4):e16291
URL: http://medinform.jmir.org/2019/4/e16291/
doi: 10.2196/16291
PMID: 31804183

Study IV

16 Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review

PLOS ONE

RESEARCH ARTICLE

Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review

Ana Luiza Dallora[1☯]*, Shahryar Eivazzadeh[2☯], Emilia Mendes[1‡], Johan Berglund[2‡], Peter Anderberg[2‡]

1 Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden, **2** Department of Health, Blekinge Institute of Technology, Karlskrona, Sweden

☯ These authors contributed equally to this work.
‡ These authors also contributed equally to this work.
* ana.luiza.moraes@bth.se

OPEN ACCESS

Citation: Dallora AL, Eivazzadeh S, Mendes E, Berglund J, Anderberg P (2017) Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. PLoS ONE 12(6): e0179804. https://doi.org/10.1371/journal.pone.0179804

Editor: Kewei Chen, Banner Alzheimer's Institute, UNITED STATES

Received: January 10, 2017

Accepted: June 5, 2017

Published: June 29, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files. The protocol of the systematic literature review is available from GitHub at: https://goo.gl/6Jddw3.

Funding: The authors of the research presented in this article received no grant from any funding agency.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Background

Dementia is a complex disorder characterized by poor outcomes for the patients and high costs of care. After decades of research little is known about its mechanisms. Having prognostic estimates about dementia can help researchers, patients and public entities in dealing with this disorder. Thus, health data, machine learning and microsimulation techniques could be employed in developing prognostic estimates for dementia.

Objective

The goal of this paper is to present evidence on the state of the art of studies investigating and the prognosis of dementia using machine learning and microsimulation techniques.

Method

To achieve our goal we carried out a systematic literature review, in which three large databases—Pubmed, Socups and Web of Science were searched to select studies that employed machine learning or microsimulation techniques for the prognosis of dementia. A single backward snowballing was done to identify further studies. A quality checklist was also employed to assess the quality of the evidence presented by the selected studies, and low quality studies were removed. Finally, data from the final set of studies were extracted in summary tables.

Results

In total 37 papers were included. The data summary results showed that the current research is focused on the investigation of the patients with mild cognitive impairment that will evolve to Alzheimer's disease, using machine learning techniques. Microsimulation studies were concerned with cost estimation and had a populational focus. Neuroimaging was the most commonly used variable.

Conclusions

Prediction of conversion from MCI to AD is the dominant theme in the selected studies. Most studies used ML techniques on Neuroimaging data. Only a few data sources have been recruited by most studies and the ADNI database is the one most commonly used. Only two studies have investigated the prediction of epidemiological aspects of Dementia using either ML or MS techniques. Finally, care should be taken when interpreting the reported accuracy of ML techniques, given studies' different contexts.

Introduction

Dementia is a complex disorder that affects the brain. It is most prevalent in the elderly population, responsible for a progressive cognitive decline severe enough to interfere with the patient's daily functioning and independence. Although decades of research have been dedicated to studying it, little is known about its mechanisms and there is still no disease modifying treatment that is able to stop or significantly delay its progression [1]. The most common form of dementia pathology is the accumulation of amyloid plaques in the brain and tau proteins inside the neurons. Amyloid plaques are very small in size (about 0.1 mm) and are formed by protein fragments Aβ, surrounded by dysfunctional neurons, whilst tau proteins accumulated inside the neurons form fibrillary tangles [2]. Together, these two factors are believed to be highly correlated to the neurodegeneration process [2].

Beyond the loss of independence, studies estimate that persons with dementia face mortality risks two times higher than that for similar groups without dementia [3] and deal with 2 to 8 additional chronic diseases that may accelerate their decline in daily functioning [1,4]. There are also consequences for the caregivers, especially for the family of the affected persons, who report low confidence in managing the condition, high levels of strain and depressive symptoms [5].

The demographic changes, with an increasing number of older people worldwide, will dramatically increase the cost in health and care programs. In 2011 the global estimated number of people with dementia was 35.6 million, and the trend points to a 100% increase within 20 years [6]. In comparison to other chronic disorders, in 2010 the global direct cost (prevention and treatment) and indirect cost (owing to mortality and morbidity) of cancer and diabetes were respectively $290 billion and $472 billion, while in 2014 the direct cost of Alzheimer's Disease (AD), in USA alone, was of $214 billion [2].

Given that dementia is a serious disorder that brings so many challenges to patients, caregivers and public entities, and for which research on treatments are still on course, it is extremely important to investigate dementia's prognosis. Prognostic estimates can aid researchers in finding patterns on disease progression, support public entities in allocating resources for the creation and maintenance of healthcare programs, and also aid patients and their caregivers in understanding more about their condition [7]. To be able to derive such useful estimates about dementia, reliable patient data is needed, like the ones from randomized clinical trials e.g. the Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER), and the Healthy Aging Through Internet Counseling in the Elderly (HATICE), or other study initiatives/consortiums e.g. the Swedish National Study on Aging and Care (SNAC), Alzheimer's Disease Neuroimaging Initiative (ADNI), and European Alzheimer's Disease Consortium Impact of Cholinergic Treatment Use (EADC-ICTUS).

The existence of health data allows for the execution of analyses that can derive several types of prognostic estimates. Two data analysis approaches that are specially focused in prediction and could be of great service to prognostic studies are: machine learning (ML) and microsimulation (MS) [8,9].

ML is already widely employed in biological domains such as genomics, proteomics, microarrays, systems biology, evolution and text mining [8]. It comprises a group of techniques that are able to learn from a set of examples (training set) to perform a task, so that this task can be performed with a completely new data set [10]. The most common learning approaches for ML techniques are supervised or unsupervised learning. When a supervised learning approach is used, the training set is composed by labeled examples (input and output variables). The most common tasks that use such approach are classification, in which the data is categorized in a finite number of classes; and regression, in which a function maps input variables to a desirable output variable [11]. Unsupervised learning happens when the data is not labeled, so the algorithm will work to find patterns that describe the data. Clustering is a common employed task, and characterizes the partitioning of a data set following a certain criteria [10]. Depending on the available health data and problem that needs to be solved, both supervised and unsupervised approaches can be used for prognostic estimates [11].

In the past there has been a number of studies that used standard statistics for disease prediction and prognosis (e.g. cancer, dementia). Such studies were feasible because "our dependency on macro-scale information (tumor, patient, population, and environmental data) generally kept the numbers of variables small enough so that standard statistical methods or even a physician's own intuition could be used to predict cancer risks and outcomes" [12]. However, the world has changed into a reality where high-throughput diagnostic and imaging technologies are used, which, as a consequence lead to an overwhelming number of molecular, cellular and clinical parameters [12]. This has been the case in cancer as well as in dementia research, amongst other diseases. In such circumstances, as clearly stated by Cruz and Wishart [12] "human intuition and standard statistics don't generally work. Instead we must increasingly rely on nontraditional, intensively computational approaches such as machine learning. The use of computers (and machine learning) in disease prediction and prognosis is part of a growing trend towards personalized, predictive medicine". Such argument is also shared by others, such as Kourou et al. [10], who have even explicitly removed from their Mini review any studies that employed conventional statistical methods (e.g. chi-square, Cox regression). Finally, we also share the same view as Cruz and Wishart [12] with respect to the advantages that ML techniques provide, when compared to standard statistics: "Machine learning, like statistics, is used to analyze and interpret data. Unlike statistics, though, machine learning methods can employ Boolean logic (AND, OR, NOT), absolute conditionality (IF, THEN, ELSE), conditional probabilities (the probability of X given Y) and unconventional optimization strategies to model data or classify patterns"; further, ML "still draws heavily from statistics and probability, but it is fundamentally more powerful because it allows inferences or decisions to be made that could not otherwise be made using conventional statistical methodologies..". [12]

One point to note is that the studies that use ML techniques for prognosis deal mostly with individuals as their unit of study. However, prognosis can also be extended beyond individuals to also include populations (e.g. studies by Suh and Shah [13] and Jagger et al [14]). To focus upon populations may be a suitable choice for example, when addressing the dementia family of diseases, as their long-term presence and considerable direct or indirect costs require significant investment, economic arrangements, and development of care facilities & infrastructure. Therefore, to address dementia prognosis in populations, we included MS methods, as this is a technique that has been traditionally used for prediction in populations.

MS models are closely related to an agent-based simulation model and aim to model individuals in a specific context though time [15,16]. The result of this simulation can give insights about the overall future of that population. MS has been used in healthcare to study how the screening programs can change morbidity and mortality rates or to estimate the economic aspects of diagnosis in specific diseases [9]. The same rationale could be applied in prognosis of dementia-related diseases and to apply MS as means to obtain insights on dementia prognostic isolation at the population level (in contrast to an individual level).

Given the abovementioned motivation, this paper aims to detail a systematic literature review that investigates the state of the art on how the ML and MS techniques are currently being applied to the development of prognostic estimates of dementia, aiming to answer the research question: "How are the machine learning and microsimulation techniques being employed by the researches on the prognosis of dementia and comorbidities?".

This paper is organized as follows: the Method section presents the approach followed to conduct the review; the Results section presents summarized data from the included studies; the Discussion section argues about the results and presents threats to validity; and the Conclusion section presents final statements and comments on future work.

Methods

A systematic literature review (SLR) identifies, evaluates and interprets a significant and representative sample of all of the pertinent primary studies of the literature concerning a topic of the research. SLRs execute a comprehensive search following a preset method that specifies focused research questions, criteria for the selection of studies and assessment of their quality, and forms to execute the data extraction and synthesis of results [17]. Among the motivations for conducting a SLR, the most common are: to summary all the evidence about a topic; to find gaps in the research; to provide a ground for a fundament to new research; and to examine how the current research supports a hypothesis. Performing a SLR comprises the following steps: (i) identify the need for performing the SLR; (ii) formulate research questions; (iii) execute a comprehensive search and selection of primary studies; (iv) assess the quality and extract data from the studies; (v) interpret the results; and (vi) report the SLR [18,19].

The SLR reported herein is part of a multidisciplinary project, in which five participants with different expertise (health, machine learning and bioinformatics) took part. Throughout the text, references to the authors will use a notation, in which A1 refers to the first author; A2 refers to the second author, and so forth.

The main research question this SLR aims to address is: "How are the machine learning and microsimulation techniques being employed by the researches on the prognosis of dementia and comorbidities?". This main question was decomposed further into five research questions:

- RQ1: Which ML and MS techniques are being used in the dementia and comorbidities research?
- RQ2: What data characteristics (variables, determinants and indicators) are being considered when applying the ML or/and MS techniques (physiological, demographic/social, genetics, lifestyle etc)?
- RQ3: What are the goals of the studies that employ ML or MS techniques for prognosis of dementia and comorbidities?
- RQ4: How is data censoring being handled in the studies?
- RQ5: Do the studies focus on individuals or populations?

Partial results for questions RQ2 and RQ3 were the subject of a previous publication by the same authors [18]. The present paper builds upon these questions and additionally presents the results of the other two additional research questions.

Further, the key terms related to comorbidities were included in the search string to ensure that relevant studies about ML or MS for the prognosis of a disease, where dementia is considered a comorbidity to that disease would also be retrieved from the database searches, even when the term dementia was not mentioned in the paper's title or abstract.

The protocol that guided the execution of this SLR is available at https://goo.gl/6Jddw3

Search strategy

To address the research questions, a search string was defined using the PICO approach, which decomposes the main question into four parts: population, intervention, comparison and outcome [19]. The comparison component was discarded because the SLR was mainly concerned with a characterization. For each of the remaining components, keywords were derived and their rationale can be represented as follows:

- **Population**: Studies that present research on dementia and comorbidities. Dementia's keywords were selected from the "Systematized Nomenclature of Medicine–Clinical Terms" and selected by A4. Comorbidities' keywords were extracted from the Marengoni et al. SLR in this topic [20].
- **Intervention**: ML or MS techniques. The ML keywords were selected from the branch "Machine Learning Approaches" of the "2012 ACM Computing Classification System". The MS keywords were selected by A2.
- **Outcome**: Prognosis on dementia and comorbidies. The prognosis keywords were provided by A4.

The automated searches were performed in the Pubmed, Web of Science and Scopus databases. Table 1 shows the search string used for the Pubmed automated search, but note that this search string was adapted to each of the other databases' search context.

Table 1. Search string used in the Pubmed automated search.

Search Date	October 23rd of 2015
("Dimentia" OR "Dementia" OR "Alzheimer" OR "Mixed Dementia" OR "Vascular Dementia" OR "Lewy Bodies" OR "Parkinson" OR "Creutzfeldt-Jakob" OR "Normal pressure hydrocephalus" OR "Huntington disease" OR "Wernicke-Korsakoff Syndrome" OR "Frontotemporal Dementia" OR "Neurosyphilis" OR "complex of Guam" OR "Subcortical leukoencephalopathy" OR "Comorbidities" OR "Comorbidity" OR "Comorbidity" OR "multimorbidity" OR "multimorbidities" OR "multi-morbidity") AND ("Machine Learning" OR "Data Mining" OR "Decision Support System" OR "Clinical Support System") AND ("Classification" OR "Regression" OR "Kernel" OR "Support vector machines" OR "Gaussian process" OR "Neural networks" OR "Logical learning" OR "relational learning" OR "Inductive logic" OR "Statistical relational" OR "probabilistic graphical model" OR "Maximum likelihood" OR "Maximum entropy" OR "Maximum a posteriori" OR "Mixture model" OR "Latent variable model" OR "Bayesian network" OR "linear model" OR "Perceptron algorithm" OR "Factorization" OR "Factor analysis" OR "Principal component analysis" OR "Canonical correlation" OR "Latent Dirichlet allocation" OR "Rule learning" OR "Instance-based" OR "Markov" OR "Stochastic game" OR "Learning latent representation" OR "Deep belief network" OR "Bio-inspired approach" OR "Artificial life" OR "Evolvable hardware" OR "Genetic algorithm" OR "Genetic programming" OR "Evolutionary robotic" OR "Generative and developmental approaches" OR "microsimulation" OR "micro-simulation" OR "microanalytic simulation" OR "agent-based modeling") AND ("prognosis" OR "prognostic estimate" OR "predictor" OR "prediction" OR "model" OR "patterns" OR "diagnosis" OR "diagnostic" OR "Forecasting" OR "projection")	

https://doi.org/10.1371/journal.pone.0179804.t001

Study selection

The first step of the study selection was the execution of an evaluation round with 100 random papers from the 593 results returned from the automated searches. These had their title and abstracts assessed by A1, A2 and A3, according to inclusion and exclusion criteria defined previously in the protocol (see Table 2). This step was mainly concerned in maintaining the consistency of the selection between the participants throughout the SLR.

The remaining 493 results had their title and abstracts assessed by A1 and A2, according to the inclusion and exclusion criteria. After the evaluations, 37 papers were selected. Then a one-iteration backward snowballing was carried out looking for possible additional studies. The 1199 new identified studies were assessed analogously as the previous ones, resulting in 41 new selected papers. Throughout the whole selection process, A3 and A4 acted in conflict resolution in the case where A1 and A2 couldn't reach an agreement.

In total, 78 papers were selected to be fully read and assessed regarding its eligibility. The ones that successfully passed the established criteria previously defined in the protocol, had their relevant data extracted.

In order to minimize the chance of selecting studies with bias evidence, a quality assessment questionnaire was used. This questionnaire was adapted from Kitchenham's guidelines [18] and can be found in the SLR protocol. If the grading attributed to a paper fell below 8 points (out of a total of 12), it would be rejected for quality reasons. The 8-point threshold was decided in the research group discussions involving all the authors. In this phase, a paper could also be rejected due to inclusion and exclusion criteria because the selection process adopted an inclusive approach. This means that during the reading of the titles and abstracts, in the case where the information provided was incomplete or too general it was selected to be fully read in the posterior phase. A common example is the case when the data analysis technique specified in the abstract was merely "classification", so it was not possible to know if any machine learning occurred.

As in the study selection, a quality assessment evaluation round was performed beforehand to ensure consistency in the evaluations. A1, A2 and A5 participated in this task.

In total, 37 studies composed the final set of included primary studies and had their relevant data extracted, 7 papers were rejected due quality reasons, and 34 papers were rejected due to failing the inclusion and exclusion criteria. One reason for the high number of the latter was the decision to exclude the papers that used solely statistical methods as data analysis techniques to build the prognostic models.

The selected studies were also assessed for the risk of cumulative evidence bias. This was done by checking, in the case of the same research group with different studies in the final set of included primary studies, if it was justified having both studies (i.e different samples).

Data collection

For the data collection, a base extraction form was defined in the protocol, but later in the study it was evolved based on the research group discussions. Table 3 lists and defines the collected variables.

Table 2. Inclusion and exclusion criteria for assessing the studies returned by the searches.

Inclusion Criteria	Exclusion Criteria
Be a primary study in English; AND address research on dementia and comorbidities; AND address at least one ML or MS technique; AND address a prognosis related to dementia and comorbidities.	Be a secondary or tertiary study; OR be written in another language other than English; OR do not address a research on dementia and comorbidities; OR do not address at least one ML or MS technique; OR do not address a prognosis related to dementia and comorbidities.

https://doi.org/10.1371/journal.pone.0179804.t002

Table 3. List of collected variables and their definitions.

Variable	Definition
Conditions Studied	For which dementia disorder is the study deriving a prognosis.
Database used in the study	Name and origin of the data source used to derive the prognosis of the studied dementia.
Dataset Categories	Classes in which the data units were divided into.
Handling of censored data	Description of the way in which censored data was handled.
Follow-up Period	Period of time, which the data units were followed.
Data Analysis Techniques	ML or MS techniques that were used to build the prognostic models.
Model Variables	The variables used in building the prognostic models.
Aim of the Study	The goal of the built prognostic models.
Focus of the Study	If the built prognostic models aim its predictions on an individual or population level.

https://doi.org/10.1371/journal.pone.0179804.t003

In addition to these variables other basic data about the studies was collected, these were: title, authors, journal/source, year and type of publication. No summary measures were used.

