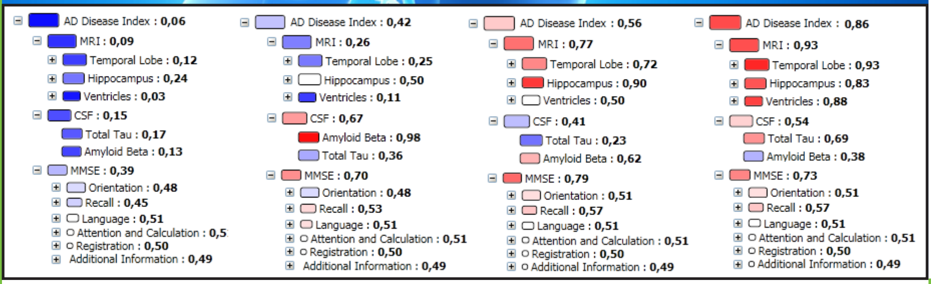
VISIONS • SCIENCE • TECHNOLOGY • RESEARCH HIGHLIGHTS

Dissertation

51



```
⊟ AD Disease Index : 0,06
   ⊟ MRI : 0,09
      ⊞ Temporal Lobe : 0,12
      ⊞ Hippocampus : 0,24
      ⊞ Ventricles : 0,03
   ⊟ CSF : 0,15
        Total Tau : 0,17
        Amyloid Beta : 0,13
   ⊟ MMSE : 0,39
      ⊞ Orientation : 0,48
      ⊞ Recall : 0,45
      ⊞ Language : 0,51
      ⊞ ○ Attention and Calculation : 0,5
      ⊞ ○ Registration : 0,50
      ⊞   Additional Information : 0,49
```

```
⊟ AD Disease Index : 0,42
   ⊟ MRI : 0,26
      ⊞ Temporal Lobe : 0,25
      ⊞ Hippocampus : 0,50
      ⊞ Ventricles : 0,11
   ⊟ CSF : 0,67
        Amyloid Beta : 0,98
        Total Tau : 0,36
   ⊟ MMSE : 0,70
      ⊞ Orientation : 0,48
      ⊞ Recall : 0,53
      ⊞ Language : 0,51
      ⊞ ○ Attention and Calculation : 0,51
      ⊞ ○ Registration : 0,50
      ⊞   Additional Information : 0,49
```

```
⊟ AD Disease Index : 0,56
   ⊟ MRI : 0,77
      ⊞ Temporal Lobe : 0,72
      ⊞ Hippocampus : 0,90
      ⊞ Ventricles : 0,50
   ⊟ CSF : 0,41
        Total Tau : 0,23
        Amyloid Beta : 0,62
   ⊟ MMSE : 0,79
      ⊞ Orientation : 0,51
      ⊞ Recall : 0,57
      ⊞ Language : 0,51
      ⊞ ○ Attention and Calculation : 0,51
      ⊞ ○ Registration : 0,50
      ⊞ ○ Additional Information : 0,49
```

```
⊟ AD Disease Index : 0,86
   ⊟ MRI : 0,93
      ⊞ Temporal Lobe : 0,93
      ⊞ Hippocampus : 0,83
      ⊞ Ventricles : 0,88
   ⊟ CSF : 0,54
        Total Tau : 0,69
        Amyloid Beta : 0,38
   ⊟ MMSE : 0,73
      ⊞ Orientation : 0,51
      ⊞ Recall : 0,57
      ⊞ Language : 0,51
      ⊞ ○ Attention and Calculation : 0,51
      ⊞ ○ Registration : 0,50
      ⊞ ○ Additional Information : 0,49
```

Disease State Index and Disease State Fingerprint

Supervised learning applied to clinical decision support in Alzheimer's disease

Jussi Mattila

VTT SCIENCE 51

Disease State Index and Disease State Fingerprint

Supervised learning applied to clinical decision support in Alzheimer's disease

Jussi Mattila

Thesis for the degree of Doctor of Technology to be presented with due permission for public examination and criticism in Tietotalo Building, Auditorium TB103 at Tampere University of Technology (Tampere, Finland) on the 9th of May 2014, at 12 noon.

ISBN 978-951-38-8119-1 (Soft back ed.)
ISBN 978-951-38-8120-7 (URL: http://www.vtt.fi/publications/index.jsp)

VTT Science 51

ISSN-L 2242-119X
ISSN 2242-119X (Print)
ISSN 2242-1203 (Online)

JULKAISIJA – UTGIVARE – PUBLISHER

VTT
PL 1000 (Tekniikantie 4 A, Espoo)
02044 VTT
Puh. 020 722 111, faksi 020 722 7001

VTT
PB 1000 (Teknikvägen 4 A, Esbo)
FI-02044 VTT
Tfn. +358 20 722 111, telefax +358 20 722 7001

VTT Technical Research Centre of Finland
P.O. Box 1000 (Tekniikantie 4 A, Espoo)
FI-02044 VTT, Finland
Tel. +358 20 722 111, fax +358 20 722 7001

Kopijyvä Oy, Kuopio 2013

Disease State Index and Disease State Fingerprint
Supervised learning applied to clinical decision support in Alzheimer's disease

Taudin tilan indeksi ja taudin tilan sormenjälki. Ohjatun koneoppimisen menetelmä kliinisen päätöksenteon tueksi Alzheimerin taudissa. **Jussi Mattila.**
Espoo 2013. VTT Science 51. 96 p. + app. 75 p.

Abstract

Due to scientific and technological advancements, investigations in modern medicine are producing more measurement data than ever before. Since a large amount of information exists, and it is also being produced at ever-increasing rates, no single person can digest all current knowledge of diseases. Data collected from large patient cohorts may contain valuable knowledge of diseases, which could be useful to clinicians when making diagnoses or choosing treatments. Making use of the large volumes of data in clinical decision-making requires ancillary help from information technologies, but such systems have not yet become widely available. This thesis addresses the challenge by proposing a computer-based decision support method that is suited to clinical use.

This thesis presents the Disease State Index (DSI), a supervised machine learning method intended for the analysis of patient data. The DSI comprehensively compares patient data with previously diagnosed cases with or without a disease. Based on this comparison, the method provides an estimate of the state of disease progression in the patient. Interpreting the DSI is made possible by its visual counterpart, the Disease State Fingerprint (DSF), which allows domain experts to gain a comprehensive view of patient data and the state of the disease at a quick glance. In the design and development of these methods, both performance and applicability in clinical use were taken into account equally.

Alzheimer's disease (AD) is a slowly progressing neurodegenerative disease and one of the largest social and economic burdens in the world today, and it will continue to be so in the future. Studies with large patient cohorts have significantly improved our knowledge of AD during the last decade. This information should be made extensively available at memory clinics to maximize the benefits for diagnostics and treatment of the disease. The DSI and DSF methods proposed in this thesis were studied in the early diagnosis of AD and as a measure of disease progression in six original publications. The methods themselves and their implementation within a clinical decision support system, the PredictAD tool, were quantitatively evaluated with regard to their performance and potential benefits in clinical use. The results show that the methods and clinical decision support tool based on these methods can be used to follow disease progression objectively and provide earlier diagnoses of AD. These, in turn, could improve treatment efficacy due to earlier interventions and make drug trials more efficient by allowing better patient selection.