Summary tables were used for the synthesis of results and no additional analyses were carried out.

Results

The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Flow chart that describes the selection of the articles is shown in Fig 1.

A total of 78 results were assessed for eligibility having 37 studies selected as part of the final set of included primary studies, 7 studies excluded for falling out of the threshold of the quality assessment (8 out of 12), 3 studies excluded for not being a primary study, 7 studies excluded for not being about a prognostic estimate, 23 studies excluded for not making use of a ML or MS technique, and 1 study excluded for dealing with cognitive decline, but not dementia specifically (see S1 Table).

Three groups of common authors had 2 papers included in the final set of included studies, these were: Zhang, Daoqiang and Shen, Dinggang; Llano, Daniel A. and Devanarayan, Viswanath; and Moradi, Elaheh, Tohka, Jussi and Gaser, Christian. After the assessment for possible bias it was found that in these cases, either the sample varied or the categories of variables changed, not representing cumulative evidence bias to the SLR.

Fig 2 shows the frequency of primary studies per year of publication. It has to be remarked that the frequency showed for the year 2015 refers to studies published until October, when the search was performed.

Identified machine learning techniques

This section presents the results that address research question RQ1: "Which ML and ML techniques are being used in the dementia and comorbidities research?"

Regarding the ML techniques, the synthesis of the extracted data shows that the most frequently used ML techniques were Support Vector Machines (SVM) (30 studies), Decision Trees (DT) (6 studies), Bayesian Networks (BN) (6 studies) and Artificial Neural Networks (ANN) (3 studies). These results are consistent with the cancer prognosis research, which also lists ANN, DT, SVM and BN as the ones most widely used [10]. In the cancer field, SVMs are relatively new algorithms that have been widely used due to its predictive performance, ANNs

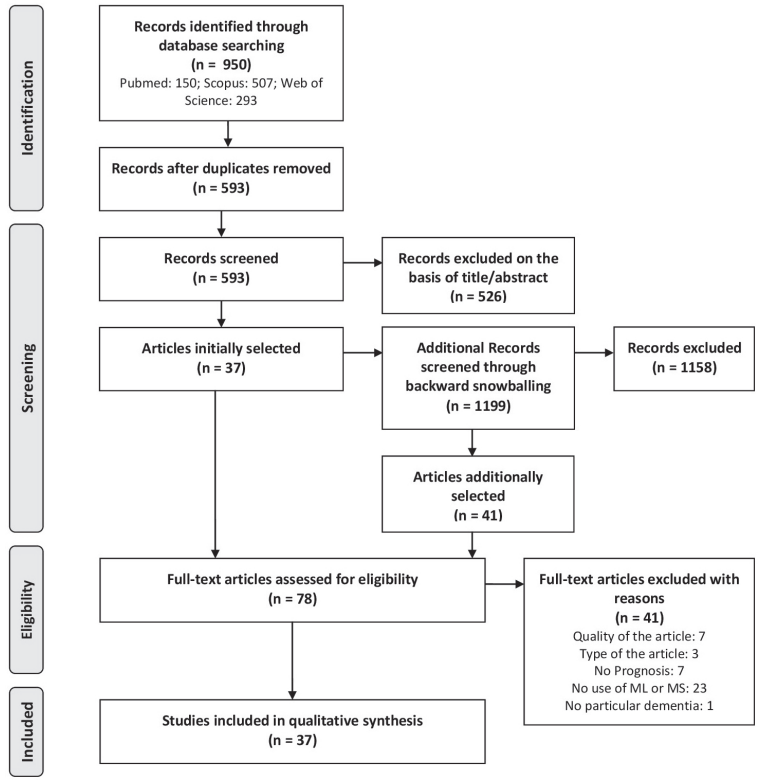


Fig 1. PRISMA flow chart.

https://doi.org/10.1371/journal.pone.0179804.g001

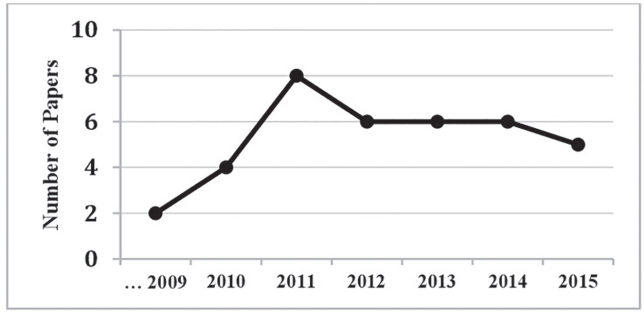


Fig 2. Frequency of published papers per year.

https://doi.org/10.1371/journal.pone.0179804.g002

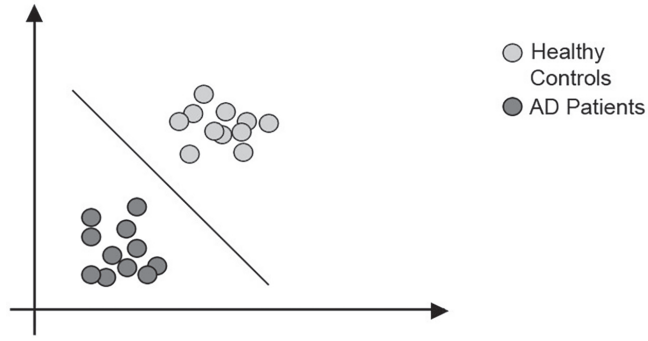


Fig 3. SVM classification example. The data points in the feature space are being classified in 2 classes.

https://doi.org/10.1371/journal.pone.0179804.g003

have being used extensively for almost 30 years; however the ideal ML technique to be used in a certain situation is dependent on the type of data to be used in the model, sample sizes, time constraints and the prediction outcome [11].

Other techniques that appeared less frequently are: Voting Feature Intervals (VFI), K-Nearest Neighbors (KNN), Nearest Shrunken Centroids (NSC) and Bagging (BA). These results will be explored in more detail next, so that firstly we provide a brief explanation of each ML technique, followed by a description on how it was applied for prognosis.

Support Vector Machine (SVM) was originally proposed as an algorithm for classification problems; it is a relatively new technique compared to the other ML approaches. The classification process consists of mapping the data points (usually the study subjects) into a feature space composed of the variables that characterize these data points, except for the outcome variable. Then, the algorithm finds patterns in this feature space by defining the maximum separation between two or more classes, depending on the problem to be solved (see Fig 3) [21]. Contrary to some regression techniques, SVMs are not dependent on a pre-determined model for data fitting, although there are still algorithm specifications to be considered (e.g. choice of a kernel function) [22]; instead, it is a data-driven algorithm that can work relatively well in a scenario where sample sizes are small compared to the number of variables, reason why it has been widely employed by prognostic studies in tasks related to the automated classification of diseases [23].

Regarding the SLR results, SVMs were present in 30/37 selected studies, in 38 proposed models, and being by far the most used machine learning technique in the dementia prognosis research. These numbers account for the traditional SVM and variations (see Table 4). In all of

Table 4. Featuring studies that applied SVMs for the prognosis of dementia.

Variations of the Technique	Featuring Studies
Support Vector Machines	[24–45]
Radial Basis Function SVM	[38, 46–49]
Multi Kernel SVM	[35, 50–52]
Semi-supervised Low Density Separation	[40,48]
Domain Transfer SVM	[28]
Laplacian SVM	[28]
Relevance Vector Machines	[53]
SVM with a Logistic Regression Loss Function	[28]
Other proposed approaches to SVM	[28]

https://doi.org/10.1371/journal.pone.0179804.t004

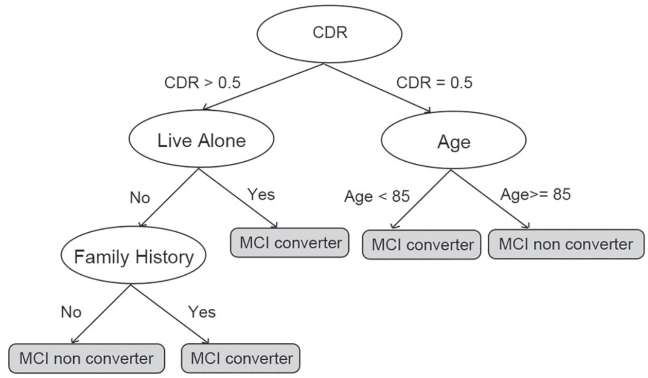


Fig 4. DT classification example. V(1–6) represent values that regulates the splits of the tree.

https://doi.org/10.1371/journal.pone.0179804.g004

the 30 selected studies the SVMs focused at binary classifications where the task was to discriminate mild cognitively impaired (MCI) patients that will or will not develop Alzheimer's Disease (AD). In the general case, the problem is posed as either MCI converters versus MCI non-converters, or progressive MCI versus stable MCI classification. This outlines a situation in which a regression problem ("when will the MCI patients convert to AD?") is formulated as a classification problem ("which MCI patients will convert to AD in X months?") to be solved. Reasons for this could be due to limitations in the data used, i.e. the limited follow-up periods of the subjects included in the studies.

A **Decision Tree** (DT) is a classification algorithm in which the learned knowledge is represented in a tree structure that can be translated to if-then rules. DT's learning process is recursive and starts by testing each input variable as to how well each of them, alone, can classify the labeled examples. The best one is selected as a root node for the tree and its descendant nodes are defined as the possible values (or relevant ratios) of the selected input variable. The training set is then classified between the descendant nodes according to the values of the selected input variable. This process is repeated recursively until no more splits in the tree are possible (see Fig 4) [54]. Like SVMs, DTs do not depend on a pre-defined model and are mostly used to find important interactions between variables. Being intuitive and easy to interpret, DTs have been used in prognostic studies as a tool for determining prognostic subgroups [55,12].

In this SLR, DTs were the second most frequently used ML technique, present in 6/37 selected studies and proposed in 7 models. The variations of this type of technique in the selected studies are shown in Table 5. It was employed for the same reason as SVM; except for one study that investigated the evolution of patients diagnosed with cognitive impairment no dementia (CIND) to AD.

Bayesian Networks (BN) are directed acyclic graphs, in which each node represents a variable and each edge represents a probabilistic dependency. This structure is useful for

Table 5. Featuring studies that applied DTs for the prognosis of dementia.

Variations of the Technique	Featuring Studies
Random Forests	[30, 38, 47, 56]
Decision Trees	[25, 57]
Boosted Trees	[30]

https://doi.org/10.1371/journal.pone.0179804.t005

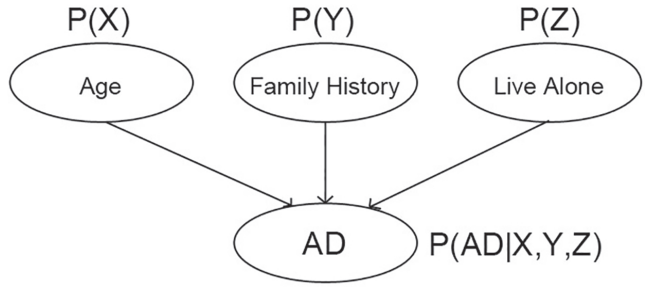


Fig 5. BN example. P(X-Z) represent probabilities and P(X-Z|X,Y,Z) represent conditional probabilities.

https://doi.org/10.1371/journal.pone.0179804.g005

computing the conditional probability of a node, given the values of the other variables or events. In a BN, the learning process is composed of two tasks: learning the structure of the graph and learning the conditional probability distribution for each node (see Fig 5) [58]. In this way, the classification in a BN estimates the posterior probability of a data point to belong to a class, given a set of features. BNs have been applied in the research for classification, knowledge representation and reasoning; however contrary to the other mentioned algorithms, BNs generally produce probabilistic estimations, rather than predictions per se [58]. A great advantage of BN in comparison to other techniques, such as ANNs and SVMs, which renders it benefits for its use in prognostic models, is that they do not require the availability of large amounts of data and can also encode the knowledge of domain experts [59]. However, a drawback in this technique is that it may not be expandable to a large number of features [60].

BNs were the second mostly used ML technique, along with DTs, present in 6/37 selected studies. The identified variations of BN algorithms are shown in Table 6. As previously mentioned, BN models were built for use in classification tasks related to the evolution of patients with MCI to Alzheimer's, with the exception of one study that used BNs for events-based modeling of the progression to AD.

An **Artificial Neural Network** (ANN) is a methodology that performs multifactorial analyses, which is desirable in the health area as medical decision-making problems are usually dependent of many factors. An ANN is composed of nodes connected by weighted edges in a multi-layer architecture that comprises: an input layer, one or more hidden layers and an output layer (see Fig 6). In the training process, inputs and outputs values are known to the network, while the weights are incrementally adjusted so that the outputs of the network are approximate to the known outputs [63]. Despite being a powerful predictor, ANNs are 'black boxes', which means that they are not able to explain their predictions in an intuitive way, contrary to DTs or BNs. Also, they require the specification of the architecture to be used beforehand (i.e. the number of hidden layers) [64].

Table 6. Featuring studies that applied BNs for the prognosis of dementia.

Variations of the Technique	Featuring Studies
Naïve Bayes	[30, 42]
Gaussian Naive Bayes	[27]
Markov Chains Monte Carlo	[61]
Bayesian Outcome prediction with Ensemble Learning	[62]
Gaussian Process Classification	[45]

https://doi.org/10.1371/journal.pone.0179804.t006

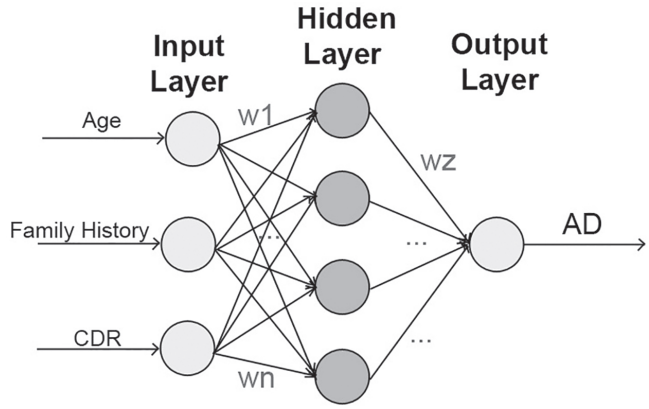


Fig 6. ANN example. The weights of the edges are represented by w(1-n).

https://doi.org/10.1371/journal.pone.0179804.g006

ANNs were present in 3/37 of the selected studies and proposed in 4 models. The identified variations of the standard ANN in the selected studies are shown in Table 7. As was the case with all previous techniques, two studies aimed to predict the development of AD in patients with MCI and one aimed to predict the stage of AD on patients according to their cognitive measures.

K Nearest Neighbors (KNN) is a classification algorithm that takes a data point from an unknown class and assigns it as an input vector in the feature space. Then, the classification process follows by assigning the unknown class data point to the class in which the majority of the K nearest data points belong to (see Fig 7) [66]. The distance between data points is usually measured by Euclidean distance, but it is possible to employ other measures. KNN is one of the simplest ML classification algorithms and have been used in a wide range of applications; however, it can be computationally expensive in a highly dimensional scenario. Further, it considers all features to be equally weighted, which can be a problem if the data has superfluous attributes [60].

The **Nearest Shrunken Centroids (NSC)** classification process starts by calculating the centroids to each of the classes that an unknown data point could belong to. In this context, the centroids are the mean feature vectors of each of the possible classes. Then, the algorithm shrinks the centroids toward the global centroid by a certain threshold [67]. This shrinkage operation acts as a form of feature selection, as it eliminates from the prediction rule, the components of the centroid that are equal to the correspondent component of the overall centroid [68]. Then, the algorithm sets the unknown data point to the class that has the shortest distance to its shrunken centroid (see Fig 7). As in KNN, the distance measure can be Euclidean or other. In the medical field, this algorithm was proposed to deal with the problem of predicting a diagnostic category by DNA microarrays, being useful in a high dimensionality scenario; yet a disadvantage of NSC is the arbitrary choice of shrinkage threshold [67].

Table 7. Featuring studies that applied ANNs for the prognosis of dementia.

Variations of the Technique	Featuring Studies
Artificial Neural Networks	[25,47,65]
Mixed Effects ANN	[65]

https://doi.org/10.1371/journal.pone.0179804.t007

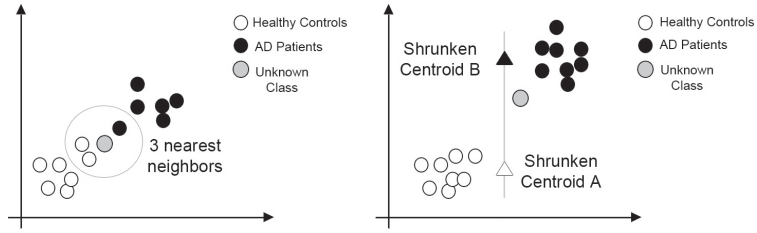


Fig 7. KNN (3-NN) and NSC examples. Both cases classify the unknown data point between 2 classes.

https://doi.org/10.1371/journal.pone.0179804.g007

Voting Feature Intervals (VFI) is an algorithm with a classification process similar to the BN, but instead of assigning probabilities to each possible class, VFI assign votes between feature intervals among the classes. The classification output is the class with the highest sum of votes [69]. One downside of this algorithm is that it is best applicable to contexts where the features are considered independent of each other, which may not always be the case [69]. On the other hand, the VFI algorithm can perform well in scenarios that may have many superfluous features to the classification task, which is also the reason why it was employed in the prognosis study [41].

Bagging (BA) or Bootstrap Aggregation is an ensemble ML technique. This means that it is actually a predictor created from an aggregation of different predictors. It uses bootstrapping to replicate the data set in new data sets that are used to make new predictions. In a classification task, different predictors will assign an unknown data point to a class. Then, it choses class it was classified in the most cases. BA is a method that is useful in the case of instable predictors to reduce the variance and prevent overfitting [70].

The studies that featured KNN, NSC, VFI and BA are shown in the Table 8. In all of these cases, studies aimed to predict the MCI evolution to AD.