Keywords supervised learning, data visualization, clinical decision support systems

Taudin tilan indeksi ja taudin tilan sormenjälki
Ohjatun koneoppimisen menetelmä kliinisen päätöksenteon tueksi Alzheimerin taudissa

Disease State Index and Disease State Fingerprint. Supervised learning applied to clinical decision support in Alzheimer's disease. **Jussi Mattila.**
Espoo 2013. VTT Science 51. 96 s. + liitt. 75 s.

Tiivistelmä

Nykyaikaisen lääketieteen tutkimuksissa kerätään uuden teknologian ansiosta enemmän mittaustuloksia kuin koskaan aiemmin. Koska tietoa on paljon ja sitä tuotetaan yhä nopeammin, yksittäisen ihmisen on mahdotonta sisäistää kaikki olemassa oleva ajantasainen tietämys eri taudeista. Suurista potilasjoukoista saadut tulokset saattavat sisältää arvokastakin tietoa, josta olisi apua kliinikoille diagnostiikassa ja hoitotoimenpiteitä päätettäessä. Suurten tietomassojen hyödyntäminen päätöksenteossa vaatii tietotekniikkaa apuvälineeksi, mutta tähän mennessä tehtävään sopivia järjestelmiä ei ole saatu laajamittaiseen käyttöön. Tämä väitöskirja vastaa tähän haasteeseen esittelemällä kliiniseen käyttöön soveltuvan tietokonepohjaisen päätöksenteon tukijärjestelmän.

Tämä väitöskirja esittelee ohjatun koneoppimisen menetelmän nimeltään Disease State Index (DSI, suom. taudin tilan indeksi), jolla potilaiden mittaustuloksia voidaan verrata kattavasti suurissa tietokannoissa oleviin aiemmin diagnosoituihin potilaisiin. Menetelmä antaa vertailun perusteella arvion potilaan taudin tilasta ja sen etenemisestä. DSI:n tulosten tulkintaan kehitettiin visualisointimenetelmä nimeltään Disease State Fingerprint (DSF, suom. taudin tilan sormenjälki), joka mahdollistaa potilaan tietojen ja tulosten nopean mutta kattavan arvioinnin. Menetelmien suunnittelussa ja toteutuksessa otettiin yhtä lailla huomioon tarkkuusvaatimukset kuin niiden soveltuvuus käyttöönottoon klinikoissa.

Alzheimerin tauti (AT) on hitaasti etenevä neurodegeneratiivinen tauti ja yksi maailman vakavista sosiaalisista ja taloudellisista ongelmista nyt ja tulevaisuudessa. Potilaista kerättyjen suurten tietomassojen avulla AT:n kuva on terävöitynyt merkittävästi kymmenen viime vuoden aikana. Tämä tieto olisi hyvä saada laajamittaisesti muistiklinikoiden käyttöön parhaan mahdollisen diagnostiikan ja hoidon varmistamiseksi. Väitöskirjassa esiteltyjen menetelmien soveltuvuutta AT:n varhaiseen diagnostiikkaan sekä taudin seurantaan tutkittiin kuudessa julkaisussa, joissa itse menetelmät sekä niiden toteutus kliinisenä päätöksenteon tukijärjestelmänä, nimeltään PredictAD tool (suom. EnnustaAT-apuväline), arvioitiin kvantitatiivisesti suorituskyvyn ja potentiaalisten hyötyjen suhteen. Tulokset näyttävät, että menetelmillä ja niiden pohjalta kehitetyllä kliinisen päätöksenteon tukityökalulla voidaan seurata potilaan taudin tilan etenemistä objektiivisesti sekä mahdollistaa AT:n varhaisempaa diagnostiikkaa. Näiden voidaan puolestaan odottaa parantavan hoitojen tehoa hoitojen aiemman aloituksen ansiosta sekä auttavan lääkekehityksessä paremmin kohdennetun potilasvalinnan myötä.

Avainsanat supervised learning, data visualization, clinical decision support systems

Preface

The research presented in this thesis was made possible by the support of many people. I would like to express my heartfelt thanks to everyone and take this opportunity to give special recognition to several key individuals.

First and foremost, I am indebted to Jyrki Lötjönen for the opportunities, ideas and supervision during the past few years. I thank him for 'hijacking' me into the medical signal and image processing group at VTT and giving me a chance to work on several interesting topics. He set up and coordinated projects where I did my research and always provided direction for the work as well. His supervision and relentless reviewing of the results were instrumental in reaching my goals.

My thesis supervisor at the Tampere University of Technology, Professor Tapio Elomaa, always supported my ambitions and demonstrated patience, even when nothing seemed to be happening. I thank him also for his extensive input on this thesis and the management of many practical issues at the university.

Among my colleagues at VTT, I am especially grateful to Juha Koikkalainen, who provided a platform I could build my own research upon. Developing new ideas and methods in close collaboration with Jyrki and Juha made the journey very enjoyable. Mark van Gils' ability and availability as a team leader, project manager, statistician extraordinaire and one to share a pint with is deeply appreciated. Also, this thesis would have been much more difficult to put together without the excellent work of Arho Virkki and Hilkka Runtti. Lastly, I want to tip my hat to all of my other colleagues at VTT. You make the workplace atmosphere extremely pleasant and something I always enjoy coming back to.

The research work for this thesis was mostly done within a single EU project, PredictAD. It was an inspiring experience, with incredible people from all over Europe coming together around a common vision. Having access to the clinical expertise of Hilkka Soininen, Sanna-Kaisa Herukka, Yawu Liu, Teemu Paajanen, Anette Hall, Merja Hallikainen and Miguel Ángel Munoz Ruiz at Kuopio University Hospital and Gunhild Waldemar, Anja Simonsen, Anne-Mette Hejl and Kristian Frederiksen of Rigshospitalet in Copenhagen was crucial for this thesis. Daniel Rueckert and Robin Wolz from Imperial College London, Lennart Thurfjell from GE Healthcare and Marcello Massimini and Silvia Casarotto from the University of Milan provided their brain imaging expertise to the group. Matej Oresic from VTT and Roman Zubarev from Karolinska Institutet did groundbreaking metabolomics and proteomics research for the project. Finally, I also want to mention the PredictAD project managers at VTT, Kari Kohtamäki and Kari Antila, whose rock solid performance allowed us to concentrate on the research.

I had the honour of having Professors Martti Juhola and Annalena Venneri as the pre-examiners of my dissertation. Their constructive comments helped improve this thesis significantly. Professor Ron Summers has kindly accepted to act as the opponent in the public defence of this thesis, for which I am very grateful.

On a more personal level, I wish to thank my parents. From an early age, I have felt that I can do anything that I want with my life. Thank you for trusting in me, I know that you are happy with the path I chose. I thank my friends for the support and distractions, always provided in a timely manner as needed. In particular, the regular Champions League nights with Samuel Lahtinen and the cutthroat board game sessions with Valtteri Pirttilä helped along the way. Most importantly of all, I have my unbelievably cool and beautiful wife to thank. Elina, without you this thesis would not exist. You showed me how it is done, supported me, studied with me and gave me good advice every step of the way. And all of this is secondary compared to how wonderful sharing my life with you is. I love you, you are all I ever wanted. My very final thank yous go to the rest of our small family playing at my feet as I'm writing this: Eero, Armas and Iines, two cats and our beautiful baby daughter whose birth was set as the ultimate deadline to writing this thesis. With that said, I will now go down on the floor myself and sing the song "Kirppu ja Härkä".

Tampere, February 2014

Jussi Mattila

Academic dissertation

Supervisor Professor Tapio Elomaa
Department of Mathematics
Tampere University of Technology
P.O. Box 553, FI-33101 Tampere
Finland

Reviewers Professor Martti Juhola
Computer Science
School of Information Sciences
33014 University of Tampere
Tampere
Finland

Professor Annalena Venneri
Department of Neuroscience
University of Sheffield
Royal Hallamshire Hospital
N floor, room N130
Glossop Road
Sheffield
S10 2JF
United Kingdom

Opponent Professor Ron Summers
Sir David Davies Building
Loughborough University
Loughborough
Leicestershire
LE11 3TU
United Kingdom

List of publications

This thesis is based on the following six original publications, which are referred to in the text as I–VI. The publications are copyright of their respective publishers. They are reused and reproduced with kind permission from the publishers.

I Mattila, J., Koikkalainen, J., Virkki, A., Simonsen, A., van Gils, M., Waldemar, G., Soininen, H. & Lötjönen, J. (2011). A disease state fingerprint for evaluation of Alzheimer's disease. *Journal of Alzheimer's Disease*, **27**(1), 163–176.

II Mattila, J., Koikkalainen, J., Virkki, A., van Gils, M. & Lötjönen, J. (2012). Design and application of a generic clinical decision support system for multiscale data. *IEEE Transactions on Biomedical Engineering*, **59**(1), 234–240.

III Runtti, H., Mattila, J., van Gils, M., Koikkalainen, J., Soininen, H. & Lötjönen, J. (2014) Quantitative evaluation of disease progression in a longitudinal mild cognitive impairment cohort. *Journal of Alzheimer's Disease*, **39**(1), 49–61.

IV Mattila, J., Soininen, H., Koikkalainen, J., Rueckert, D., Wolz, R., Waldemar, G. & Lötjönen, J. (2012). Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects. *Journal of Alzheimer's Disease*, **32**(4), 969–979.

V Liu, Y., Mattila, J., Ruiz, M. Á. M., Paajanen, T., Koikkalainen, J., van Gils, M., Herukka, S.-K., Waldemar, G., Lötjönen, J. & Soininen, H. (2013). Predicting AD conversion: comparison between prodromal AD guidelines and computer assisted PredictAD tool. *PLOS ONE*, **8**(2), e55246.

VI Simonsen, A. H., Mattila, J., Hejl, A. M., Frederiksen, K. S., Herukka, S.-K., Hallikainen, M., van Gils, M., Lötjönen, J., Soininen, H. & Waldemar, G. (2012). Application of the PredictAD Software Tool to Predict Progression in Patients with Mild Cognitive Impairment. *Dementia and geriatric cognitive disorders*, **34**(5–6), 344–350.

Author's contributions

Author's contributions to the original publications in this thesis are as follows:

I The author had the main responsibility for designing the algorithms, visualization methods, and the study, with supervision from J. Lötjönen and J. Koikkalainen. Input regarding clinical, statistical and mathematical aspects was provided by the rest of the authors. Main responsibility for data analysis and writing the publication were with the author of this thesis.

II For Publication II, the author was responsible for the design and development of the software components, study design, data analysis, and writing the publication. One illustration and verification of equations were provided by A. Virkki and all authors provided input on the final version.

III The author was responsible for method development with J. Lötjönen and for study design together with J. Lötjönen and H. Runtti. H. Runtti did data analysis and had the main responsibility for writing the publication. The other authors provided input in clinical and statistical issues.

IV The author developed the methods, designed the study, analysed data, and wrote the publication. J. Lötjönen and J. Koikkalainen provided input to method development and study design. Input regarding clinical and imaging issues was provided by the other authors of the publication.

V The author's main responsibilities for this publication were method and software development, and providing data and tools to allow implementation of the study. Main responsibility for the study design, data analysis, and writing the publication were with Y. Liu, with support from the thesis author, H. Soininen, and J. Lötjönen.

VI The study in this publication was designed by the author together with G. Waldemar, H. Soininen, J. Lötjönen, and A. Simonsen. Additionally, the author had the main responsibility for developing the application used in the study, organizing the study, analysing the data, and writing relevant parts of the publication. Local support at the clinical sites was provided by A. Simonsen and S.-K. Herukka. Writing responsibilities were shared equally between the author and A. Simonsen.

Contents

Appendices

List of abbreviations

AD	Alzheimer’s disease
ADAS	Alzheimer’s disease assessment scale
ADNI	Alzheimer’s Disease Neuroimaging Initiative
aMCI	Amnestic MCI
ANN	Artificial neural network
API	Application programming interface
APOE	Apolipoprotein E
AUC	Area under the receiver operating characteristic curve
BBN	Bayesian belief network
CDSS	Clinical decision support system
CSF	Cerebrospinal fluid
DCM	Dilated cardiomyopathy
DLB	Dementia with Lewy bodies
DMSS-R	Dementia Management Support System – Revised
DPS	Disease progression score
DSF	Disease State Fingerprint
DSI	Disease State Index
DSM-IV	Diagnostic and Statistical Manual of Mental Disorders – 4th edition
DSP	Disease state parameter
DT	Decision tree
FDG	Fluorodeoxyglucose

FTD	Frontotemporal dementia
GP	Gaussian process
HCI	Hypometabolic convergence index
HIS	Hospital information system
HL7	Health Level 7
LR	Logistic regression
MCI	Mild cognitive impairment
MMSE	Mini mental state examination
MRI	Magnetic resonance imaging
naMCI	Non-amnestic MCI
PACS	Picture archiving and communication system
PDF	Probability density function
PET	Positron emission tomography
PMCI	Progressive MCI – diagnosis converts from MCI to AD
RF	Random Forest
SMCI	Stable MCI – diagnosis remains as MCI
SVM	Support vector machine
TMT	Trail making test
VaD	Vascular dementia

1. Introduction

Technology has revolutionized the way people work in medical research and in clinical practice. Modern medical devices allow recording of enormous amounts of patient data, which adds to the knowledge of diseases and provides opportunities for better healthcare. Computer-based information systems simplify patient management and enable more efficient use of resources. But is healthcare using the deluge of data produced by modern medicine to maximal benefit? Unfortunately, the answer is a resounding no [Fasano 2013]. Clinicians are often unable to benefit from the knowledge of diseases that exists in large patient data sets, as it is buried within the large volumes of data. Sometimes clinicians are even unable to cope with the plethora of data measured from a single patient when making diagnostic decisions. There is a need for computer-based methods that allow making use of the existing large heterogeneous data sets to provide objective knowledge about the condition of a patient. In other words, there is a need for clinical decision support methods.