Identified microsimulation techniques

This section presents the results for RQ1 that relate to MS techniques.

In a typical MS model, a database of samples from a population exists. Each record in the database represents an individual and their associated states. Thus, in each time-step of the simulation a record at a time is being updated by applying a collection of rules [16]. The updated database at each time step shows the course and trajectory of changes in the population and therefore aggregative indicators can be extracted from this database. MS models contrast with other aggregative simulation models in the way they represent individuals in a population rather than aggregative variables and collective representations. Further, MS differ from agent-based simulations, as the focus of the first is on the trajectory and independent reaction of each individual unit and it is assumed that the units are independent [9, 16].

Table 8. Featuring studies that applied other machine learning techniques for the prognosis of dementia.

Variations of the Technique	Featuring Studies
K Nearest Neighbors	[30,47]
Bagging	[47]
Nearest Shrunken Centroids	[30]
Voting Feature Intervals	[42]

https://doi.org/10.1371/journal.pone.0179804.t008

In the selected studies in our SLR, two papers have used MS techniques: Furiak et al. [71] through a simulation of **Time-to-Event (TTE)** for individuals; and Stallard et al. [72] through a **Grade of Membership (GM)** approach.

Furiak et al. [71] uses TTE to simulate the impact of future hypothetical screening and treatment interventions in delaying AD in a specific population. Stallard et al. [72] applies a GM approach to represent a multidimensional multi-attribute model of AD progress [72]. The term "microsimulation" is not used in this study; however, approaches similar to microsimulation were applied. In this study the impact of a future hypothetical successful intervention in slowing AD progress rate on MEDICARE and MEDICAID programs in the USA has been simulated by aggregating predictions on individual levels.

Any of these studies can be an example of instantiation of the MS approach in population level prognosis. For example in the study by Furiak et al. [71], a baseline population of individuals was created according to the relevant incident data from available studies. Each simulated individual, a record in a table, goes through a risk of developing AD and then after is being exposed to a time-to-event of being diagnosed as AD through screening (in different strategies). The diagnosed simulated individuals delay their progress to AD by receiving a hypothetical treatment. The aggregated values of individual progressions toward AD can be compared to real world situation in order to have an understanding of the performance of different screening strategies in the presence of a possible delaying treatment.

Data characteristics used in the models

Table 9 shows the summary of data regarding RQ2: "What data characteristics (variables, determinants and indicators) are being considered when applying the ML or and MS techniques (physiological, demographic/social, genetics, lifestyle etc)?"

ML methods try to learn the relationship between a set of variables, i.e. variates, and the result variable, i.e. covariate. The studies in our collection used variables as variates while they were focused mostly on a binary prognosis variable (usually indicating development/no development to AD). This binary variable was accompanied with degrees of MMSE (Mini Mental State Examination) and ADAS-cog (Alzheimer's Disease Assessment Scale-cognitive subscale) cognitive scores in the study by Zhang *et al.* [52].

Table 9 also summarizes the connection between variables and studies. Note that variables that were considered in a study but did not contribute to its final result are not shown. Variables were categorized as neuroimaging, cognitive measures (neuropsychological), genetic, lab test, and demographic. Regarding the neuroimaging variables, in all of the included studies, automatic feature extraction techniques were used, either in the form of a software tool (e.g. FreeSurfer for MRI scans) or via their own ML technique being applied to create the models. A reason for this could be the interest in the implementation of more automated methods for identifying the development of AD in MCI patients.

Further, studies that examined more than one variable contributed more than once to the subcategories, while their contribution to categories was counted only once; therefore, the number shown for categories might be smaller than the total sum of numbers for their subcategories. Also, the smaller subcategories were grouped together. Finally, for the two MS studies, demographic and cognitive scores variables were considered as the input variables.

Goals of the prognosis studies on dementia

Table 10 shows the summary data concerning the RQ3: "What are the goals of the studies that employ ML or MS techniques for prognosis of dementia and comorbidities?".

Table 9. Identified data characteristics in the included studies.

Variable Category	Variable Subcategory	Number of Studies	Featuring Studies
Neuroimaging: 28 ML studies	**MRI**	27	[24–26, 28, 31–37, 39–46, 48–53, 61, 62]
	PET	8	[27, 28, 33, 35, 45, 50–52]
Cognitive Measures: 5 ML studies, 1 MS study	**MMSE**	2	[34, 65]
	ADAS-cog	2	[38, 56]
	Other (CDR, FAQ, Buschke Cued Recall)	3	[32, 57, 65]
Genetic: 5 ML studies	**ApoE**	5	[33, 38, 41, 44, 45]
	Family History	1	[41]
Lab Test: 10 ML studies	**CSF**	8	[28, 30, 32, 33, 35, 45, 45, 51]
	Other Lab tests	4	[38, 44, 45, 47]
Demographic: 5 ML studies, 2 MS studies	**Age**	4	[30, 40, 53, 65]
	Other demographic	3	[38, 71, 72]

Abbreviations: MRI: Magnetic Resonance Imaging; **PET**: Positron Emission Tomography; **MMSE**: Mini Mental State Examination; **ADAS-cog**: Alzheimer's Disease Assessment Scale-cognitive subscale; **CDR**: Clinical Dementia Rating; **FAQ**: Functional Activities Questionnaire; **CSF**: Cerebrospinal Fluid

https://doi.org/10.1371/journal.pone.0179804.t009

The aggregated data shows that the majority of studies, 32/37, intended to investigate the progression of MCI to AD, posing the problem as either MCI converters versus MCI non-converters, or progressive MCI versus stable MCI. These studies identified the need for the AD prognosis and proposed models for its prediction, usually within 6 to 36 months until development. Exceptions are Adaszewski *et al.* [24], and Chen *et al.* [62], which constructed models for 48 and 60 months prediction, respectively. Most of these studies used data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) in their models, what could explain the limitation of the prediction time as the follow-up period is limited.

A variation of this goal was the prediction of AD, but from Cognitive Impairment No Dementia, instead of MCI. This study [57] was conducted using data from the Canadian Study of Health and Aging (CSHA).

The goal of Tandon *et al.* [65] was the use of patients' longitudinal data of the Layton Aging and Alzheimer's Research Center (LAARC) to model the time-course of AD in terms of their cognitive functions. The chosen variable was the MMSE score (Mini Mental State Examination). Likewise, Fonteijn *et al.* [61] also had disease progression modeling as the goal, but in

Table 10. Goals of the studies in respect to the prognosis of dementia.

Study Goals	Count	Conditions Studied	Type of Data Analysis	Featuring Studies
Predict the development of Alzheimer's Disease from Mild Cognitive Impairment	32	Alzheimer's Disease, Mild Cognitive Impairment	Machine Learning	[24–53, 56, 62]
Predict the development of Alzheimer's Disease from Cognitive Impairment No Dementia	1	Alzheimer's Disease, Cognitive Impairment No Dementia	Machine Learning	[57]
Model disease stage through Mini Mental State Examination score	1	Alzheimer's Disease	Machine Learning	[65]
Events-based disease progression modeling	1	Alzheimer's Disease, Huntington's Disease	Machine Learning	[61]
Estimate the clinical course of mild Alzheimer's Disease to Alzheimer's Disease to death, and estimate costs (MEDICARE and MEDICAID)	1	Alzheimer's Disease, Mild Cognitive Impairment	Microsimulation	[72]
Evaluate screening and treatment to delay Alzheimer's Disease	1	Alzheimer's Disease	Microsimulation	[71]

https://doi.org/10.1371/journal.pone.0179804.t010

this study the model is characterized as an events-based model. Two types of events were considered: transitions to later clinical status (i.e. presymptomatic AD to MCI) and atrophy events measured by MRI scans. This is also the only study that approached a dementia disorder other than AD, building models for both AD, and Huntington's Disease.

Stallard *et al.* [72] explored disease progression through MS aiming the estimation and comparison of costs (MEDICARE and MEDICAID) for slowing AD advancement in both patients with mild AD and moderate AD. Finally, Furiak *et al.* [71], applied MS to provide a framework for the screening of AD, which investigated treatment interventions for delaying the AD progression in a population.

Handling of censored data

Data censoring happens when the information about the individual time to the event of interest is unknown for some participants [73]. This can occur when information on the time to event is unavailable due to loss to follow-up or due to the non-occurrence of the outcome event before the trial end [73][74]. In this SLR, only two out of the 37 selected studies addressed censored data, explicitly. The Craig-Schapiro et al. study [30] used Cox proportional hazard models (CPHM) to assess which baseline biomarkers should be considered in the ML multivariate models targeting at their ability to predict the conversion from cognitive normalcy (CDR 0) to very mild or mild dementia (CDR 0.5 and 1). They stated that participants who did not develop very mild or mild dementia during the follow-up were statistically censored.

In the Plant et al. study [42] data censoring is addressed as a threat to validity, arguing that for shorter follow-ups (30 months in this study), there may be patients classified as MCI with a MRI pathological pattern who had not yet developed AD during the follow-up.

The remaining studies included herein did not make any explicit statements about data censoring. Note that the studies by Escudero et al. [33], Moradi et al. [40] and Gaser et al. [53] performed a survival analysis to estimate a hazard ratio for the MCI conversion to AD using CPHM, which is a technique that is able to deal with censored data; however, no specific remark about data censoring was made.

Focus of the studies

The last research question of this SLR (RQ5) was: "Do the studies focus on individuals or populations?".

Out of the 37 included studies, only the two studies that used the MS methods [71,72] focused on populations and the rest of papers, all using ML, focused on the prognosis of dementia in an individual level.

Discussion

Discussion of the current evidence

The main findings from this SLR are summarized in the following points: (i) most of the research is focused on the use of neuroimaging for predicting the development of AD from MCI, using the ADNI database; (ii) estimations are usually made up to 36 months before the development to AD; (iii) lifestyle and socioeconomic variables were absent in the assessed models; (iv) data censoring is not addressed in the vast majority of the studies included in our SLR; (v) the focus of the research is mostly on individual level.

There is an indication that North America is leading the research on treatments for the preclinical stage of AD whilst Europe leads the lifestyle interventions for the prevention of

dementia [2]. In what concerns studies that make use of high-level computational techniques, such as ML and MS, the findings of this SLR are consistent with the first part of this sentence.

Most of the research dedicated to the prognosis of dementia and that make use of ML techniques is focused on using neuroimaging to predict the development of AD from MCI, particularly making use of the ADNI database. A consequence of this scenario is the almost exclusive focus of the recent research in validating biomarkers to be used in treatment trials, since this is the overall objective of ADNI [75]. This intensive research on biomarkers is important to make the pharmaceutical research faster and to reduce the exposure of ineffective experimental drugs [76].

Another important aspect regarding the overall preference for the neuroimaging variables, throughout the included studies in this SLR, is that most of the prognosis research is concerned with a single aspect of dementia, and prognostic estimates should consider a multivariable approach. The reason for that is the variability among patients that could make the single predictor variable not very effective [77]. Furthermore, ADNI does not include in its database subjects with important comorbidities, considering dementia and the elderly population, like cancer and heart failure [75]. Additionally, as ADNI is not an epidemiologic study, there is a risk that the utility of methods used in the studies would be tailored to the ADNI's specific conditions.

Still on the topic of the development of AD in persons with MCI, another important aspect to be discussed is how much time beforehand the proposed models in the current research can predict this conversion. The majority of them are set out to do this task in a period of up to 36 months. Putting aside the accuracy aspect of these predictions (which of course is of great importance), would this time constraint be enough to employ preventive strategies (pharmaceutical or non-pharmaceutical) on screened patients so to delay their progression to AD? An important consideration in the research for treatments is that they prolong the time patients spend in the most amenable stages of dementia and shorten the time they stay in the most severe stages, in which they suffer from a very low quality of life and when care is most costly [2].

The absence of studies using lifestyle and socioeconomic variables in models could point out a gap in the current research for more holistic approaches to the prognosis of dementia. Another possibility is that studies that are investigating these may simply apply other data analysis methods in deriving their predictions that are not ML or MS.

The fact that the handling of censored data was not made clear in all, except for two studies, can raise some concerns, as in most of the studies' demographics the participants were divided into classes of normal controls, AD patients, MCI converters and MCI non-converters. The class of MCI non-converters could entail the case when a participant did not experience the event in evidence during the follow-up, characterizing a right-censoring scenario.

Lastly, with regard to the focus of the primary studies, the lack of studies that utilize MS for prognosis of dementia in individuals can be due to this method being usually applied to populations (rather than individuals), and also due to this method be based on simulating masses of individuals. However, in relation to ML methods, the lack of studies focusing on predicting the epidemiology of dementia in populations can be interpreted as a study gap.

Methodological issues

When discussing the techniques being used to derive prognostic estimates for dementia, one interesting aspect to note is the comparison between them, in relation to which one(s) performed best. This task proved challenging. Reasons for that are due to several limitations in interpreting such results, detailed next.

Firstly, the studies have used different validation procedures and this can make their comparison difficult. Even in the studies that used the same method for accuracy calculation, the difference in some parameters (e.g. the number of folds in cross validation) or how they calculated distance of a prediction to the test case could have an impact on the reported accuracy.

Further on, the reports on accuracy are based on different datasets, and for those who share the same dataset (such as ADNI) each might use a different number of records, which can impact the reported accuracy. Also, the majority of studies have considered MRI or PET images, where the quality of images or the image pre-processing applied before the ML method can impact their reported accuracy.

Regardless of the applied method, different variables have different predictive powers and when two papers that have used the same method report different variables, they should be compared considering these reported variables. Accuracy and other related indicators have their values compared to a golden standard that determines the existence of AD (or any other progression). In some studies they have chosen other indicators rather than the golden standard (i.e. cerebral biopsy/autopsy versus CDR/GDS/ADL/ADAS-cog/MMSE numbers). Lastly, the follow-up period is different in the studies and maybe longer follow-ups would result in higher sensitivity reports.

Further, there are three commonly used types of accuracy that can be attributed to prognosis models: discrimination, calibration, and reclassification [78]. The discriminatory accuracy is the ability of the prognosis model in separating individuals regarding the outcome, while the calibration accuracy is how much the prognosis model risk prediction complies with observed risks in a population. In reclassification, one is interested in measuring the added predictive ability by a new biomarker [79,80]. With regard to the primary studies in this SLR, they only reported the discriminatory accuracy of prognosis. The overall accuracy was the index most commonly reported (see S1 Table). Alongside that, in 18 studies the AUC was indicated. The predications mostly concern a binary discrimination between converting and non-converting MCI, while in one of the studies (Zhang et al. [52]) predication of MMSE and ADAS-cog scores was considered. It is noteworthy that the presentation of prognosis accuracy in this SLR's primary studies contrasts with similar prediction studies in the cancer field [81].

Limitations

Regarding the limitations of this SLR it can be addressed the issue of whether a suitable large representative sample of all the possible relevant primary studies were included in the final set, and also the non-medical background of the two researchers on the study team, who screened most of the papers (A1 and A2). To mitigate the first issue a more inclusive selection strategy has been taken. This means that in papers in which there were poor indications of the inclusive criteria in their title or abstract, the content of study was further investigated for a possible inclusion. For the second issue, if it was not clear the fit of the paper for prognosis, a member of research team with medical education background (A4) was consulted.

Future perspectives

The results of this SLR presented research trends and gaps that should be addressed in future research on the prognosis of dementia. Based on these findings, further research should explore different combinations of ML and MS techniques, using a multivariable approach that includes the identified data characteristics as well as lifestyle and social factors.

Conclusion

Through the SLR, 37 studies that focus on the prognosis of dementia by using ML or MS techniques were selected. These studies were summarized in terms of different aspects including types of techniques, variables or goals and focus of the studies.

Our findings pointed out that most of the studies were concerned about predicting the development of AD in individuals with MCI using one of ML techniques. Neuroimaging data was the most common data to be fed into ML techniques. Only two studies focused on prediction regarding populations, and those were the only two studies that applied MS techniques. We identified only a limited number of datasets are being used in the studies (most notably, the ADNI database).

Supporting information

S1 Table. Final set of included studies.
(PDF)

S2 Table. PRISMA checklist.
(PDF)

Author Contributions

Conceptualization: Ana Luiza Dallora, Shahryar Eivazzadeh, Johan Berglund, Peter Anderberg.

Data curation: Ana Luiza Dallora, Shahryar Eivazzadeh.

Formal analysis: Ana Luiza Dallora, Shahryar Eivazzadeh.

Investigation: Ana Luiza Dallora, Shahryar Eivazzadeh.

Methodology: Ana Luiza Dallora, Shahryar Eivazzadeh, Emilia Mendes, Johan Berglund, Peter Anderberg.

Project administration: Emilia Mendes, Johan Berglund.

Supervision: Emilia Mendes, Johan Berglund, Peter Anderberg.

Validation: Ana Luiza Dallora, Shahryar Eivazzadeh, Emilia Mendes, Johan Berglund, Peter Anderberg.

Visualization: Ana Luiza Dallora, Shahryar Eivazzadeh.

Writing – original draft: Ana Luiza Dallora, Shahryar Eivazzadeh.

Writing – review & editing: Emilia Mendes, Johan Berglund, Peter Anderberg.