Clinical decision support can be defined as: "providing clinicians or patients with computer-generated clinical knowledge and patient-related information, intelligently filtered or presented at appropriate times, to enhance patient care" [Osheroff et al. 2005]. In recent decades, the ever-increasing medical knowledge has grown too large for any individual to master. This has led to suboptimal patient care that is inefficient and potentially dangerous to patients, to a lower level of quality in healthcare practices and to clinical processes that are not well coordinated [Institute of Medicine (US) 2001, Legido-Quigley et al. 2008]. A remedy proposed for these issues is clinical decision support systems (CDSS). CDSS is defined as software that provides clinical decision support at the point of care. With a CDSS, the characteristics of patients are analysed by a computer and, based on the findings, assessments or recommendations are presented to the clinician to simplify the decision-making process [Sim et al. 2001]. The healthcare community has slowly but steadily begun embracing CDSSs. Though not all initiatives succeed, recent studies show that a majority of CDSSs have had a positive impact on practitioner performance and, as a result, research into CDSSs is accelerating [Jaspers et al. 2011].

There are many types of clinical tasks that a CDSS can support [Berner 2007, Greenes 2011]. A CDSS can warn of changes in a patient's condition – e.g., while in a critical care unit or during surgery – or send reminders and warnings for laboratory test results, dosage errors, drug-to-drug interactions, and conflicts with allergies. In complex or rare diseases a CDSS can propose likely diagnoses based on patient data and the system's knowledge base of diseases. Subsequently, a CDSS can formulate treatment suggestions based on guidelines or treatment efficacy models.

The provision of CDSSs should follow the principles of evidence-based medicine, which has been defined as "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients" and also as "the use of mathematical estimates of probability and risk" [Sackett et al. 1996, Donald & Greenhalgh 2000]. The motivation for evidence-based medicine is to prevent unsafe practices that lack empirical support, to reduce unacceptable individual variance in diagnoses and treatments and, ultimately, to increase efficiency, quality and equality of healthcare [Donald & Greenhalgh 2000]. In short, evidence-based medicine is about seeking the best scientific evidence for questions concerning diagnosis and treatment of diseases.

To make use of existing large data sets in an evidence-based manner, supervised learning can be used. This is a field of machine learning in which a function or model is inferred from labelled training data. An early and influential example of supervised learning is the perceptron model by Rosenblatt [1958], a building block of artificial neural networks (ANN). Supervised learning predicts the output for any valid input after having been presented with training examples with known outcomes, i.e. pairs of patient data and corresponding diagnoses. According to Alpaydin [2010], supervised learning is especially useful for prediction, extracting knowledge (data mining), compressing information and detecting outliers. Supervised learning is used routinely in medical research to provide the results needed for evidence-based medicine and has been applied to the study of cancers [Liu et al. 2010, Steinfort et al. 2010, Floares et al. 2011, Bhat et al. 2012], infections [Pillai 2011, Yao et al. 2011], cardiac diseases [Sitar-Taut et al. 2009] and neurodegenerative diseases [Devanand et al. 2008, Hinrichs et al. 2009, Walhovd et al. 2010, Ewers et al. 2012, Kruczyk et al. 2012], among many other diseases.

Though extensively applied in medical research, supervised learning methods have not yet become widely available as CDSSs to enable evidence-based medicine at clinics [Wu et al. 2006, Greenes 2011]. Interpretation of the results provided by supervised learning methods can be challenging for clinicians. Requirements on data can cause difficulties in realistic clinical settings in which information collected from patients is of varying completeness and only a few patients have exactly the same tests administered. Additionally, diagnostic decisions rely not only on comprehensive analysis of patient data but also on patient and caregiver interviews, consideration of confounding factors and drawing on past experience. Thus, including the analysis results from supervised learning methods to the holistic assessment of a patient's condition is very challenging.

Quantification of disease state is a concept that has been gaining traction in the medical research community. It relates to a process in which a method, based on patient data, not only labels the patient as having (or not having) a disease but also estimates the clinical condition of the patient [Escudero et al. 2012, Jedynak et al. 2012]. These methods, based on patient data, comprehensively assess the condition of a patient in relation to a disease that usually has several progression stages. Results of the analysis indicate where, in the continuum of the disease, the patient is at. Disease state quantification methods are promising for use in clinical decision support since they can be linked to the clinical states of the disease, allowing clinicians and researchers to more easily incorporate results from the machine learning methods into their decision-making processes.

Despite all the recent efforts, disease state quantification, supervised learning and CDSSs in general are underused in clinical decision-making. There is a need for novel methods and CDSSs that work with realistic clinical data sets, perform robustly, provide clinical benefits and are widely accepted by the end-uses. When such methods become available, they could launch a new era of evidence-based medicine, improving both the quality and equality of healthcare.

1.1 Objectives

The context of this thesis work was the early diagnostics of Alzheimer’s disease (AD). AD is a neurodegenerative disease that causes irreversible death of brain cells, which accelerates with disease progression [Braak & Braak 2012]. The goal was to provide methods and tools that could be used as a CDSS in a situation in which a clinician has to determine whether a subject with mild memory problems is at an early stage of AD. This is a complex problem that currently causes long delays in the diagnosis of AD. Specifically, the clinical objectives of this thesis work were to:

1) provide a comprehensive and objective data-driven estimate of the patient’s disease state for clinical decision-making and

2) enable earlier and more accurate diagnosis of Alzheimer’s disease.

The primary technical objective of this thesis was to design and implement a supervised learning method and a data visualization method that can fulfil the clinical objectives. This required considering the requirements for such methods and working with heterogeneous and sparse data obtained from various medical data sources. The second objective was to validate the performance of the method, ensuring that it reaches a clinically acceptable level and is suitable for a variety of problems. The final goal was to design and develop a CDSS software tool using these methods and to validate its utility with clinicians. In summary, the technical objectives of this thesis were to:

1) design and implement a supervised learning method and data visualization method for estimating the disease state of a patient,

2) validate that the methods perform at an acceptable level, and

3) develop a software tool using the methods and evaluate its clinical utility.

1.2 Outline of the thesis

This thesis comprises six original publications of the author's research concerning supervised learning methods for clinical decision support. All research was conducted between the years 2008 and 2013, including the design, implementation, evaluation and validation of the methods and the related software tools. The work was partially funded by PredictAD, a research project in the 7th EU Framework (FP7 – 224328) in which a consortium of technical and clinical partners aimed to provide standardized and objective solutions for enabling earlier diagnoses of AD, improved monitoring of treatment efficacy, easier patient selection for drug trials and improved cost-effectiveness of diagnostic protocols.

The research work for this thesis consisted of the development of a supervised learning method called Disease State Index (DSI), the development of an accompanying visualization method called Disease State Fingerprint (DSF) and an implementation of a CDSS intended for early diagnosis of AD, called the PredictAD tool. The aim of the PredictAD tool is to help clinicians form an objective view of the state of AD progression in a patient using the DSI and DSF methods.

Publication I presents the DSI and DSF methods, investigating their characteristics and performance when analysing early clinical symptoms and biomarkers of AD. This publication introduces the concepts behind the methods and justifies the development of new supervised learning and visualization methods for clinical decision support.

Publication II describes a software implementation of the DSI and DSF methods as a generic clinical decision support tool. Reusable libraries and the tool were implemented on a modern software development platform, supporting various use cases, and evaluated with regard to accuracy and performance using several data sets.

Analyses of the temporal dynamics of the DSI and DSF methods, when quantifying the progression of Alzheimer's disease, are provided in Publication III. This paper evaluates whether the longitudinal changes in a patient's disease state are reflected by the DSI and whether differences in varying patient profiles are clearly revealed by the DSI and DSF.

Publication IV presents a decision support methodology – enabled by the DSI method – in which the classification problem is constrained by first setting a clinically meaningful target accuracy that must be reached when predicting future progression to AD. Having defined the target accuracy, the number of patients who could be classified with the target accuracy at an early phase of AD are then determined.

In Publications V and VI, the PredictAD tool, which implements the DSI and DSF methods, is evaluated with clinicians who use the PredictAD tool for decision support. The first study compares the clinical performance of using the tool with the current diagnostic guidelines of AD. The latter study compares the use of the PredictAD tool with the current situation in clinical diagnostics in which no CDSS is available.

The rest of the thesis is organized as follows. Chapter 2 motivates the need for clinical decision support, focusing on CDSSs based on supervised learning. In Chapter 3, an overview of Alzheimer's disease diagnostics in the context of clinical decision support is given. Chapter 4 provides a description of the methods and the computer-based decision support tools developed during this research work. Chapter 5 contains a summary of the publications in this thesis, covering their goals, the methods applied, the main results and conclusions. A discussion of the results and consideration of topics remaining for future research are provided in Chapter 6. Finally, concluding remarks are drawn in Chapter 7.

2. Clinical decision support systems

In a landmark paper by McDonald [1976], it was argued that the amount of information required to practise medicine had become so expansive that no human could provide perfect care unaided, but some ancillary aid, like a computer, was needed. Now, almost forty years later, medical research is producing new findings faster than ever, with technological advancements providing a deluge of data and new information. The challenges that were recognized decades ago have only grown greater, and no human can absorb all the knowledge in medicine, which is also being produced at ever-increasing rates. CDSSs are seen as part of the solution to this problem.

2.1 History

Clinical decision support has a long history dating from the early years of computing. Even though the initial systems in the 1950s were purely mathematical or statistical and did not rely on computational power provided by computers, they laid a foundation for the first computer-based CDSSs of the following decades [Ledley & Lusted 1959, Warner et al. 1961]. The first pioneering computerized CDSS described a system for the differential diagnosis of a large number of diseases based on a number of questions and answers fed to the computer using a stack of cards [Collen et al. 1964]. This system was initially applied to the screening of bronchial asthma. Soon after, research was conducted on more focused systems, each supporting one disease area and enforcing strict data collection protocols [Bleich 1969, De Dombal et al. 1972, Peck et al. 1973, Shortliffe et al. 1975]. Some researchers also began modelling the thinking processes behind diagnostic decisions [Pople et al. 1975, Pauker et al. 1976].

By the 1980s, the availability and performance of computers had increased significantly. Artificial intelligence and machine learning had also become active fields of science, allowing several CDSSs for diagnostics and treatment planning to be developed, some with broader scope than before [Miller et al. 1982, Miller 1986]. One of the best-known systems from this era is DXplain, a differential diagnostic system based on relationships between symptoms and diseases [Barnett et al. 1987]. By this time the evolution of hospital information systems (HIS) was also accelerating, allowing more seamless integration of CDSSs with the clinical workflows [McDonald et al. 1977, Pryor et al. 1983, Sittig et al. 1989]. According to Greenes [2011], the period from approximately 1960 to 1985 was a “long infatuation” phase, when there was great enthusiasm for clinical decision support, many research initiatives and a wealth of new ideas.

After the first successes of the 1980s, CDSS research has expanded considerably, with hundreds of articles describing and evaluating a plethora of systems, and many of them also showing evidence of being beneficial to practitioners [Garg et al. 2005, Kawamoto et al. 2005, Jaspers et al. 2011]. Despite all the activity and the current technology-driven medical environment, it is perhaps surprising to learn that CDSSs have had very limited impact on healthcare outside a handful of mostly academic medical centres and highly integrated service providers [Wu et al. 2006, Greenes 2011]. Especially in the diagnostics of diseases, the common procedure is still to manually evaluate all patient data – relying on thresholds or statements by specialists, if available – and to exclude other possible causes of the symptoms. This leads to a question: why, compared with the amount of active research, are there so few practical applications of CDSSs providing decision support for clinicians making the diagnoses and deciding treatments?

The lack of CDSS deployments in wide practice can be attributed to several issues according to [Greenes 2011]: images and signals produced in medical studies are important to diagnostics, but their automatic and robust quantification is difficult. Methods and tools that work in controlled research settings are not always applicable in realistic clinical settings; it takes a long time to test and approve medical devices and a large scale-up effort to move from an initial implementation to one providing on-going decision support. Healthcare as a field is moving forward at a fast pace but, perplexingly, is also very conservative: new ideas are not always embraced immediately. The development of clinical decision support systems for healthcare should also consider the environment in which the system will be used [Kaplan 2001]. Healthcare systems, organizations and clinical processes are not very well coordinated in general. Data produced by medical examinations are often incomplete and can contain errors [Little et al. 2012]. There are also philosophical and language barriers between engineers and clinicians. If the results are difficult to interpret, clinicians may not take them into account. Even when the tools and methods fit clinical workflows, the lack of standards makes interoperability with existing systems difficult. Simply accessing existing data is problematic. Electronic health records are still a work in progress in several nations and they will continue to be so for years to come. Regarding many of these issues, the situation is constantly improving. But the progress is slow and all-encompassing solutions to the technical, organizational and practical challenges in CDSS deployments are not expected in the near future. All in all, the lack of successful CDSS deployments is a sign that more work is needed to bridge medical research and clinical practice.

2.2 Categorization and standards

CDSSs have been categorized into four distinct architectural groups based on their evolution through the years [Wright & Sittig 2008]:

1) Stand-alone decision support systems, beginning in 1959
2) Integrated systems, beginning in 1967
3) Standards-based systems, beginning in 1989
4) Service models, beginning in 2005.