References

1. Melis RJF, Marengoni A, Rizzuto D, Teerenstra S, Kivipelto M, Angleman SB, et al. The influence of multimorbidity on clinical progression of dementia in a population-based cohort. PloS One. 2013; 8(12): e84014. https://doi.org/10.1371/journal.pone.0084014 PMID: 24386324
2. Winblad B, Amouyel P, Andrieu S, Ballard C, Brayne C, Brodaty H, et al. Defeating Alzheimer's disease and other dementias: a priority for European science and society. Lancet Neurol. 2016 Apr; 15(5):455–532. https://doi.org/10.1016/S1474-4422(16)00062-4 PMID: 26987701
3. van de Vorst IE, Vaartjes I, Geerlings MI, Bots ML, Koek HL. Prognosis of patients with dementia: results from a prospective nationwide registry linkage study in the Netherlands. BMJ Open. 2015 Oct 1; 5(10):e008897. https://doi.org/10.1136/bmjopen-2015-008897 PMID: 26510729

4. Poblador-Plou B, Calderón-Larrañaga A, Marta-Moreno J, Hancco-Saavedra J, Sicras-Mainar A, Soljak M, et al. Comorbidity of dementia: a cross-sectional study of primary care older patients. BMC Psychiatry. 2014; 14:84. https://doi.org/10.1186/1471-244X-14-84 PMID: 24645776
5. Jennings LA, Reuben DB, Evertson LC, Serrano KS, Ercoli L, Grill J, et al. Unmet needs of caregivers of individuals referred to a dementia care program. J Am Geriatr Soc. 2015 Feb; 63(2):282–9. https://doi.org/10.1111/jgs.13251 PMID: 25688604
6. WHO | Dementia: a public health priority [Internet]. WHO. [cited 2016 Aug 9]. http://www.who.int/mental_health/publications/dementia_report_2012/en/
7. Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform. 2001 Dec; 34(6):428–39. https://doi.org/10.1006/jbin.2002.1038 PMID: 12198763
8. Larrañaga P, Calvo B, Santana R, Bielza C, Galdiano J, Inza I, et al. Machine learning in bioinformatics. Brief Bioinform. 2006 Mar; 7(1):86–112. PMID: 16761367
9. Louridas P, Ebert C. Machine Learning. IEEE Softw. 2016 Sep; 33(5):110–5.
10. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015; 13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005 PMID: 25750696
11. Boman M, Holm E. Multi-agent systems, time geography, and microsimulations. In: Systems approaches and their application [Internet]. Springer; 2005 [cited 2015 Jul 9]. p. 95–118. http://link.springer.com/chapter/10.1007/1-4020-2370-7_4
12. Cruz JA, Wishart DS. Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Inform. 2007 Feb 11; 2:59–77. PMID: 19458758
13. Suh G-H, Shah A. A review of the epidemiological transition in dementia—cross-national comparisons of the indices related to Alzheimer's disease and vascular dementia. Acta Psychiatrica Scandinavica. 2001 Jul 1; 104(1):4–11. PMID: 11437743
14. Jagger C, Andersen K, Breteler MM, Copeland JR, Helmer C, Baldereschi M, et al. Prognosis with dementia in Europe: A collaborative study of population-based cohorts. Neurologic Diseases in the Elderly Research Group. Neurology. 2000; 54(11 Suppl 5):S16–20. PMID: 10854356
15. Gilbert GN. Agent-Based Models. SAGE; 2008. 113 p.
16. Rutter CM, Zaslavsky AM, Feuer EJ. Dynamic microsimulation models for health outcomes: a review. Med Decis Mak Int J Soc Med Decis Mak. 2011 Feb; 31(1):10–8.
17. Kitchenham B, Charters S. Guidelines for performing Systematic Literature Reviews in Software Engineering. 2007.
18. Dallora AL, Eivazzadeh S, Mendes E, Berglund J, Anderberg P. Prognosis of Dementia Employing Machine Learning and Microsimulation Techniques: A Systematic Literature Review. Procedia Computer Science. 2016; 100:480–8.
19. Pai M, McCulloch M, Gorman JD, Pai N, Enanoria W, Kennedy G, et al. Systematic reviews and meta-analyses: an illustrated, step-by-step guide. Natl Med J India. 2004 Apr; 17(2):86–95. PMID: 15141602
20. Marengoni A, Angleman S, Melis R, Mangialasche F, Karp A, Garmen A, et al. Aging with multimorbidity: a systematic review of the literature. Ageing Res Rev. 2011 Sep; 10(4):430–9. https://doi.org/10.1016/j.arr.2011.03.003 PMID: 21402176
21. Cortes C, Vapnik V. Support-Vector Networks. Mach Learn. 1995 Sep; 20(3):273–97.
22. Burges CJC. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. 1998 Jun 1; 2(2):121–67.
23. Yu W, Liu T, Valdez R, Gwinn M, Khoury MJ. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Med Inform Decis Mak. 2010; 10:16. https://doi.org/10.1186/1472-6947-10-16 PMID: 20307319
24. Adaszewski S, Dukart J, Kherif F, Frackowiak R, Draganski B, Alzheimer's Disease Neuroimaging Initiative. How early can we predict Alzheimer's disease using computational anatomy? Neurobiol Aging. 2013 Dec; 34(12):2815–26. https://doi.org/10.1016/j.neurobiolaging.2013.06.015 PMID: 23890839
25. Aguilar C, Westman E, Muehlboeck J-S, Mecocci P, Vellas B, Tsolaki M, et al. Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment. Psychiatry Res-Neuroimaging. 2013 May 30; 212(2):89–98.
26. Aksu Y, Miller DJ, Kesidis G, Bigler DC, Yang QX. An MRI-derived definition of MCI-to-AD conversion for long-term, automatic prognosis of MCI patients. PloS One. 2011; 6(10):e25074. https://doi.org/10.1371/journal.pone.0025074 PMID: 22022375
27. Cabral C, Morgado PM, Campos Costa D, Silveira M, Alzheimer's Disease Neuroimaging Initiative. Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal stages. Comput Biol Med. 2015 Mar; 58:101–9. https://doi.org/10.1016/j.compbiomed.2015.01.003 PMID: 25625698

28. Cheng B, Liu M, Zhang D, Munsell BC, Shen D. Domain Transfer Learning for MCI Conversion Prediction. IEEE Trans Biomed Eng. 2015; 62(7):1805–17. https://doi.org/10.1109/TBME.2015.2404809 PMID: 25751861

29. Costafreda SG, Dinov ID, Tu Z, Shi Y, Liu C-Y, Kloszewska I, et al. Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment. NeuroImage. 2011 May 1; 56(1):212–9. https://doi.org/10.1016/j.neuroimage.2011.01.050 PMID: 21272654

30. Craig-Schapiro R, Kuhn M, Xiong C, Pickering EH, Liu J, Misko TP, et al. Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and prognosis. PloS One. 2011; 6(4):e18850. https://doi.org/10.1371/journal.pone.0018850 PMID: 21526197

31. Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert M-O, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage. 2011 May 15; 56(2):766–81. https://doi.org/10.1016/j.neuroimage.2010.06.013 PMID: 20542124

32. Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T, et al. Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. PloS One. 2011; 6(7):e21896. https://doi.org/10.1371/journal.pone.0021896 PMID: 21814561

33. Escudero J, Ifeachor E, Zajicek JP, Alzheimer's Disease Neuroimaging Initiative. Bioprofile analysis: a new approach for the analysis of biomedical data in Alzheimer's disease. J Alzheimers Dis JAD. 2012; 32(4):997–1010. https://doi.org/10.3233/JAD-2012-121024 PMID: 22886027

34. Guerrero R, Wolz R, Rao AW, Rueckert D, Alzheimer's Disease Neuroimaging Initiative (ADNI). Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO. NeuroImage. 2014 Jul 1; 94:275–86. https://doi.org/10.1016/j.neuroimage.2014.03.036 PMID: 24657351

35. Hinrichs C, Singh V, Xu G, Johnson SC, Alzheimers Disease Neuroimaging Initiative. Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population. NeuroImage. 2011 Mar 15; 55(2):574–89. https://doi.org/10.1016/j.neuroimage.2010.10.081 PMID: 21146621

36. Kloeppel S, Peter J, Ludl A, Pilatus A, Maier S, Mader I, et al. Applying Automated MR-Based Diagnostic Methods to the Memory Clinic: A Prospective Study. J Alzheimers Dis. 2015; 47(4):939–54. https://doi.org/10.3233/JAD-150334 PMID: 26401773

37. Komlagan M, Ta V-T, Pan X, Domenger J-P, Collins DL, Coupe P. Anatomically Constrained Weak Classifier Fusion for Early Detection of Alzheimer's Disease. In: Wu G, Zhang D, Zhou L, editors. Machine Learning in Medical Imaging (mlmi 2014). 2014. p. 141–8.

38. Li H, Liu Y, Gong P, Zhang C, Ye J, Alzheimers Disease Neuroimaging Initiative. Hierarchical interactions model for predicting Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD) conversion. PloS One. 2014; 9(1):e82450. https://doi.org/10.1371/journal.pone.0082450 PMID: 24416143

39. Li Y, Wang Y, Wu G, Shi F, Zhou L, Lin W, et al. Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features. Neurobiol Aging. 2012 Feb; 33(2):427.e15–30.

40. Moradi E, Pepe A, Gaser C, Huttunen H, Tohka J, Alzheimer's Disease Neuroimaging Initiative. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. NeuroImage. 2015 Jan 1; 104:398–412. https://doi.org/10.1016/j.neuroimage.2014.10.002 PMID: 25312773

41. Nho K, Shen L, Kim S, Risacher SL, West JD, Foroud T, et al. Automatic Prediction of Conversion from Mild Cognitive Impairment to Probable Alzheimer's Disease using Structural Magnetic Resonance Imaging. AMIA Annu Symp Proc AMIA Symp AMIA Symp. 2010; 2010:542–6.

42. Plant C, Teipel SJ, Oswald A, Böhm C, Meindl T, Mourao-Miranda J, et al. Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease. NeuroImage. 2010 Mar; 50(1):162–74. https://doi.org/10.1016/j.neuroimage.2009.11.046 PMID: 19961938

43. Salvatore C, Cerasa A, Battista P, Gilardi MC, Quattrone A, Castiglioni I, et al. Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach. Front Neurosci. 2015; 9:307. https://doi.org/10.3389/fnins.2015.00307 PMID: 26388719

44. Ye J, Farnum M, Yang E, Verbeeck R, Lobanov V, Raghavan N, et al. Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data. BMC Neurol. 2012; 12.

45. Young J, Modat M, Cardoso MJ, Mendelson A, Cash D, Ourselin S, et al. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage Clin. 2013; 2:735–45. https://doi.org/10.1016/j.nicl.2013.05.004 PMID: 24179825

46. Ferrarini L, Frisoni GB, Pievani M, Reiber JHC, Ganzola R, Milles J. Morphological hippocampal markers for automated detection of alzheimer's disease and mild cognitive impairment converters in magnetic resonance images. J Alzheimers Dis. 2009; 17(3):643–59. https://doi.org/10.3233/JAD-2009-1082 PMID: 19433888

47. Llano DA, Devanarayan V, Simon AJ. Evaluation of Plasma Proteomic Data for Alzheimer Disease State Classification and for the Prediction of Progression From Mild Cognitive Impairment to Alzheimer Disease. Alzheimer Dis Assoc Disord. 2013 Sep; 27(3):233–43. https://doi.org/10.1097/WAD.0b013e31826d597a PMID: 23023094

48. Moradi E, Tohka J, Gaser C. Semi-supervised learning in MCI-to-ad conversion prediction—When is unlabeled data useful? DeepDyve [Internet]. 2014 Jun 4 [cited 2016 Feb 10]; https://www.deepdyve.com/lp/institute-of-electrical-and-electronics-engineers/semi-supervised-learning-in-mci-to-ad-conversion-prediction-when-is-8CEJUEmnTf

49. Ota K, Oishi N, Ito K, Fukuyama H. A comparison of three brain atlases for MCI prediction. J Neurosci Methods. 2014; 221:139–50. https://doi.org/10.1016/j.jneumeth.2013.10.003 PMID: 24140118

50. Liu F, Wee C-Y, Chen H, Shen D. Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification. NeuroImage. 2014; 84:466–75. https://doi.org/10.1016/j.neuroimage.2013.09.015 PMID: 24045077

51. Suk H-I, Shen D. Deep learning-based feature representation for AD/MCI classification. Med Image Comput Comput-Assist Interv MICCAI Int Conf Med Image Comput Comput-Assist Interv. 2013; 16(Pt 2):583–90.

52. Zhang D, Shen D, Alzheimer's Disease Neuroimaging Initiative. Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. PloS One. 2012; 7(3):e33182. https://doi.org/10.1371/journal.pone.0033182 PMID: 22457741

53. Gaser C, Franke K, Klöppel S, Koutsouleris N, Sauer H, Alzheimer's Disease Neuroimaging Initiative. BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease. PloS One. 2013; 8(6):e67346. https://doi.org/10.1371/journal.pone.0067346 PMID: 23826273

54. Mitchell T. Decision Tree Learning. In: Machine Learning. McGraw-Hill Science/Engineering/Math; 1997. p. 432. 50

55. N, Kudus A. Decision Tree for Prognostic Classification of Multivariate Survival Data and Competing Risks. In: A M, editor. Recent Advances in Technologies [Internet]. InTech; 2009 [cited 2016 Aug 10]. http://www.intechopen.com/books/recent-advances-in-technologies/decision-tree-for-prognostic-classification-of-multivariate-survival-data-and-competing-risks 51

56. Llano DA, Laforet G, Devanarayan V. Derivation of a new ADAS-cog composite using tree-based multivariate analysis: Prediction of conversion from mild cognitive impairment to alzheimer disease. Alzheimer Dis Assoc Disord. 2011; 25(1):73–84. 53 https://doi.org/10.1097/WAD.0b013e3181f5b8d8 PMID: 20847637

57. Ritchie LJ, Tuokko H. Clinical Decision Trees for Predicting Conversion from Cognitive Impairment No Dementia (CIND) to Dementia in a Longitudinal Population-Based Study. Arch Clin Neuropsychol. 2011 Feb; 26(1):16–25. 54 https://doi.org/10.1093/arclin/acq089 PMID: 21147863

58. Cheng J, Greiner R. Comparing Bayesian Network Classifiers. In: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence [Internet]. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 1999 [cited 2016 Aug 10]. p. 101–8. (UAI'99). http://dl.acm.org/citation.cfm?id=2073796.2073808

59. van Gerven MAJ, Taal BG, Lucas PJF. Dynamic Bayesian networks as prognostic models for clinical patient management. J Biomed Inform. 2008 Aug; 41(4):515–29. https://doi.org/10.1016/j.jbi.2008.01.006 PMID: 18337188

60. Phyu TN. Survey of classification techniques in data mining. In: Proceedings of the International MultiConference of Engineers and Computer Scientists [Internet]. 2009 [cited 2017 Apr 24]. p. 18–20.

61. Fonteijn HM, Modat M, Clarkson MJ, Barnes J, Lehmann M, Hobbs NZ, et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage. 2012 Apr 15; 60(3):1880–9. https://doi.org/10.1016/j.neuroimage.2012.01.062 PMID: 22281676

62. Chen R, Young K, Chao LL, Miller B, Yaffe K, Weiner MW, et al. Prediction of conversion from mild cognitive impairment to Alzheimer disease based on bayesian data mining with ensemble learning. Neuroradiol J. 2012 Mar; 25(1):5–16. https://doi.org/10.1177/197140091202500101 PMID: 24028870

63. Dayhoff JE, DeLeo JM. Artificial neural networks: opening the black box. Cancer. 2001 Apr 15; 91(8 Suppl):1615–35. PMID: 11309760

64. Ripley BD, Ripley RM. Neural networks as statistical methods in survival analysis. In: Clinical Applications of Artificial Neural Networks. Cambridge University Press; 2007.

65. Tandon R, Adak S, Kaye JA. Neural networks for longitudinal studies in Alzheimer's disease. Artif Intell Med. 2006 Mar; 36(3):245–55. https://doi.org/10.1016/j.artmed.2005.10.007 PMID: 16427257

66. Keller JM, Gray MR, Givens JA. A fuzzy K-nearest neighbor algorithm. IEEE Trans Syst Man Cybern. 1985 Jul; SMC-15(4):580–5.

67. Tibshirani R, Hastie T, Narasimhan B, Chu G. Class Prediction by Nearest Shrunken Centroids, with Applications to DNA Microarrays. Stat Sci. 2003; 18(1):104–17.

68. Struyf J, Dobrin S, Page D. Combining gene expression, demographic and clinical data in modeling disease: a case study of bipolar disorder and schizophrenia. BMC Genomics. 2008; 9:531. https://doi.org/10.1186/1471-2164-9-531 PMID: 18992130

69. Demiröz G, Güvenir HA. Classification by Voting Feature Intervals. In: van Someren M, Widmer G, editors. Machine Learning: ECML-97 [Internet]. Springer Berlin Heidelberg; 1997 [cited 2016 Aug 10]. p. 85–92. (Lecture Notes in Computer Science). http://link.springer.com/chapter/10.1007/3-540-62858-4_74

70. Breiman L. Bagging Predictors. Mach Learn. 24(2):123–40.

71. Furiak NM, Klein RW, Kahle-Wrobleski K, Siemers ER, Sarpong E, Klein TM. Modeling screening, prevention, and delaying of Alzheimer's disease: An early-stage decision analytic model. BMC Med Inform Decis Mak. 2010; 10(1).

72. Stallard E, Kinosian B, Zbrozek AS, Yashin AI, Glick HA, Stern Y. Estimation and validation of a multiattribute model of alzheimer disease progression. Med Decis Making. 2010; 30(6):625–38. https://doi.org/10.1177/0272989X10363479 PMID: 21183754

73. Prinja S, Gupta N, Verma R. Censoring in Clinical Trials: Review of Survival Analysis Techniques. Indian J Community Med. 2010 Apr; 35(2):217–21. https://doi.org/10.4103/0970-0218.66859 PMID: 20922095

74. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. Br J Cancer. 2003 Jul 21; 89(2):232–8. https://doi.org/10.1038/sj.bjc.6601118 PMID: 12865907

75. Weiner MW, Aisen PS, Jack CR, Jagust WJ, Trojanowski JQ, Shaw L, et al. The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimers Dement J Alzheimers Assoc. 2010 May; 6(3):202–11.e7.

76. Strimbu K, Tavel JA. What are Biomarkers? Curr Opin HIV AIDS. 2010 Nov; 5(6):463–6. https://doi.org/10.1097/COH.0b013e32833ed177 PMID: 20978388

77. Moons KGM, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: what, why, and how? BMJ. 2009 Feb 23; 338:b375. https://doi.org/10.1136/bmj.b375 PMID: 19237405

78. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Statistical methods for the assessment of prognostic biomarkers (Part I): discrimination. Nephrol Dial Transplant. 2010 May; 25(5):1399–401. https://doi.org/10.1093/ndt/gfq018 PMID: 20139066

79. Tripepi G, Jager KJ, Dekker FW, Zoccali C. Statistical methods for the assessment of prognostic biomarkers(part II): calibration and re-classification. Nephrol Dial Transplant. 2010 May; 25(5):1402–5. https://doi.org/10.1093/ndt/gfq046 PMID: 20167948

80. Pencina MJ, D'Agostino RB, D'Agostino RB, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008 Jan 30; 27(2):157–172–212. https://doi.org/10.1002/sim.2929 PMID: 17569110

81. Mallett S, Royston P, Waters R, Dutton S, Altman DG. Reporting performance of prognostic models in cancer: a review. BMC Medicine. 2010; 8:21. https://doi.org/10.1186/1741-7015-8-21 PMID: 20353579

S1 Table. Final set of included studies.

Title	Study	Conditions Studied	Database Name	Data Analysis Technique	Best Accuracy Achieved
Automated hippocampal shape analysis predicts the onset of dementia in mild cognitive impairment	Costafreda, S. et al. [29]	AD, MCI	AddNeuroMed	ML	85.0%
Different multivariate techniques for automated classification of MRI data in Alzheimer's disease and mild cognitive impairment	Aguilar, C. et al. [25]	AD, MCI	AddNeuroMed	ML	84.9%
An MRI-derived definition of MCI-to-AD conversion for long-term, automatic prognosis of MCI patients	Aksu, Y. et al. [26]	AD, MCI	ADNI	ML	83.00%
Discriminant analysis of longitudinal cortical thickness changes in Alzheimer's disease using dynamic and network features	Li, Y. et al. [39]	AD, MCI	ADNI	ML	81.7%
BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease	Gaser, C. et al. [53]	AD, MCI	ADNI	ML	81.0%
Domain Transfer Learning for MCI Conversion Prediction	Cheng, B. et al. [28]	AD, MCI	ADNI	ML	79.4%
Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers	Zhang, D. et al. [52]	AD, MCI	ADNI	ML	78.4%
Deep learning-based feature representation for AD/MCI classification	Suk, Heung-Il; Shen, Dinggang [51]	AD, MCI	ADNI	ML	75.80%
Anatomically constrained weak classifier fusion for early detection of Alzheimer's disease	Komlagan, M. et al. [37]	AD, MCI	ADNI	ML	75.6%
Hierarchical interactions model for predicting Mild Cognitive Impairment (MCI) to Alzheimer's Disease (AD) conversion	Li, H. et al. [38]	AD, MCI	ADNI	ML	74.76%
Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects	Moradi, E. et al. [40]	AD, MCI	ADNI	ML	74.74%
Automatic Prediction of Conversion from Mild Cognitive Impairment to Probable Alzheimer's Disease using Structural Magnetic Resonance Imaging	Nho, K. et al. [41]	AD, MCI	ADNI	ML	72.3%

Manifold population modeling as a neuro-imaging biomarker: application to ADNI and ADNI-GO	Guerrero, R. et al. [34]	AD, MCI	ADNI	ML	71.0%
Bioprofile analysis: A new approach for the analysis of biomedical data in Alzheimer's disease	Escudero, J. et al. [33]	AD, MCI	ADNI	ML	68.2%
Inter-modality relationship constrained multi-modality multi-task feature selection for Alzheimer's Disease and mild cognitive impairment identification	Liu, F. et al. [50]	AD, MCI	ADNI	ML	67.83%
Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors	Cui, Y. et al. [32]	AD, MCI	ADNI	ML	67.13%
Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment	Young, J. et al. [45]	AD, MCI	ADNI	ML	66.7%
Magnetic resonance imaging biomarkers for the early diagnosis of Alzheimer's disease: a machine learning approach	Salvatore, C. et al. [43]	AD, MCI	ADNI	ML	66.0%
How early can we predict Alzheimer's disease using computational anatomy?	Adaszewski, S. et al. [24]	AD, MCI	ADNI	ML	63.7%
Evaluation of Plasma Proteomic Data for Alzheimer Disease State Classification and for the Prediction of Progression From Mild Cognitive Impairment to Alzheimer Disease	Llano, D. et al. [47]	AD, MCI	ADNI	ML	62.0%
Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database	Cuingnet, R. et al. [31]	AD, MCI	ADNI	ML	-
Semi-supervised learning in MCI-to-ad conversion prediction - " When is unlabeled data useful?"	Moradi, E. et al. [48]	AD, MCI	ADNI	ML	-
Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data	Ye, J. et al. [44]	AD, MCI	ADNI	ML	-
Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population	Hinrichs, C. et al. [35]	AD, MCI	ADNI	ML	-
Predicting conversion from MCI to AD with FDG-PET brain images at different prodromal	Cabral, C. et al. stages [27]	AD, MCI	ADNI	ML	-

Derivation of a new ADAS-cog composite using tree-based multivariate analysis: Prediction of conversion from mild cognitive impairment to alzheimer disease	Llano, D. et al. [56]	AD, MCI	ADNI	ML	-
Applying Automated MR-Based Diagnostic Methods to the Memory Clinic: A Prospective Study	Kloeppel, S. et al. [36]	AD, MCI	ADNI, Freiburg	ML	65.0%
Prediction of conversion from mild cognitive impairment to Alzheimer disease based on bayesian data mining with ensemble learning	Chen, R. et al. [62]	AD, MCI	ADNI, PCD	ML	81%
Automated detection of brain atrophy patterns based on MRI for the prediction of Alzheimer's disease	Plant, C. et al. [42]	AD, MCI	Clinic of Psychiatry at the Ludwig Maximilian University of Munich	ML	95.83%
An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease	Fonteijn, H.M. et al. [61]	AD, HD	Cognitive Disorders Clinic at the National Hospital for Neurology and Neurosurgery, Multidisciplinary HD Clinic at the National Hospital for Neurology and Neurosurgery	ML	-
Clinical Decision Trees for Predicting Conversion from Cognitive Impairment No Dementia (CIND) to Dementia in a Longitudinal Population-Based Study	Ritchie, Lesley J.; Tuokko, Holly [57]	AD, CIND	CSHA	ML	70.2%
Neural networks for longitudinal studies in Alzheimer's disease	Tandon, R. et al [65]	AD, MCI	LAARC	ML	87.0%
Morphological hippocampal markers for automated detection of alzheimer's disease and mild cognitive impairment converters in magnetic resonance images	Ferrarini, L. et al. [46]	AD, MCI	Laboratory of Epidemiology, Neuroimaging, and Telemedicine, at the IRCSS San Giovanni di Dio-FBF in Brescia	ML	80.0%
A comparison of three brain atlases for MCI prediction	Ota, K. et al. [49]	AD, MCI	SEAD-J	ML	77.9%
Multiplexed immunoassay panel identifies novel CSF biomarkers for Alzheimer's disease diagnosis and prognosis	Craig S. et al. [30]	AD, MCI	WU-ADRC	ML	-
Modeling screening, prevention, and delaying of Alzheimer's disease: An early-stage decision analytic model	Furiak, N.M. et al. [73]	AD	-	MS	No Prognostic accuracy
Estimation and validation of a multiattribute model of alzheimer disease progression	Stallard, E. et al. [74]	AD	Predictor's Study (1989-2001) and Predictor's Study 2 (1997-2007)	MS	No Prognostic accuracy

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	2
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	3-7
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	7
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	9
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	9-11
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	10
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	9-10
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	10-12
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	12-13
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	12-13
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	11-12
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	12-13
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	13

PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	13
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	13
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	14
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	12-24
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	12
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	13-27
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	13-27
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	13
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	27
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	31
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	32
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	32

From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit: **www.prisma-statement.org**.

Study V

17 A decision tree multifactorial approach for predicting dementia in a 10 years' time

Manuscript

A decision tree multifactorial approach for predicting dementia in a 10 years' time

Ana Luiza Dallora, Leandro Minku, Emilia Mendes, Mikael Rennemark, Peter Anderberg, Johan Sanmartin Berglund

Abstract

Background: Dementia is a complex neurological disorder, which little is known about its mechanisms. It affects the older adults population causing a progressive cognitive decline that can become severe enough to impair an individual's independence and functioning. To date no therapeutic treatment was identified to revert or to alleviate its symptoms. Therefore, prognosis research aiming to identify variable risk factors of dementia within a large enough time frame to prevent or delay its onset is greatly important.

Objective: This study investigates the use of a decision tree multifactorial approach for the prognosis of dementia in individuals within a 10-year time frame, using data on individuals who were not diagnosed with this disorder at baseline.

Methods: This study used data from the Swedish National Study on Aging and Care, consisting of 726 subjects (313 males and 413 females), of which 91 presented a diagnosis of dementia at the 10-year study mark. A K-nearest neighbors multiple imputation method was employed to handle the missing data; a wrapper feature selection was utilized to select the best features subset from a set of 75 variables; characterizing factors are related to demographic, social, lifestyle, medical history, biochemical test, physical examination, psychological assessment and diverse health instruments relevant to dementia evaluation. Lastly, a cost-sensitive decision tree approach was applied in order to build predictive models in a stratified nested cross-validation experimental setup.

Results: The proposed approach achieved an AUC of 0.745 and Recall of 0.722 for the prognosis of dementia within a 10-year time frame. Our findings showed that most of the variables selected by the tree are related to modifiable risk factors, of which physical strength was an important factor across all ages of the sample. Also, there was a lack of variables related to the health instruments routinely used for the

dementia diagnosis, suggesting that perhaps they may not be sensitive enough to predict the conversion to dementia within a 10-year window.

Conclusions: The proposed model identified possible modifiable factors that could be used to predict conversion to dementia within a 10-year time frame; possible interventions on such factors could be investigated in order to delay or prevent the dementia onset.

1. Introduction

1.1 Background

Dementia refers to a set of complex neurological disorders that are characterized by progressive cognitive deterioration and increasing disabilities that affect the individual's independence, social and professional functioning [1,2]. It affects mostly the older adult population, where early symptoms are characterized by difficulty in remembering recent events, and evolve to more serious ones like disorientation, mood swings, aggravated memory loss, confusion, changes in behavior, impaired gait, and difficulty in speaking and swallowing, all of which lead to very poor quality of life [1,3]. Additionally, mortality risks for individuals with dementia are two times higher than for persons without dementia; and they may also have to handle comorbidities that may play a role in accelerating their decline in functioning [4]. To aggravate this scenario, there is no known cure for dementia and the available symptomatic treatments show limited benefits in regards to the cognition [3].

There is still significant uncertainty surrounding dementia pathology, such as what exactly triggers it and how its progression unfolds. Such uncertainty makes the development of dementia treatments and interventions much more difficult [1].

Another difficult aspect of dementia disorder is that it also brings negative consequences to the caregivers (usually next of kin) of the individuals affected by it, whose burden cannot be overlooked [5]. These persons deal with this complex condition for years and report low confidence in managing it, in addition to also being prone to negative health outcomes themselves, like high levels of strain, depressive symptoms, risk of alcohol-related problems, etc [1,5]. The families of the individuals affected by dementia also suffer considerable financial impact due to the costs of care and the family's reduction in income [6].

In the public health sphere, the dementia disorder represents a significant cost with prospects to increase in the coming decades. In 2015, the global sum of direct medical costs, social and informal care for dementia was estimated to be US$ 818 billion, which amounted to 1.1% of the global gross domestic product [6]. Since old age is

believed to be the biggest risk factor for dementia, and life expectancy is growing worldwide, the global costs for dementia are estimated to rise to US$ 2 trillion by the year of 2030. This can place a huge strain on healthcare systems around the world [6]. With regard to the health economics of dementia, nowadays, the main costs associated with this disorder are directed towards compensating for the decline in an individual's functioning, instead of treatment or efforts in prevention [1].

1.2 Addressing the dementia epidemic

Dementia research has been very focused upon identifying and validating biomarkers to predict the development of dementia from its prodromal stage i.e. mild cognitive impairment (MCI) [7]. One of the most significant initiatives in this direction is the Alzheimer's Disease Neuroimaging Initiative (ADNI), which aims to identify biomarkers to be used in treatment trials and pharmaceutical research [8]. Research related to dementia treatments is important not only to search of a cure, but also to develop drugs aimed to delay the progression and to shorten the time spent in the most severe stages of this disorder, characterized by the lowest quality of life for the patients and families and increased costs for healthcare systems [1].

Note, however, that the abovementioned research focuses on persons who are already at a major risk of developing dementia, since reported conversion rates from MCI to dementia can be as high as 40% [9–11]. An alternative and also very important research direction to address this epidemic is the prevention of new dementia cases. Prevention and risk reduction are pointed out by the World Health Organization as key elements to focus on, so to decrease the worldwide impact of dementia [1,6]. In this sense, the identification of modifiable risks is imperative and prognostic estimates become more and more important. Prognostic estimates are useful to identify patterns of disease progression, to support public entities in creating and maintaining healthcare programs to address the epidemic, and to aid patients in understanding their condition in order to participate in shared decisions with health providers [12].

However, any research investigating a disorder as complex as dementia is difficult, especially in regard to traditional trials, as they suffer from major obstacles. For instance, certain health conditions that have high prevalence on older persons (e.g. cardiovascular disease) cannot be left untreated in a control group; blinding is a challenge in trials related to lifestyle interventions; and since dementia is a multifactorial disorder, the assumption of one-dimensionality may hinder the results [1,13]. Another significant challenge in such research is that evidence shows that changes in the brain of individuals who come to develop dementia can start from 10 to 15 years before the diagnosis; therefore, preventive interventions should take this

into account, despite the risk that such considerations could make the research efforts costly [14].

1.3 Decision trees for the prognosis of dementia using the SNAC database

Due to the aforementioned challenges, the research on modifiable risks and prevention of dementia, could benefit from data-centered approaches. In regard to prognosis, a technique that is very useful is the Decision Tree (DT). The DT creates a structure for representing knowledge in the form of IF-THEN rules that are intuitive and easy to interpret, and it is widely used in medical research for studying prognostic subgroups [15]. In the context of health data, the DT creates subgroups with minimal intra-variability and maximal inter-variability [16].

Since the DT is a data-centered approach, the quality and length of a possible prognostic estimate is very dependent on the data used to build such types of models. The Swedish National Study on Aging and Care (SNAC) encompasses a longitudinal cohort that is collecting multifactorial data from the older adult population in Sweden for more than 15 years, thus offering the potential for the investigation of long-term dementia prognosis. The SNAC project was designed to study the aging population's health as well as the provided social care, and contains a database with data regarding physical examination, psychological assessment, social factors, lifestyle factors, medical history, etc. The SNAC cohort contains a subset of the aging population in Sweden that comprehends individuals aged 60 years and higher in defined age groups (60, 66, 72, 78, 81, 84, 87, 90, 93 and 96+ years) [17]. The SNAC data is collected from four sites, which represent two Swedish counties, one borough and one municipality. They are respectively Skåne, Blekinge, Kungsholmen and Nordanstig. Such range aims to gather representative national data from both urban and rural areas.

1.4 Study's aim

This study investigates a DT approach for the prognosis of dementia, which considers multiple types of factors as predictors, including social, lifestyle, medical history, blood test, physical examination, demographic (age, gender) psychological and the assessment of diverse health instruments relevant to the dementia evaluation. This DT approach aims at investigating the prognosis of the older individuals in the SNAC cohort, who did not present a diagnosis of dementia at baseline (2000 to 2003) and their development (or not) to dementia at the 10 year mark of the study (2010 to 2013).

2. Material and methods

2.1 Population

The population used herein was the baseline examination of the SNAC-Blekinge, with data ranging from 2001 to 2003, and characterizing subjects based in a urban area in the south-east part of Sweden. The recruited subjects underwent extensive examinations and interviews by physicians, nurses and psychologists [17]. In case of disabilities, an examination team did the examinations and interviews at the subject's home and proxy interviews were conducted with relatives of the subjects whenever necessary, and upon the subjects' consent [17].

The following criteria were used to determine the exclusion of subjects in this study: (i) subjects who already had dementia at baseline; (ii) subjects who had missing values at the outcome variable (dementia diagnosis); (iii) subjects who presented more than 10% of missing values in the input variables; (iv) subjects deceased before the 10 years study mark; and (v) subjects who were diagnosed with dementia before the 10 years mark, as they could already have advanced progress of dementia at baseline.

The study sample consisted of 726 subjects (313 males and 413 females), of which 91 (12.5%) were given a diagnosis of dementia at the 10-year mark. The demographics of the study sample are shown in table 1.

Table 1. Demographics of the study sample

		Age at baseline								
Diagnosis	**Gender**	**60**	**66**	**72**	**78**	**81**	**84**	**87**	**90+**	**Total**
No dementia at 10 years mark	**Male**	81	72	47	36	27	19	2	0	284
	Female	81	92	67	38	35	27	7	4	351
Dementia at 10 years mark	**Male**	1	3	3	5	8	7	2	0	29
	Female	1	3	7	12	11	15	12	1	62

2.2 Ethics and data privacy

This study was carried out in accordance with the Declaration of Helsinki and was approved by the Research Ethics Committee at Lund (LU 604-00, LU 744-00). Written informed consent was collected from all subjects. All data was anonymized and stratified by age and gender.

2.3 Outcome variable: diagnosis of dementia at the SNAC 10-year mark

The outcome variable that the DT aims to estimate is the dementia diagnosis given by physicians 10 years from the SNAC baseline. The given diagnosis of dementia followed the guidelines of the International Statistical Classification of Diseases and Related Health Problems - 10th Revision (ICD-10) [18] and the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) [19].

No distinctions between subtypes of dementia were made since mixed pathologies are common and uncommon subtypes tend to be misdiagnosed by Alzheimer's disease [20].