Systems in categories one (1) and two (2) are the most common, and these categories include most of the historical systems mentioned previously. The fundamental difference between stand-alone and integrated systems is that integrated systems are a part of a larger whole, usually implemented as a component within a HIS. Stand-alone systems on the other hand have limited interactions with systems or services outside the immediate environment of the CDSS.

CDSSs in category three (3) include standards for representing, encoding, storing and sharing knowledge. They strive to overcome some of the disadvantages of proprietary decision support systems, especially those concerning interoperability. For example, Arden Syntax is a language for generating automatic alerts and messages [Hripcsak et al. 1994]. A system named Gello formalizes decision criteria and can be used for providing alerts and reminders or creating guidelines for complex clinical workflows [Sordo et al. 2004]. The healthcare standards body Health Level 7 (HL7) has accepted both of these as standards, and Arden Syntax is also accepted by the American National Standards Institute (ANSI).

Service models, the fourth (4) and most recent category, shares the goal of interoperability with the standards-based systems, but they achieve this goal by standardizing an application programming interface (API) instead of providing interoperable data formats. There are two approaches to implementing CDSS in the service models category, depending on where the API exists. If an API is specified in front of a clinical system, then any CDSS supporting this data access API can query data from the clinical system and use it for analyses. An effort using this approach is the Shareable Active Guideline Environment (SAGE), which standardizes the vocabularies used to access and process medical records. Unfortunately, SAGE severely constrains the types of decision support methods that can be implemented on top of it [Ram et al. 2003]. The alternative option for service models is to have the API in front of the CDSS, allowing clinical systems to push information into the CDSS for analyses. The System for Evidence-Based Advice through Simultaneous Transaction with an Intelligent Agent across a Network (SEBASTIAN) is such a system and is being developed by HL7 as the HL7 Decision Support Service [Kawamoto & Lobach 2005].

2.3 Supervised learning in clinical decision support

Supervised learning is a technique in which a mapping from the inputs to the correct outputs is learned through examples provided by a supervisor [Alpaydin 2010]. Supervised learning has been used extensively in medical research to enable evidence-based medicine, i.e. to produce disease models that allow clinicians to assess risks and benefits of diagnostic tests and treatments (including lack of treatment) based on existing evidence. Well-known methods such as ANNs, decision trees (DT), Bayesian belief networks (BBN) and support vector machines (SVM) have been applied to a great variety of medical problems. Since the field is extremely active, this section provides only an overview of the extent of medical problems to which supervised learning methods have recently been applied with regard to disease prediction and diagnostics. CDSSs and supervised learning methods specific to AD are described in more detail in the next chapter. In addition to supervised learning methods, unsupervised and reinforcement learning methods have been utilized in medical contexts. Unsupervised learning methods differ from supervised learning methods by looking for patterns in unlabelled training data [Alpaydin 2010]. Reinforcement learning methods produce actions to maximize a cumulative reward, unlike the individual input/output pairs presented to supervised learning methods [Alpaydin 2010]. Unsupervised or reinforcement learning methods are not considered in this thesis except for those specifically applied to AD, described in the next chapter.

DTs are among the most popular methods for CDSSs due to their simplicity and understandable rules [Alpaydin 2010]. DTs have been employed, e.g., to diagnosis of cancers [Liu et al. 2010, Steinfort et al. 2010], heart diseases [Sitar-Taut et al. 2009] and cerebrovascular disease [Yeh et al. 2011]. BBNs are a probability-based inference model used for knowledge representation when reasoning under uncertainty [Alpaydin 2010]. They have been applied to a range of medical applications, including treatment prioritization during emergencies [Sadeghi et al. 2006] and early diagnosis of sepsis [Gultepe et al. 2012]. ANNs are especially suited to modelling the complex and fuzzy cognitive processes of making diagnoses [Alpaydin 2010]. The applicability of ANNs has been demonstrated with the diagnosis of cancers [Floares et al. 2011, Bhat et al. 2012], tumours [Săftoiu et al. 2012, Streba et al. 2012] and glaucoma [Andersson et al. 2012]. SVM is a relatively recent classification and regression technique [Cortes & Vapnik 1995]. Today, due to its good performance, SVM is one of the most frequently used algorithms in machine learning. SVMs have been used in clinical decision support in, for example, tuberculosis infection [Pillai 2011], pulmonary infections [Yao et al. 2011] and for predicting brain infarcts [Huang et al. 2011]. Recently, medical research has begun applying Gaussian processes (GP), a stochastic method that was primarily designed to solve regression problems but also allows probabilistic classification. GPs have been used, at least, for the prediction of psychiatric disorders [Mourão-Miranda et al. 2012] and estimation of child mortality rates [Rajaratnam et al. 2010].

In addition to the widely used methods above, a large number of other supervised learning methods and hybrid approaches have been used when dealing with particular problems. For example, there are ensemble methods that combine the decisions of several classifiers trained to solve the same problem. One of the most well-known of these is the Random Forest (RF) classifier, which builds several DTs to solve a single problem and selects the outcome based on votes from all the different DTs [Breiman 2001].

Based on the considerable amount of research on the topic, it is quite obvious that there is huge interest in applying supervised learning to questions in medical research. The publications also often state intentions to make the methods available to healthcare providers as CDSSs, but very few of them are eventually deployed to clinics for validation studies, let alone commercialized as tools for daily practice. So the question remains: why, with all the research on CDSSs using supervised learning, are there so few practical applications of such systems providing evidence-based decision support for clinicians? To improve the situation, this thesis proposes a supervised learning method and a data visualization method that could be made available at clinics with reasonable effort and hopefully be also accepted by end-users.

3. Diagnosis of Alzheimer's disease

AD is the most common reason for dementia, accounting for approximately two-thirds of all 34 million dementia cases worldwide [Geldmacher & Whitehouse 1996, Wimo & Prince 2010]. Dementia is a general term that refers to a group of symptoms indicating problems with memory or cognition. The term *dementia* does not reveal the original reason for the problems and dementia is not a disease: it is the clinical manifestation of a disease. There are many possible causes of dementia, some of which are reversible [Hughes et al. 2011]. AD is a neurodegenerative disease that causes dementia due to irreversible death of brain cells, which also accelerates as the disease progresses [Braak & Braak 2012]. As the primary cause of dementia, AD is a huge economic and social challenge. The accumulated costs of social care, unpaid care and medical bills globally are estimated at 1% of all gross domestic products. The costs are projected to more than triple by 2050 when there will be well over 100 million people with dementia, due to the aging population [Wimo & Prince 2010].

AD develops slowly over several years. Initially, there are no visible symptoms. This asymptomatic phase may last up to a decade or even longer [Morris 2005]. When the first clinical symptoms appear, the disease has already been active for several years. Figure 1 illustrates the onset and slow progression of AD.