2.4 Input variables

The input variables used to build the DT prognostic estimate models were selected by senior researchers specialized in geriatrics and gerontology (authors 4 and 6) who adopted a broad approach in order to consider variables of diverse types that could influence the onset of dementia. In total, 75 variables were selected, which encompassed the following categories: demographic, social, lifestyle, medical history, blood test, physical examination, psychological and the assessment of multiple health instruments relevant to the dementia evaluation, at the study baseline (2000-2003). Table 2 shows the list of the selected variables. Descriptions, possible values and number of missing values of all input variables are shown in Appendix 1.

2.5 Data preparation

The data preparation step was characterized by the imputation of missing data. To this end, a K-Nearest Neighbors (KNN) multiple imputation approach was employed [31]. This technique works by finding the K entries of data that are the most similar (near) to an entry that contains missing data. Then, the KNN imputation fills its missing values with the mean (in the case of numeric variables) or the most frequent value of the K most similar neighbors (in the case of categorical variables) [31].

In the present study, the KNN imputation was applied separately for entries of the majority class (no dementia at 10 years mark) and the minority class (dementia at 10 years mark). This was done because the sample used to build the prognostic estimates presented a pronounced class imbalance (12.5% on the minority class to 87.5% on the majority class), so the risk of contaminating the minority class with data from the majority class was mitigated.

The number of nearest neighbors used in the KNN imputations was set to K=3. This choice was based on literature findings that argue about limiting the number of K to avoid distortion of the original variability of the data, as in the critical evaluation of Nearest Neighbors Imputation on medical datasets by Beretta and Santaniello that recommend K=3 [32].

Table 2. List of input variables that were selected by the dementia specialists

Variable Type	Variables
Demographic	Age, Gender,
Social	Education, Holds or not a Religious Belief, Participation in Religious Activities, Voluntary Association, Social Network, Support Network, Loneliness
Lifestyle	Light Exercise, Alcohol Consumption, Alcohol Quantity, Working State at 65 years, Physical Workload, Present Smoker, Past Smoker Number of Cigarettes a Day, Social Activities, Physically Demanding Activities, Leisure Activities
Medical History	Number of Medications, Family History of Importance, Myocardial Infarction, Arrhythmia, Heart Failure, Stroke, TIA/RIND, Diabetes Type 1, Diabetes Type 2, Thyroid Disease, Cancer, Epilepsy, Atrial Fibrillation, Cardiovascular Ischemia, Parkinson's Disease, Depression, Other Psychiatric Diseases, Snoring, Sleep Apnea, Hip Fracture, Head Trauma, Developmental Disabilities, High Blood Pressure
Biochemical Test	Haemoglobin Analysis, C-Reactive Protein Analysis
Physical Examination	Body Mass Index (BMI), Pain in the last 4 weeks, Heart Rate Sitting, Heart Rate Lying, Blood Pressure on the Right Arm, Hand Strength in Right Arm in a 10s Interval, Hand Strength in Left Arm in a 10s Interval, Feeling of Safety from Rising from a Chair, Assessment of Rising from a Chair, Single-Leg Standing with Right Leg, Single Leg Standing with Left Leg, Dental Prosthesis, Number of Teeth
Psychological	Memory Loss, Memory Decline, Memory Decline 2, Abstract Thinking, Personality Change, Sense of Identity
Health Instruments	Sense of Coherence [21], Digit Span Test [22], Backwards Digit Span Test [22], Livingston Index [23], EQ5D Test [24], Activities of Daily Living [25], Instrumental Activities of Daily Living [26], Mini-Mental State Examination [27], Clock Drawing Test [28], Mental Composite Score of the SF-12 Health Survey [29], Physical Composite Score of the SF-12 Health Survey [29], Comprehensive Psychopathological Rating Scale [30]

2.6 Model building

The DT algorithm was applied to predict dementia 10 years from the baseline. This prediction is given by building a binary classifier, which classifies subjects in the

sample as to 'no dementia' (negative class) or 'dementia' (positive class), as specified by the outcome variable.

The DT approach starts by testing each input variable to their capacity of classifying the right examples in regard to an outcome. The best one is selected to be the root node of the tree and the descendent nodes are defined as its possible values (or ratios), thus creating a rule. The data is then partitioned following this rule. This process continues for each new created node, recursively, until no more splits are possible [33]. The nodes created by the last splits in the tree are called terminal nodes and they provide the class prediction.

In order to build the DT, two main problems that are common regarding studies with medical databases arose: (i) a high class imbalance (12.5% on the minority class and 87.5% on the majority class); and (ii) a high number of input variables in relation to the number of entries. Both are known to hinder the performance of DT algorithms and need to be addressed [34,35]. To address the first problem, a cost-sensitive learning approach [36] was used with the DT, and for the second a wrapper feature selection was employed in order to reduce the number of variables [34]. These are detailed in the following sections.

2.6.1 Cost-Sensitive Learning

The consequence of having an imbalance database is that the trained classifier tends to be biased towards the majority class. The cost-sensitive learning approach works by applying a heavier penalty on misclassifying the minority class. In doing so, it attributes higher weights to the minority class in order to change the class distribution. Weights were calculated as follows:

$$w_i = \frac{1}{n_i} \times 0.5 \quad (1)$$

where w_i represents the weight attributed to a class i and n_i is the number of entries of the class i.

The weights in the DT algorithm are used in the criteria for finding splits and also on the terminal nodes in the class prediction, which is given by the weighted majority vote [36].

2.6.2 Wrapper Feature Selection

To address the high number of input variables in relation to the number of entries, the Recursive Feature Elimination (RFE) feature selection method was employed in order to select the most important variables for the classification. This wrapper method assesses multiple different models composed of different combinations of input variables in order to find the optimal subset of variables to maximize a performance metric of choice [34].

2.7 Experimental Setup

Class weighting and RFE were performed before training the DT models. For the specific DT algorithm implementation in R (version 3.4.1) used in this study, we refer to boo by Kuhn and Johnson [34]. The R package employed for training the DT and feature selection was the caret, and imputation was performed with the VIM package.

All experiments were performed using a stratified nested cross-validation setup. This approach performs outer and inner cross-validations in a way that in each iteration one fold of the outer cross-validation is used for testing and the remaining for the inner cross-validation, which is responsible for hyperparameter tuning [37,38]. Using the nested approach for cross-validation produces a more reliable estimate of error, since data being used to estimate the model's performance is not being used for optimization. This avoids the high risk of producing over-optimistic results which occurs when using only one test set, as done in more traditional machine learning experimental setups [37,38]. After performing nested cross-validation and reporting the values of the evaluation metrics for each of the outer cross-validation test sets, we select the model which produces the median performance results of the outer cross-validation test sets for further analysis.

The stratified part of the proposed approach means that for all the folds of both inner and outer cross-validations, the proportions of both minority and majority classes remain the same as in the original sample. Due to class imbalance, the experiments were conducted in a 5-fold outer, 4-fold inner nested cross-validation setup in order to have enough examples of the minority class in each fold.

2.8 Evaluation metrics

The evaluation metrics considered for the experiments were: Area Under the Curve (AUC), Accuracy' and Recall and Precision on the positive class (dementia at the 10 year mark of the study).

3. Results

The performance metrics obtained in each of the outer cross-validation test sets of the nested stratified cross-validation procedure are shown in table 3. General lower values for the Precision metric were a consequence of the class weighting approach employed to deal with the significant class imbalance of the used sample. This makes the misclassification of the positive class more costly than the inverse, minimizing the gravest error. In a diagnostic tool, this would be a poor scenario, but in the present

study the aim is to conduct a prognostic analysis on the factors that could influence the development of dementia, 10 years later, so this type of error has a reduced magnitude.

Table 3. Performance results in each of the 5 outer cross-validation test sets

Model	AUC	Accuracy	Recall	Precision
1	0.718	0.664	0.790	0.250
2 (median)	**0.735**	**0.745**	**0.722**	**0.289**
3	0.827	0.738	0.944	0.315
4	0.763	0.752	0.778	0.304
5	0.712	0.662	0.778	0.237

The median model, which presents the median AUC, was model 2 with an AUC of 0.735, Accuracy of 0.745, Recall of 0.722 and Precision of 0.289. The DT that visually represents this model is shown in figure 1. A description of the variables, with possible values, selected by the final tree is shown in table 4.

Table 4. Summary of the variables present in the DT

Variable	Description	Values
Age	Subject's age at baseline	Numeric (years)
StandingTest_Left	Single leg standing test with left leg. Best value in seconds of three tries	Numeric (Seconds)
PastSmoker_CigarsDay	Cigarettes/day, on average, before quitting smoking.	Numeric
Diabetes_type2	Medical history of diabetes type 2	Yes; No
Alcohol_Consumption	Alcohol consumption frequency	Never; Once a month or more rarely; 2-4 times a month ; 2-3 times a week; more than 4x a week
Medications	Number of medications taken regularly by the subject	Numeric
HandStrength_Right	The subject's hand strength, measured in an interval of 10 seconds, for the right hand.	Numeric (Newtons)
HandStrength_Left	The subject's hand strength, measured in an interval of 10 seconds, for the left hand.	Numeric (Newtons)
BMI	Subject's BMI	Numeric (kg/m^2)
DigitSpan_Backwards	The number of correct sequences on the Backwards Digit Span Test [22].	Numeric
PresentSmoker	Subject's habit of smoking at baseline.	No, never smoked; no, quit smoking; yes, smoke sometimes; Yes, smoke regularly

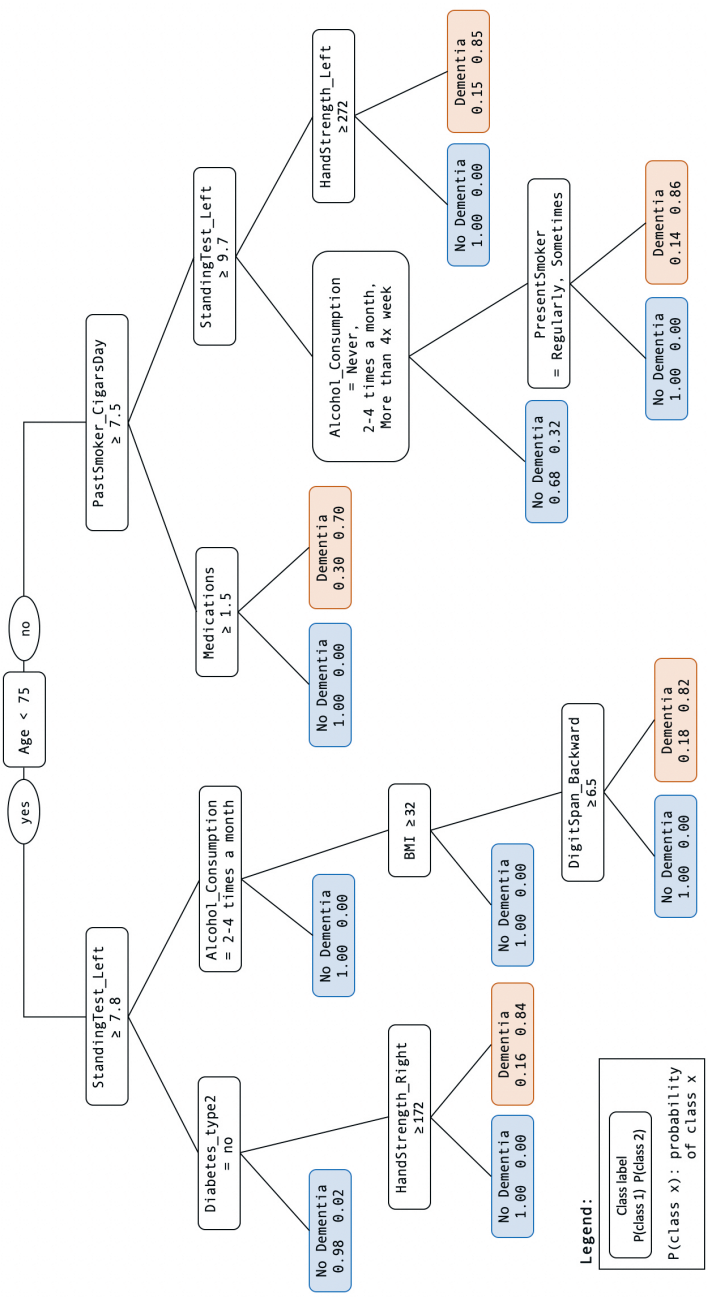


Figure 1. Plot of the Decision Tree of the median model.

Age was chosen as the DT's root node, at a threshold of 75 years, so indicating high importance for the classification. This is in line with the literature that attributes age as the major risk factor for dementia, pointing out increased risk for the ages older than 65 years, which accounts for 95% of the cases [1]. The threshold chosen by the DT was 75 years, probably due to the study's sample, comprising subjects older than 60 years (at the baseline), and excluding subjects on the only onset who developed dementia before the 10 year mark of the study, as stated by the study's exclusion criteria.

Using the threshold of 75 years, the following sections will detail the DT classification for branches of subjects older and younger than 75 years.

3.1 Prediction of dementia for subjects 75 years and older at the baseline

The most striking factor of the branch of subjects 75 years and older is the presence of two variables related to smoking, which not only relate to whether an individual is a present smoker at baseline but also if they smoked in the past but not anymore. Smoking is a known risk factor for dementia, and literature findings point out that smoking in midlife and late-life increases the risk of developing dementia (when compared with non-smokers) [39], and may also be related to an accelerated cognitive decline (assessed by the Mini-Mental State Examination [27]) [40]. The variable associated with the number of medications taken regularly by the subjects is related to the past heavy smokers, and it could be interpreted as an indicator of comorbidities (possibly connected to the heavy smoking itself). The presence of comorbidities may be associated to an accelerated decline in functioning in demented individuals [2]. However, the DT pointed out that individuals who on average took less regular medications were leading to dementia 10 years later, which might indicate individuals prone to not seeking medical care, and therefore with a higher chance of presenting undiagnosed conditions.

The inclusion of the hand strength (grip test) and single-leg balancing point out that physical strength is protective against dementia. A poor value on these variables could be related to frailty or sedentary life, which is known to be a risk factor of dementia [1].

The last variable on the branch related to the subjects 75 years and older is the alcohol consumption frequency. The evidence about the protective benefits of alcohol consumption are not very consolidated in the literature, as there are no randomized clinical trials performed on the subject, but observational studies suggest that a moderate consumption seems to be protective against cardiovascular disease (a

known risk factor of dementia [1]), while abstinence or heavy drinking seems to be a factor of risk [41]. This is in line with the thresholds chosen by the tree.

3.2 Prediction of dementia for subjects younger than 75 years at baseline

The variables in the branch related to the subjects younger than 75 years are somewhat similar to the older subjects in relation to hand strength (grip test), single-leg balancing and alcohol consumption frequency.

Another important variable in this branch is the subject's medical history in regard to diabetes Type 2. The risk of dementia in individuals with this condition is reported to be higher than in individuals without it [42], and there is a theory being actively researched as to consider the Alzheimer's Disease as a diabetes Type 3, as it seems that there are common molecular and cellular features in diabetes Types 1 and 2, which are associated with cognitive decline in the older adult population [43]. It is also interesting to notice that in the branch of individuals with diabetes Type 2 the following node that relates to hand strength has a threshold much lower when comparing to the branches of subjects 75 years or older, with a difference of 100 Newtons. This might be a direct consequence of the weakening effects of diabetes Type 2.

The BMI on the threshold of 32 kg/m2 was also selected as a split in the tree, which is commonly referred to as overweight. The relationship between BMI and dementia, and it being a risk factor or not, is not consolidated in the literature and the findings are conflicting, where trials on the older adult population showed that to higher BMI values being a protective and a risk factor [44,45]. In the present study, lower values of BMI were predictive factors of dementia in a 10 year time. The only health instrument that was selected as a predictive factor of dementia was the Backwards Digit Span test. This is a neuropsychological instrument that is used to assess the working memory [46]. Since mild memory loss is one of the earliest symptoms of dementia, a poor result in this test might indicate the start of the development of the disorder, but no literature was found to relate scores in Backward Digit Span test and the risk of development of dementia.

4. Discussion

4.1 Main findings

This study proposed a decision tree approach for the 10-year prognosis of dementia, which achieved an AUC of 0.745 and Recall of 0.722. The proposed approach investigated multiple domains in order to derive prognostic estimates, assessing 75

variables related to: demographic, social, lifestyle, medical history, biochemical tests, physical examination, psychological assessment and diverse health instruments relevant to the dementia disorder.

The main findings of our assessment are the following: (i) most of the variables selected by the DT are related to modifiable factors; (ii) physical strength was an important factor in all ages of the study sample; and (iii) the DT selected almost no variables related to psychological assessment health instruments that are used nowadays in the assessment of dementia.

With the exception of age, past smoking, number of medications and the Backwards Digit Span test, the remaining variables present in the prognostic tree can be considered modifiable factors. These are related to physical strength, present smoking, BMI, diabetes Type 2 and alcohol consumption. These results are promising as 10 years could be a considerable time frame for the implementation of interventions for delaying or even preventing dementia, which could mean major individual, societal and financial benefits. However, without further investigation trials on these areas it is not possible to assert that the proposed factors are in fact protective against dementia.

Another important aspect to be discussed is that among the sub-branches of the tree that leads to dementia in 10 years from the baseline, in all of them but one, there was the presence of a node indicating poor physical strength, which may indicate that this is an especially important factor to be considered in further studies.

Additionally, the lack of variables related to psychological assessment and health instruments that are routinely used for the assessment of dementia points out that these instruments may not be sensitive enough for the long-term prediction of dementia, and also be as well insensitive to mild cases that are possibly at the beginning of their development. The only exception was the Backwards Digit Span test, which assesses the working memory of an individual and may present the potential to be used in detecting early memory loss, one of the earlier symptoms of dementia. This is important as most of the cases of dementia are diagnosed when the disorder is at an already advanced stage [6].

Lastly, considering the results, especially in relation to the modifiable factors, they are all already recommended in some form to prevent chronic illnesses and to maintain a healthy life. This could also refer to dementia theories that state that dementia is not an unavoidable consequence of aging and measures to delay and to prevent its onset should be investigated [1].