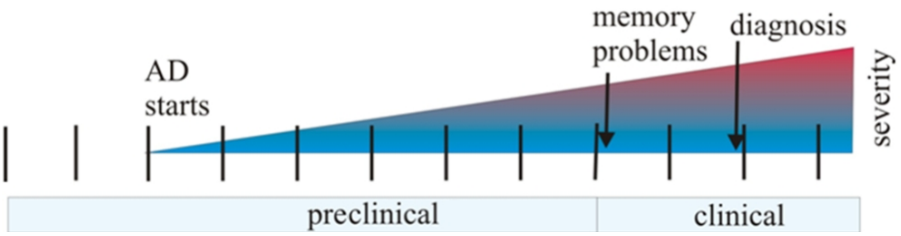


Figure 1. Onset and slow progression of AD.

While the mechanisms of AD are not yet fully understood, the pathophysiological process of AD can be detected *in vivo* already in the preclinical phase from several subtle changes [Jack et al. 2013]. The first measurable change indicating early

AD is the decreased concentration of amyloid β protein in the cerebrospinal fluid (CSF). This is followed by an accumulation of amyloid β in the brain, detectable using positron emission tomography (PET) amyloid imaging. Next, the accumulation of tau proteins in the CSF and brain tissue changes brain metabolism, causing functional and structural neurodegeneration. These are detectable using fluorodeoxyglucose (FDG) PET and structural magnetic resonance imaging (MRI), respectively. Throughout the pathological process, subtle cognitive alterations and impairment occur, and finally the memory and cognition of the person deteriorate towards clinical dementia [Sperling et al. 2011]. Predicting progression to clinical AD from early observations is not simple. Genetic predisposition towards AD, age, gender, education and comorbidities confuse the situation, and some individuals with the pathophysiological process of AD may not become symptomatic during their lifetime at all.

There are two major diagnostic challenges of dementia and AD. One is to differentiate AD from the much more uncommon causes of dementia, e.g. vascular dementia (VaD), dementia with Lewy bodies (DLB), frontotemporal dementia (FTD) and mixed dementia to which multiple diseases contribute. There are also other conditions that cause dementia for which the main symptoms are different – e.g., Parkinson's disease and Huntington's disease – which must be considered before the final diagnosis is given. The second challenge in dementia diagnostics is to determine whether the initial minor memory problems are truly early manifestations of AD, or are they temporary or even part of normal aging. Currently, clinicians rely mostly on their experience and behavioural assessment to dissociate between the different types of dementia. This is done within a framework provided by clinical criteria and national guidelines that allow the fitting of patients' data within agreed standards to reach a diagnosis. Making the diagnosis is a time-consuming exercise and needs to be done manually. Finding the answers to the clinical questions as soon as possible would minimize the delays in current diagnostic work and also improve patient selection in drug trials. In this thesis work, research concentrated on the latter problem: the identification of early AD. The primary reason for concentrating on this issue was the good availability of large data sets containing comprehensive data on early AD development.

3.1 Mild cognitive impairment

A central theme in the early diagnosis of AD – and in this thesis – is to recognize subjects with mild cognitive impairment (MCI) who will develop AD. MCI is a term referring to persons who do not fulfil the criteria for dementia but who exhibit some form of cognitive impairment [Palmer et al. 2003, Petersen 2004]. MCI is associated with an increased risk of developing AD, especially when the MCI is related to memory problems [Petersen et al. 1999]. MCI is a heterogeneous condition that can remain stable (stable MCI, SMCI) or even revert to a cognitively normal state [Petersen & Negash 2008]. Though often caused by early AD or another neurodegenerative disease, MCI can also be caused by a vascular burden, metabolic disturbances, medication interactions, infections, vitamin shortages, malnutrition, drug/alcohol abuse and other physiological and psychological disorders. Thus, identifying MCI patients who will progress to AD (subjects with progressive MCI, PMCI) is not a simple diagnostic problem.

Accurately predicting MCI progression to AD would allow early application of disease-modifying treatments to slow AD progression at a point where the clinical manifestations are still limited. A combination of results from neuropsychological testing, MRI, CSF and genetic testing can aid in the prediction of which patients with MCI will progress to AD [Farlow et al. 2004, Diniz et al. 2008, Landau et al. 2010, Madureira et al. 2010, Lötjönen et al. 2011]. Even a modest delay of one year in the disease onset and progression could drastically reduce the burden of AD on society [Brookmeyer et al. 2007]. The current symptomatic treatments, non-pharmacological interventions and medications for treating AD are most effective at the earliest stages of the disease, underlining the importance of early diagnosis of AD at the MCI phase [Osborn & Saunders 2010]. Currently, there is no medication that would reverse or stop AD progression altogether, but when one is found, it is anticipated to provide greatest benefit if started early.

3.2 Diagnostic criteria of AD

The criteria most commonly used for the diagnosis of AD in clinical practice are the DSM-IV criteria from 1994 (Diagnostic and Statistical Manual of Mental Disorders, 4th edition [Spitzer et al. 2002]) and the NINCDS-ADRDA criteria from 1984 (National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer’s Disease and Related Disorders Association [McKhann et al. 1984]). The clinical diagnoses defined by these criteria are *possible AD* and *probable AD*, a diagnosis of *definite AD* is only available by histopathologic confirmation, meaning, in practice, post-mortem microscopic examination of brain tissue. According to these criteria, diagnosis of *probable AD* is established by clinical and neuropsychological examination. Cognitive impairments have to be progressive and present in two or more areas of cognition. The onset of the deficits must not have occurred at a young age and finally there must be an absence of other diseases capable of producing a dementia syndrome. Currently, diagnosis of AD in Europe takes on average 20 months from the first clinical symptoms, although AD pathology is known to start years before these first symptoms even appear [Cattel et al. 2000, Speechly et al. 2008, Bond et al. 2005].

The criteria mentioned above are still widely used, but they are falling behind current knowledge of AD pathophysiology, which has leaped forward in recent years due to advances in imaging and laboratory technologies. The old criteria also have deficiencies that limit drug development, medical research and clinical practice. For example, they are insensitive in the early phases of AD, when clinical symptoms are not easily detectable [Johnson et al. 2009]. As a result, proposals for new diagnostic criteria of AD and its pre-dementia and preclinical phases (MCI phases) have appeared [Dubois et al. 2007, Dubois et al. 2010, Jack et al. 2011, McKhann et al. 2011, Albert et al. 2011, Sperling et al. 2011]. These put more focus on *biomarkers*, which are defined as measurable physical changes that respond to changes in the progressing disease state. In the new diagnostic criteria of AD, biomarkers are considered an important component alongside neuropsychological tests and clinical assessment. The generally accepted AD biomarkers are as follows:

- low β-amyloid protein levels and/or elevated tau protein levels measured in the CSF,
- atrophy of the temporal lobe revealed by MRI,
- temporo-parietal hypometabolism as assessed with 18-labeled FDG PET or identification of amyloid accumulation in the brain with Pittsburgh Compound B (PIB) PET, and
- known causative genetic mutations in the immediate family, notably the apolipoprotein E (APOE) gene found on chromosome 19.