4.2 Related Work

In regard to related work, we could not find in the literature another study that predicted dementia in 10 years' time, through the use of decision trees, with such broad approach as ours. By broad approach we mean an study that addresses multiple domains (demographic, social, lifestyle, medical history, biochemical tests, physical examination, psychological examination, and diverse health instruments) in a longitudinal population-based sample.

The study by Ritchie and Tuokko (2011) [47] proposed a multi domain longitudinal approach for the prognosis of dementia. However, the study sample subjects already presented a degree of cognitive impairment at baseline; furthermore, its aim was to investigate the conversion to dementia in an interval of 5 years, which achieved an accuracy of 0.702.

However, there were randomized clinical trials (RCT) that proposed multi domain lifestyle interventions in order to investigate the benefits in cognition or dementia incidence. They are detailed next:

The Finnish Geriatric Intervention Study to Prevent Cognitive Impairment and Disability (FINGER) was an RCT which aimed at investigating the effects on cognition of a 2-year lifestyle multi domain intervention that encompassed nutritional guidance, exercise, cognitive training, social activity, and management of metabolic and vascular risk factors [48]. The population of the study was composed of 1200 individuals (60 to 77 years) already at risk of cognitive decline. The FINGER results showed that the individuals in the intervention group had a positive effect in cognition (score in the Neuropsychological Test Battery [49]), even when corrected to sociodemographic, socioeconomic, cognitive, or cardiovascular factors at baseline [50].

The Multidomain Alzheimer Preventive Trial (MAPT) is another RCT that aimed to investigate changes in cognitive functions of 1680 subjects (mean age of 75.3 years) enrolled at memory clinics in a period of 3 years. Three interventions were studied: (i) omega-3 supplementation; (ii) nutritional counseling, physical exercise and cognitive stimulation; and (iii) a combination of the i and ii interventions [51]. The results from the MAPT study presented no significant benefit in regards to any of the interventions in the memory function [52] (assessed by the Free and Cued Selective Reminding test [53]).

The Prevention of Dementia by Intensive Vascular care (preDIVA) RCT also investigated the prevention of the dementia disorder with lifestyle interventions in an interval of 6 years. The population of the study was composed of 3526 randomly selected individuals (70 to 78 years) from health care centers. The intervention

comprised 18 visits to a practice nurse who assessed smoking habits, diet, physical activity, weight, blood pressure, and blood glucose and lipid assessments every 2 years; and offered tailored lifestyle advice according to the study protocol. The primary outcome was measured in the incidence in all-cause dementia, in which the proposed intervention had no significant benefit [54].

All of the three mentioned RCTs proposed interventions in the form of nurse visits who provided specific guidance in regard to lifestyle. The only one that showed significant benefits was the FINGER study.

4.3 Limitations

The high false-positive ratio indicated by the low precision achieved by the proposed DT could be seen as a limitation of the proposed method. This feature would be undesired, especially if our objective was to propose a diagnostic tool; however, the main goal of the proposed approach was to identify factors for the long-term prediction of dementia, minimizing the misclassification of the demented cases, which was moderately achieved.

It would be novel and interesting to build separate models for male and female subjects and investigate gender-specific predictors for dementia, but the number of male subjects which developed dementia in the 10 year mark of the study was very low.

The low number of subjects could be argued to be too low to be generalizable, however, the stratified nested cross-validation method was employed to avoid the selection of an overfitted or overoptimistic model.

The results presented in the present study relate to a Swedish urban sample and may not be generalizable to other populations with different overall socioeconomic status; therefore, further investigation is needed in order to establish the effect of this factor on the results here presented.

4.4 Future work

Future work efforts should be directed to investigate the gender-specific prognosis of dementia, especially with the imbalance shown on the incidence between male and female subjects on the 10 year mark. Also this study is based on a single site of the SNAC study (Blekinge), further studies should be directed to validate the findings on this study in the other SNAC sites.

5. Conclusion

This paper proposed a DT model for the prediction of dementia using data from a longitudinal, population-based study (SNAC-Blekinge), in a broad approach that investigated social, lifestyle, medical history, blood examination, physical and psychological factors, as well as diverse health instruments that are currently in use for the assessment of dementia, achieving an AUC of 0.745 and Recall of 0.722. The model identified diverse modifiable factors that could potentially be intervened in order to attempt a delay or prevention of the dementia onset.

6. Acknowledgements

This study was accomplished within the context of the Swedish National Graduate School for Competitive Science on Ageing and Health (SWEAH) funded by the Swedish Research Council. The PhD student's learning process was supported by the Swedish National Graduate School for Competitive Science on Ageing and Health (SWEAH) funded by the Swedish Research Council.

References

1. Winblad B, Amouyel P, Andrieu S, Ballard C, Brayne C, Brodaty H, et al. Defeating Alzheimer's disease and other dementias: a priority for European science and society. Lancet Neurol. 2016;15: 455–532.
2. Melis RJF, Marengoni A, Rizzuto D, Teerenstra S, Kivipelto M, Angleman SB, et al. The influence of multimorbidity on clinical progression of dementia in a population-based cohort. PLoS One. 2013;8: e84014.
3. Furiak NM, Klein RW, Kahle-Wrobleski K, Siemers ER, Sarpong E, Klein TM. Modeling screening, prevention, and delaying of Alzheimer's disease: an early-stage decision analytic model. BMC Med Inform Decis Mak. 2010;10: 24.
4. Poblador-Plou B, Calderón-Larrañaga A, Marta-Moreno J, Hancco-Saavedra J, Sicras-Mainar A, Soljak M, et al. Comorbidity of dementia: a cross-sectional study of primary care older patients. BMC Psychiatry. 2014;14: 84.
5. Jennings LA, Reuben DB, Evertson LC, Serrano KS, Ercoli L, Grill J, et al. Unmet needs of caregivers of individuals referred to a dementia care program. J Am Geriatr Soc. 2015;63: 282–289.
6. Dua T, Seeher KM, Sivananthan S, Chowdhary N, Pot AM, Saxena S. WORLD HEALTH ORGANIZATION'S GLOBAL ACTION PLAN ON

THE PUBLIC HEALTH RESPONSE TO DEMENTIA 2017-2025. Alzheimer's & Dementia. 2017. pp. P1450–P1451. doi:10.1016/j.jalz.2017.07.758
7. Dallora AL, Eivazzadeh S, Mendes E, Berglund J, Anderberg P. Machine learning and microsimulation techniques on the prognosis of dementia: A systematic literature review. PLoS One. 2017;12: e0179804.
8. Disease Neuroimaging A 's. The Alzheimer's disease neuroimaging initiative: progress report and future plans. Alzheimers Dement. 2010. Available: https://alzjournals.onlinelibrary.wiley.com/doi/abs/10.1016/j.jalz.2010.03.007
9. Mitchell AJ, Shiri-Feshki M. Rate of progression of mild cognitive impairment to dementia -meta-analysis of 41 robust inception cohort studies. Acta Psychiatrica Scandinavica. 2009. pp. 252–265. doi:10.1111/j.1600-0447.2008.01326.x
10. Luis CA, Barker WW, Loewenstein DA, Crum TA, Rogaeva E, Kawarai T, et al. Conversion to Dementia among Two Groups with Cognitive Impairment. Dement Geriatr Cogn Disord. 2004;18: 307–313.
11. Geslani DM, Tierney MC, Herrmann N, Szalai JP. Mild Cognitive Impairment: An Operational Definition and Its Conversion Rate to Alzheimer's Disease. Dementia and Geriatric Cognitive Disorders. 2005. pp. 383–389. doi:10.1159/000084709
12. Ohno-Machado L. Modeling medical prognosis: survival analysis techniques. J Biomed Inform. 2001;34: 428–439.
13. Iqbal K, Grundke-Iqbal I. Alzheimer's disease, a multifactorial disorder seeking multitherapies. Alzheimers Dement. 2010;6: 420–424.
14. Dubois B, Hampel H, Feldman HH, Scheltens P, Aisen P, Andrieu S, et al. Preclinical Alzheimer's disease: Definition, natural history, and diagnostic criteria. Alzheimers Dement. 2016;12: 292–323.
15. A. N, Kudus A. Decision Tree for Prognostic Classification of Multivariate Survival Data and Competing Risks. Recent Advances in Technologies. 2009. doi:10.5772/7429
16. Esteban C, Arostegui I, Moraza J, Aburto M, Quintana JM, Pérez-Izquierdo J, et al. Development of a decision tree to assess the severity and prognosis of stable COPD. Eur Respir J. 2011;38: 1294–1300.
17. Lagergren M, Fratiglioni L, Hallberg IR, Berglund J, Elmståhl S, Hagberg B, et al. A longitudinal study integrating population, care and social services data. The Swedish National study on Aging and Care (SNAC). Aging Clin Exp Res. 2004;16: 158–168.

18. Organization WH, Others. ICD-10 version: 2010 [Internet]. Geneva (CH): World Health Organization. 2010.
19. Castillo RJ, Carlat DJ, Millon T, Millon CM, Meagher S, Grossman S, et al. Diagnostic and statistical manual of mental disorders. American Psychiatric Association Press, Washington, DC. 2007. Available: http://dissertation.argosy.edu/chicago/fall07/pp7320_f07schreier.doc
20. Organization WWH. Dementia: a public health priority. WHO Geneva; 2012.
21. Antonovsky A. The structure and properties of the sense of coherence scale. Soc Sci Med. 1993;36: 725–733.
22. Wechsler Adult Intelligence Scale (All Versions). SpringerReference. Berlin/Heidelberg: Springer-Verlag; 2011.
23. Livingston G, Blizard B, Mann A. Does sleep disturbance predict depression in elderly people? A study in inner London. Br J Gen Pract. 1993;43: 445–448.
24. Brooks R. EuroQol: the current state of play. Health Policy. 1996;37: 53–72.
25. Katz S. Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living. J Am Geriatr Soc. 1983;31: 721–727.
26. Lawton MP, Brody EM. Assessment of older people: self-maintaining and instrumental activities of daily living. Gerontologist. 1969;9: 179–186.
27. Folstein MF, Folstein SE, McHugh PR. Mini-mental state': A practical method for grading the cognitive state of patients for the clinician. J Psychiatry Res 1975; 12: 189--198. External Resources Pubmed/Medline (NLM) CrossRef (DOI) Chemical Abstracts Service (CAS). 1962.
28. Agrell B, Dehlin O. The clock-drawing test. Age Ageing. 1998;27: 399–403.
29. Jenkinson C, Layte R. Development and testing of the UK SF-12. J Health Serv Res Policy. 1997;2: 14–18.
30. Montgomery SA, Asberg M. A new depression scale designed to be sensitive to change. Br J Psychiatry. 1979;134: 382–389.
31. Zhang S. Nearest neighbor selection for iteratively kNN imputation. J Syst Softw. 2012;85: 2541–2552.
32. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. BMC Med Inform Decis Mak. 2016;16 Suppl 3: 74.
33. Mitchell TM, Others. Machine learning. McGraw-hill New York; 1997. Available: https://profs.info.uaic.ro/~ciortuz/SLIDES/2017s/ml0.pdf
34. Kuhn M, Johnson K. Applied Predictive Modeling. Springer, New York, NY; 2013.

35. Kai Ming Ting. An instance-weighting method to induce cost-sensitive trees. IEEE Trans Knowl Data Eng. 2002;14: 659–665.
36. Chen C, Liaw A, Breiman L, Others. Using random forest to learn imbalanced data. University of California, Berkeley. 2004;110: 24.
37. Wainer J, Cawley G. Nested cross-validation when selecting classifiers is overzealous for most practical applications. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1809.09446
38. Krstajic D, Buturovic LJ, Leahy DE, Thomas S. Cross-validation pitfalls when selecting and assessing regression and classification models. J Cheminform. 2014;6: 10.
39. Ohara T, Ninomiya T, Hata J, Ozawa M, Yoshida D, Mukai N, et al. Midlife and Late- Life Smoking and Risk of Dementia in the Community: The Hisayama Study. J Am Geriatr Soc. 2015;63: 2332–2339.
40. Anstey KJ, von Sanden C, Salim A, O'Kearney R. Smoking as a risk factor for dementia and cognitive decline: a meta-analysis of prospective studies. Am J Epidemiol. 2007;166: 367–378.
41. Vogel RA. Alcohol, heart disease, and mortality: a review. Rev Cardiovasc Med. 2019;3: 7–13.
42. Biessels GJ, Staekenborg S, Brunner E, Brayne C, Scheltens P. Risk of dementia in diabetes mellitus: a systematic review. Lancet Neurol. 2006;5: 64–74.
43. Kandimalla R, Thirumala V, Reddy PH. Is Alzheimer's disease a Type 3 Diabetes? A critical appraisal. Biochim Biophys Acta Mol Basis Dis. 2017;1863: 1078–1089.
44. Xu WL, Atti AR, Gatz M, Pedersen NL, Johansson B, Fratiglioni L. Midlife overweight and obesity increase late-life dementia risk: a population-based twin study. Neurology. 2011;76: 1568–1574.
45. Qizilbash N, Gregson J, Johnson ME, Pearce N, Douglas I, Wing K, et al. BMI and risk of dementia in two million people over two decades: a retrospective cohort study. Lancet Diabetes Endocrinol. 2015;3: 431–436.
46. Choi HJ, Lee DY, Seo EH, Jo MK, Sohn BK, Choe YM, et al. A normative study of the digit span in an educationally diverse elderly population. Psychiatry Investig. 2014;11: 39–43.
47. Ritchie LJ, Tuokko H. Clinical decision trees for predicting conversion from cognitive impairment no dementia (CIND) to dementia in a longitudinal population-based study. Arch Clin Neuropsychol. 2011;26: 16–25.
48. Kivipelto M, Solomon A, Ahtiluoto S, Ngandu T, Lehtisalo J, Antikainen R, et al. The Finnish Geriatric Intervention Study to Prevent Cognitive

Impairment and Disability (FINGER): Study design and progress. Alzheimer's & Dementia. 2013. pp. 657–665. doi:10.1016/j.jalz.2012.09.01
49. Harrison J, Minassian SL, Jenkins L, Black RS, Koller M, Grundman M. A neuropsychological test battery for use in Alzheimer disease clinical trials. Arch Neurol. 2007;64: 1323–1329.
50. Rosenberg A, Ngandu T, Rusanen M, Antikainen R, Bäckman L, Havulinna S, et al. Multidomain lifestyle intervention benefits a large elderly population at risk for cognitive decline and dementia regardless of baseline characteristics: The FINGER trial. Alzheimers Dement. 2018;14: 263–270.
51. Vellas B, Carrie I, Gillette-Guyonnet S, Touchon J, Dantoine T, Dartigues JF, et al. MAPT STUDY: A MULTIDOMAIN APPROACH FOR PREVENTING ALZHEIMER'S DISEASE: DESIGN AND BASELINE DATA. J Prev Alzheimers Dis. 2014;1: 13–22.
52. Yassine HN, Schneider LS. Lessons from the Multidomain Alzheimer Preventive Trial. The Lancet Neurology. 2017. pp. 585–586. doi:10.1016/s1474-4422(17)30227-2
53. Grober E, Buschke H, Crystal H, Bang S, Dresner R. Screening for dementia by memory testing. Neurology. 1988;38: 900–903.
54. van Charante EPM, Richard E, Eurelings LS, van Dalen J-W, Ligthart SA, van Bussel EF, et al. Effectiveness of a 6-year multidomain vascular care intervention to prevent dementia (preDIVA): a cluster-randomised controlled trial. Lancet. 2016;388: 797–805.

Appendix 1 – Input variables

Table A1. List of the input variables with description, possible values and number of missing values

Variable Type	#	Variable	Description	Values	# Missing Values
Basic	1	Gender	Subject's gender	• Male • Female	0
	2	Age	Subject's age	numeric	0
	3	BMI	Subject's BMI	numeric	0
Social	4	Education	Subject's highest level of education	• Unfinished elementary school • Elementary school • High school • Vocational training • University • Doctorate	51 (5.2%)
	5	Religion	"Do you have a religious belief?"	• Yes • No	57 (5.8%)
	6	ReligiousActivities	"Do you participate in religious activities?"	• Not at all • Sometimes • Often	39 (4%)
	7	VoluntaryAssociation	"If you are a member of a voluntary association, could you say that you feel strong associations with this association and its members?"	• Not a member of the association • Highly • To some extent • Not especially • Not at all	56 (5.7%)
	8	SocialNetwork	Assesses the subject's social network in terms of personal relationships and social interactions. See "Description 1"	• Very bad social network • Bad social network • Normal social network	57 (5.8%)

	9	SupportNetwork	Assesses the subject's support network in terms of people that can help them with life issues. See "Description 2"	numeric	53 (5.4%)
	10	Loneliness	Assesses the subject's feeling of loneliness. See "Description 3"	• Very high level of loneliness • Some level of loneliness • Normal level of loneliness	76 (7.8%)
Lifestyle	11	Exercise	Frequency of light exercise in last 12 months.	• Wheelchair or balance problems • Never • Once a month • 2-3 times a month • More than 3 times a month • Every day	99 (10.1%)
	12	Alcohol_Consumption	"How often do you drink alcohol"	• Never • Once a month or more rarely • 2-4 times a month • 2-3 times a week • 4 or more times a week	46 (4.7%)
	13	Alcohol_Quantity	"How many "glasses" do you drink on a typical day when you drink alcohol?"	• Do not drink alcohol • 1-2 • 3-4 • 5-6 • 7-9 • 10+	94 (9.6%)
	14	Working65	"When did you stop working?"	• Stopped working before 65 years	0

				• Worked until 65 years • Still working/worked after 65 years	
	15	PresentSmoker	Do you smoke?	• Yes, smoke regularly • Yes, smoke sometimes • No, quitted smoking • No, never smoked	34 (3.5%)
	16	PastSmoker_CigarsDay	If you have quitted smoking, how many cigarettes/day did you smoke on average before you stopped?	numeric	63 (6.4%)
	17	SocialActivities	Assesses the subject's engagement in sociocultural activities in the past 12 months. See "Description 4".	numeric	85 (8.6%)
	18	PhysicallyDemandingActivities	Assesses the subject's engagement in physically demanding activities in the past 12 months. See "Description 5".	numeric	85 (8.6%)
	19	LeisureActivities	Assesses the subject's engagement in leisure and hobby activities in the past 12 months. See "Description 6".	numeric	96 (9.8%)
Medical History	20	Medications	Number of medications taken regularly by the subject	numeric	0
	21	FamilyHistory	Subject's family (first degree relatives) medical history of importance, regarding cardiovascular disease, Parkinson's disease and dementia	• Yes • No	2 (≈0%)
	22	Infarct	Subject's history of infarct	• Yes • No	8 (≈0%)
	23	Arrhythmia	Subject's history of arrhythmia	• Yes • No	13 (1.3%)

	24	HeartFailure	Subject's history of heart failure	• Yes • No	1 (≈0%)
	25	Stroke	Subject's history of stroke	• Yes • No	6 (≈0%)
	26	TIARIND	Subject's history of Transient Ischemic Attacks or Reversible Ischemic Neurological Deficit (TIA/RIND)	• Yes • No	11 (1.1%)
	27	Diabetes_type1	Subject's history of diabetes type 1	• Yes • No	1 (≈0%)
	28	Diabetes_type2	Subject's history of diabetes type 2	• Yes • No	2 (≈0%)
	29	ThyroidDisease	Subject's history of thyroid disease	• Yes • No	5 (≈0%)
	30	Cancer	Subject's history of cancer	• Yes • No	1 (≈0%)
	31	Epilepsy	Subject's history of epilepsy	• Yes • No	0
	32	AtrialFibrillation	Subject's history of atrial fibrillation	• Yes • No	75 (7.7%)
	33	IschemicSigns	Subject's history of ischemic signs	• Yes • No	76 (7.8%)
	34	Parkinsons	Subject's history of Parkinson's disease	• Yes • No	3 (≈0%)
	35	Depression	Subject's history of depression	• Yes • No	0
	36	OtherPsychiatricDiseases	Subject's history of other psychiatric diseases	• Yes • No	5 (≈0%)
	37	Snoring	Subject's history of snoring	• Yes • No	3 (≈0%)
	38	SleepApnea	Subject's history of sleep apnea	• Yes • No	7 (≈0%)

	39	HipFracture	Subject's history of hip fracture	• Yes • No	10 (1.0%)
	40	HeadTrauma	Subject's history of head trauma	• Yes • No	8 (≈0%)
	41	DevelopmentalDisabilities	Subject's history of developmental disabilities	• Yes • No	0
	42	HighBloodPressure	Subject's history of high blood pressure	• Yes • No	11 (1.1%)
Blood Test	43	HB	Blood test analysis of the amount of haemoglobin in the blood (g/L)	numeric	19 (1.9%)
	44	CRP	Blood test analysis of the amount of C-reactive protein in the blood (mg/L)	numeric	30 (3.0%)
Physical Examination	45	Pain	"Have you had pain in the last 4 weeks?"	• Yes • No	53 (5.4%)
	46	HeartRate_Sitting	Subject's heart rate in beats per minute, while sitting	numeric	16 (1.6%)
	47	HeartRate_Lying	Subject's heart rate in beats per minute, while lying	numeric	25 (2.5%)
	48	BloodPressure_Right	Subject's systolic blood pressure measured on the right arm, while lying (mmHg).	numeric	21 (2.1%)
	49	HandStrength_Right	Subject's right hand strength in Newtons, during an interval of 10s, measured by the Grippit instrument.	numeric	85 (8.6%)
	50	HandStrength_Left	Subject's left hand strength in Newtons, during an interval of 10s, measured by the Grippit instrument.	numeric	88 (8.9%)
	51	Rise_Safe	"Does it feel "safe" for you to rise from a chair without using your arms?"	• Yes • No, it feels unsafe • Cannot stand up	63 (6.4%)
	52	Rise_How	Rising from the chair. How?	• Got up without using their arms	67 (6.8%)

				• Got up, but used their arms • Tried but couldn't • Not tried for security reasons • Not tried as there was no suitable chair • On a wheelchair	
	53	WeightLoss_3months	Any weight loss during the last 3 months?	• Yes, more than 3kg • Don't know • Yes, more than 1 kg, but less than 3 kg • No weight loss	16 (1.6%)
	54	StandingTest_Right	Single leg standing with right leg. Best value in seconds of three tries.	numeric	72 (7.3%)
	55	StandingTest_Left	Single leg standing with left leg. Best value in seconds of three tries.	numeric	72 (7.3%)
	56	Dental_Prothesis	Assessment via x-ray of the subject's jaws in regards to their own teeth and prosthesis.	• Only own teeth • Own teeth and removable dentures • Own teeth as well as removable prosthesis in one tooth jaw, or toothless and whole prosthesis in one tooth jaw • Completely toothless • Completely toothless and complete denture in one or both jaws • With implants	4 (≈0%)

	57	Dental_TeethNumber	Assessment via x-ray of the subject's jaws in regards to the number of own teeth.	numeric	96 (9.8%)
Psychological	58	MemoryLoss	Assessment of the subjects' memory in daily life situations. See "Description 7".	numeric	91 (9.3%)
	59	MemoryDecline	"Do you think your memory has gotten worse?"	• No • Somewhat • A lot	10 (1.0%)
	60	MemoryDecline2	"Does anyone in your circle think that your memory has gotten worse?"	• Yes • No	18 (1.8%)
	61	AbstractThinking	"Explain the following phrase: 'The apple does not fall far from the tree'"	• Wrong answer • Wrong, only concrete answer • Wrong abstract answer • Right answer	46 (4.7%)
	62	PersonalityChange	Assesses if the subject experienced changes regarding personality traits. See "Description 8"	• Yes • No	19 (1.9%)
	63	Identity	First, the subject is asked questions about their identity: first name, last name, year of birth, date of birth and age. Then, the S_Psychological_Identity index is calculated as the sum of correct answers given by the subject.	numeric	6 (≈0%)
Health Instruments	64	SOC	Sense of Coherence [1]: assesses the subject's comprehensibility (how they perceive events as making logical sense), manageability (how they feel they can cope with situations), and meaningfulness (how they feel that life makes sense	numeric	56 (5.7%)

			and challenges are worthy overcomming).		
	65	DigitSpan_Forward	Forward Digit Span Test [2]: The tester say to the subject a sequence of numbers and the subject has to repeat it in the way they hear it. It starts with two 3-number sequence, going up to two 9-number sequence. Each correct sequence said by the subject counts as 1 point.	numeric	12 (1.2%)
	66	DigitSpan_Backwards	Backward Digit Span Test [2]: The tester say to the subject a sequence of numbers and the subject has to repeat it backwards. It starts with two 3-number sequence, going up to two 9-number sequence. Each correct sequence said by the subject counts as 1 point.	numeric	14 (1.4%)
	67	Livingston	Livingston Index [3]: a sleep disorder scale, composed of eight items regarding difficulty falling asleep or staying asleep, sleep medication usage, sleep interrupted at night, moods or tension, difficulty sleeping owing to pain or itching, inability to return to sleep after walking at night, waking up too early, or feeling tired more than two hours a day.	• No sleeping problems • Presence of sleeping problems	31 (3.2%)
	68	EQ5D	EuroQol (EQ-5D) Index [4]: a generic instrument (non-disease specific) that aims to assess physical, mental and social functioning. The instrument is filled	• High quality of life • Low quality of life	77 (7.9%)

			by the subject who describes their own health-related quality of life in regards to mobility, self-care, usual activities, pain/discomfort and anxiety/distress. We used a dichotomised version of EQ-5D by the lower quartile index values.		
	69	Index_Katz	Katz Index of Independence in Activities of Daily Living (ADL) [5]: Assesses functional status of the subject in regards to their capacity to perform activities of daily living independently.	• Severe impairment • Moderate impairment • Full function	17 (1.7%)
	70	IADL	Lawton Instrumental Activities of Daily Living (IADL) [6]: Assesses independent living skills of the subjects. Considers more complex skills than the ADL index.	• Dependent • Independent	21 (2.1%)
	71	MMSE	Mini-Mental State Examination (MMSE) [7]: Assesses the cognitive aspects of mental functions.	numeric	5 (≈0%)
	72	ClockTest_Sum	Clock Drawing test [8]: Assesses cognitive impairment of a subject, in regards to verbal understanding, memory, spatially coded knowledge and construction skills. The 10-point score version was used in this study.	numeric	37 (3.7%)
	73	MCS12	Dichotomised Mental Composite Score of the SF-12 Health Survey [9]. The SF-12 is composed of 12 weighted questions that assess mental and physical functioning, and health-related quality of life. The Mental Composite Score is	• Low level of health • High Level of health	98 (10.0%)

			calculated from the designated questions and compared to the age-specific mean. An age-specific mean difference score of -5.5 points indicates a low level of health.		
	74	PCS12	Dichotomised Physical Composite Score of the SF-12 Health Survey [9]. The SF-12 is composed of 12 weighted questions that assess mental and physical functioning, and health-related quality of life. The Physical Composite Score is calculated from the designated questions and compared to the age-specific mean. An age-specific mean difference score of -5.5 points indicates a low level of health.	• Low level of health • High level of health	98 (10.0%)
	75	CPRS	Comprehensive Psychopathological Rating Scale [10]: It assesses the psychiatric state of the subject as to their level of depression.	• Absence of depression • Mild depression • Moderate depression • Severe depression	13 (1.3%)

Description 1: SocialNetwork

This is a categorical index that assesses the subjects' social network into one of the following categories: "Very bad social network", "Bad social network" and "Normal social network". It is built upon the questions described below, whose alternatives are attributed to a value held in parenthesis. The sum score of the values determines the category, as the following: a sum score of 5 characterizes a "Very bad social network"; a sum score in the range of 1 to 4 characterizes a "Bad social network"; a sum score of 0 characterizes a "Normal social network".

Questions:

Do you think your number of friends is enough?

- Too few (1)
- Enough (0)
- Too many (0)

How many people do you think you know well and can talk about most of the time?

- No one (1)
- 1-3 (1)
- 4-6 (0)
- 7-9 (0)
- 10-15 (0)
- 16-30 (0)
- More than 30 (0)

Do you have someone who you feel you can be yourself in, who accepts you with all your merits and flaws?

- Yes, without a doubt (0)
- Yes, probably (0)
- No, probably not (1)
- Not at all (1)

Do you feel close to your family (other than your husband, spouse, partner and children)?

- Missing relatives (1)
- Highly (0)
- To some extent (0)
- Not especially (1)
- Not at all (1)

Description 2: SupportNetwork

This index was built as the sum score of the questions' value regarding the subjects' social support network. The highest the score the worse social support network the subject has. The alternatives for the questions (with their respective values in parenthesis) are the following: Yes, without a doubt (1); Yes, probably (2); No, probably not (3); Not at all (4).

Questions:

- Can you get help from someone or someone in case of illness or other practical problems?
- Do you know someone or someone who can help you to write an official letter or appeal a government decision?
- Do you know that you have someone or someone who can provide you with proper personal support to cope with the stress and the problems of life?

Description 3: Loneliness

This is a categorical index that assesses the subjects' felling of loneliness into one of the following categories: "Very high level of loneliness", "Some level of loneliness" and "Normal level of loneliness". It is built upon the questions described below, whose alternatives are attributed to a value held in parenthesis. The sum score of the values determines the category, as the following: a sum score of 4 characterizes a "Very bad social network"; a sum score in the range of 1 to 3 characterizes a "Some level of loneliness"; a sum score of 0 characterizes a "Normal level of loneliness".

Questions:

- Do you feel lonely?
- Yes, often (1)
- Yes, sometimes (0)
- No, rarely (0)
- No, never (0)

When you look back on the last five years of your life, which of the following options best suits you?

- I have not felt loneliness at any time in the past 5 years (0)
- I have experienced occasional occasions with loneliness (0)
- I have experienced recurrent periods of loneliness (1)
- I have lived with a more or less constant feeling of loneliness (1)

Do you feel a strong affinity with your local community?

- Highly (0)
- To some extent (0)
- Not especially (1)
- Not at all (1)

Are you in a group of friends who have or do something in common?

- Yes (0)
- No (1)

Description 4: SocialActivities

The subjects were asked if they engaged in the following sociocultural activities in the past 12 months from the date they responded the questionnaire: “Cinema, theatre, or concert”; “Restaurant, café, or pub”; “Church or religious meetings”; and “Study circle or course of some kind”. The possible answers were “yes” or “no”. The “SocialActivities” index was built as the number of “yes” answers given by the subject.

Description 5: PhysicallyDemandingActivities

The subjects were asked if they engaged in the following physically demanding activities in the past 12 months from the date they responded the questionnaire: “Gardening”; “Taking walks outside”; “Picking berries or mushrooms”; “Hunting or fishing”; “Knit, weave or sew”; “Painting, drawing or sculpting”; “Home repairs”; and “Repairing cars or other mechanical equipment”. The possible answers were “yes” or “no”. The “PhysicallyDemandingActivities” index was built as the number of “yes” answers given by the subject.

Description 6: LeisureActivities

The subjects were asked if they engaged in the following leisure activities in the past 12 months from the date they responded the questionnaire: “Reading the newspaper”; “Reading magazines”; “Reading books”; “Watching television”; “Playing games or cards”; “Playing musical instruments”; “Listening to music”; and “Using the internet or playing computer games”. The possible answers were “yes” or “no”. The “LeisureActivities” index was built as the number of “yes” answers given by the subject.

Description 7: MemoryLoss

This index was built as the sum score of the questions' value about the subjects' memory in daily life situations. The highest the score means the highest memory decline of the subject. The alternatives for the questions (with their respective values in parenthesis) are the following: Never (1); Rarely (2); Sometimes (3); Often (4); Always (5).

Questions:

- Do you happen to come to the store and have forgotten what to trade?
- Do you have trouble remembering what happened the day before?
- Do you lose or place things?
- Do you find it hard to know where you are?
- Do you find it difficult to find the right home / at department?
- Do you find it difficult to find the store / post office?
- Do you find it difficult to find in a foreign environment?

Description 8: PersonalityChange

The tester asks the subject if they felt like they changed in regards to the items in the questions below. A positive answer to 2 or more items defines a change in personality.

Questions:

- More or less talkative?
- More or less grumpy?
- More or less agitated?
- More or less withdrawn?
- More or less apathetic?
- More or less worried?
- More difficult than before to make decisions?
- More difficult than before to take the initiative

References

[1] Antonovsky, Aaron. "The structure and properties of the sense of coherence scale." Social science & medicine 36.6 (1993): 725-733. NBR 6023

[2] Wechsler, D. (1997). The Wechsler adult intelligence scale-III. San Antonio, TX: Psychological Corporation.

[3] Livingston, G., B. Blizard, and A. Mann. "Does sleep disturbance predict depression in elderly people? A study in inner London." Br J Gen Pract 43.376 (1993): 445-448. NBR 6023

[4] Brooks, Richard, and EuroQol Group. "EuroQol: the current state of play." Health policy 37.1 (1996): 53-72.

[5] Katz, Sidney. "Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living." Journal of the American Geriatrics Society 31.12 (1983): 721-727.

[6] Lawton, M. Powell, and Elaine M. Brody. "Assessment of older people: self-maintaining and instrumental activities of daily living." The gerontologist 9.3_Part_1 (1969): 179-186. NBR 6023

[7] Folstein, Marshal F., Susan E. Folstein, and Paul R. McHugh. "“Mini-mental state”: a practical method for grading the cognitive state of patients for the clinician." Journal of psychiatric research 12.3 (1975): 189-198. NBR 6023

[8] Agrell, Berit, and Ove Dehlin. "The clock-drawing test." Age and ageing 27.3 (1998): 399-403. NBR 6023

[9] Jenkinson, Crispin, and Richard Layte. "Development and testing of the UK SF-12." Journal of health services research & policy 2.1 (1997): 14-18.

[10] Montgomery, Stuart A., and M. A. R. I. E. Åsberg. "A new depression scale designed to be sensitive to change." The British journal of psychiatry 134.4 (1979): 382-389.

ABSTRACT

Healthcare is an important and high cost sector that involves many decision-making tasks based on the analysis of data, from its primary activities up till management itself. A technology that can be useful in an environment as data-intensive as healthcare is machine learning. This thesis investigates the application of machine learning in healthcare contexts as an applied health technology (AHT). AHT refers to application of scientific methods for the development of interventions targeting practical problems related to health and healthcare.

The two research contexts in this thesis regard two pivotal activities in the healthcare systems: diagnosis and prognosis. The diagnosis research context regards the age assessment of the young individuals, which aims to address the drawbacks in the bone age assessment research, investigating new age assessment methods. The prognosis research context regards the prognosis of dementia, which aims to investigate prognostic estimates for older individuals who came to develop the dementia disorder, in a time frame of 10 years. Machine learning applications were shown to be useful in both research contexts.

In the diagnosis research context, study I summarized the state of the art evidence in the area of bone age assessment with the use of machine learning, identifying both automated and non-automated approaches for age assessment. Study II investigated a non-automated approach based on the radiologists' assessment and study III investigated an automated approach based on deep learning. Both studies used magnetic resonance imaging. The results showed that the radiologists' assessment as input was not precise enough for the estimation of age. However, the deep learning method was able to extract more useful features from the images and provided better diagnostic performance for the age assessment.

In the research context of prognosis, study IV conducted a review on the relevant evidence in on the prognosis of dementia with machine learning techniques, identifying a focus on the research on neuroimaging studies dedicated to validating biomarkers for pharmaceutical research. Study V proposed a multifactorial decision tree approach for the prognosis of dementia in older individuals as to their development or not of dementia in 10 years. Achieving consistent performance results, it provided an interpretable prognostic model identifying possible modifiable and non-modifiable risk factors and possible patient subgroups of importance for the dementia research.

BLEKINGE TEKNISKA HÖGSKOLA · BTH ·

ISSN: 1653-2090

ISBN: 978-91-7295-405-2

2020:06