Though they are important for clinical diagnostics, the sensitivity and specificity of all current biomarkers of AD are rather poor when considered individually [Jack et al. 2012]. Another recent development made possible by novel biomarker acquisition technologies is a hypothetical model of AD progression by Jack et al. [2010, 2013], showing the temporal dynamics of different biomarkers in relation to the clinical disease stages (see Figure 2). This hypothetical model provides a temporal aspect to biomarkers that should also be considered in the diagnostic process.

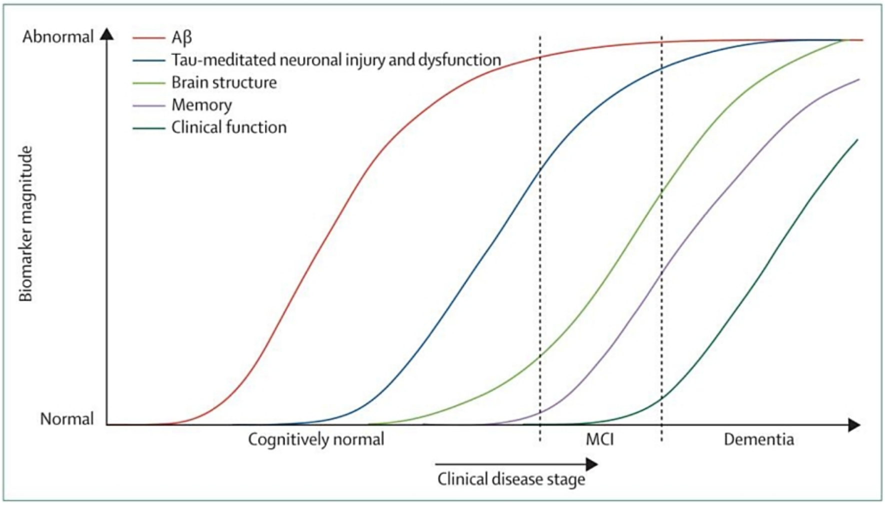


Figure 2. Dynamic biomarkers of the Alzheimer's pathological cascade. This graph illustrates timings of key biomarkers as the subjects' transition through stages of AD (cognitively normal, MCI, dementia). Reproduced from [Jack et al. 2010] with permission from Elsevier © 2010 Elsevier.

Ultimately, making an AD diagnosis could require a clinician to observe hundreds of individual data points from different tests and measurements, read statements from radiology and neuropsychology specialists, interview the patient and caregivers and rule out other possible causes of memory problems. Weighing in each piece of evidence appropriately while also taking into account the latest findings from AD research is an extremely complicated process that requires experience, intuition and tenacity. As a result, the diagnoses come rather late and the overall accuracy of clinical AD diagnosis compared with neuropathological confirmation is relatively low, with the agreement being only 70–90% [Lim et al. 1999, Petrovitch et al. 2001, Kazee et al. 1993].

3.3 Clinical decision support systems in Alzheimer’s disease

Due to the complexity of making an AD diagnosis, it is a problem well suited to clinical decision support. A need for such systems was reflected in a survey of European memory clinics that found that none of the clinics had a CDSS available and that 85% of the respondents wanted a tool that combines all available patient information and provides a risk score for Alzheimer’s disease [unpublished data collected in project PredictAD 2009].

CDSSs targeting diagnosis of AD can be considered to exist in two flavours: *singlemodal* and *multimodal*. Singlemodal CDSSs analyse data from a single measurement modality, e.g. from MRI or PET imaging, genetic testing or neuropsychological testing, and produce a disease probability, estimate of the disease state or a diagnostic suggestion based on the data. Single-modal methods are important for evidence-based medicine and for evaluating the roles of biomarkers in the progression of AD. They can also have a role in clinical decision support although they do not provide a holistic view of the patient’s situation. On the other hand, multimodal CDSSs accept a wide set of measurements and strive to provide comprehensive analysis results, usually with the goal to help in early diagnostics of AD or in differential diagnosis of multiple possible diseases contributing to dementia. Considering that the recently updated diagnostic criteria emphasize observing evidence from several measurement modalities, this thesis focuses on *multimodal* CDSS methods that bear close resemblance to clinical practice.

CDSSs targeting diagnosis of AD can also be divided into two subtypes according to the decision support approach: *expert systems* and *machine learning* methods. Expert systems have a knowledge base built together with domain experts, e.g. with clinicians specialized in neurodegenerative diseases. They include an inference engine that uses the knowledge base to derive suggestions based on new inputs. Machine learning methods, on the other hand, are data-driven and process a training data set to create a model of the disease and then evaluate the model with previously unseen patient data.

DemNet was among the first expert system CDSSs for the diagnosis of dementia. It used a BBN built painstakingly with clinicians to provide classification of patients based on a wide array of neuropsychological test results and demographic data

[Oteniya et al. 2005]. DemNet targeted nurses and general practitioners involved in the primary level assessment of patients suspected of having dementia. PathNet by the same research group extended DemNet to support differential diagnostics of dementia, providing likelihoods of having one of the several possible causes of dementia supported by the system [Oteniya 2008]. The performance of these systems was never evaluated in a clinical setting. The models were reviewed by domain experts during system development and were found reasonable.

The Dementia Management Support System and its revised version (DMSS, DMSS-R) are expert systems that guide the collection of information and allow hypotheses to be made in differential diagnostics of dementia [Lindgren 2008, Lindgren 2011a]. The system is driven by the clinical guidelines of DSM-IV and applies if-then rules to a knowledge base to support diagnostic reasoning. In common cases of dementia, the system provides a categorization of the types of pathologies that meet the guideline criteria, given the evidence presented. In atypical cases, in which the system is unsure of the diagnosis, the system presents all the evidence to the physician. The physician then attempts to infer a diagnosis based on the evidence presented. An evaluation study showed that the DMSS-R system's interpretation of available patient data has good compliance (84%) with the physician's view on the patient case [Lindgren 2011b]. There are also plans to commercialize the DMSS-R system.

An expert system for finding new knowledge and emergent rules to support the diagnostic process of dementia was recently introduced [Sanches et al. 2011, Toro et al. 2012]. This system is described not only as a CDSS but also as a research tool that could help clinicians to determine the most relevant parameters for diagnosis of AD and its cause. Clinical evaluation of the system will be performed over 15 months starting in June 2013 [Sanches et al. 2013].

A hybrid method employing both an expert system and machine learning was employed by Marling and Whitehouse [2001], with the system recommending neuroleptic medication to AD patients based on patient data. Qualitative evaluation of the system was performed with clinical partners. Perhaps the most interesting finding in this study was that the data-driven machine learning module was considered more evidence-based, and therefore more trustworthy, than the expert system module with hand-crafted rules.

A number of machine learning methods have recently been applied to holistic multimodal analysis of data from early AD patients. Duchesne et al. [2010] presented a CDSS for the diagnosis of AD that coded multimodal information measured from patients as binary strings. The strings were used for finding similar patients from a training data set using the *k*-nearest neighbours algorithm and providing a classification based on the findings. Another CDSS applied SVM to classify healthy elderly controls, SMCIs, PMCIs and ADs using data from imaging and CSF biomarkers [Zhang et al. 2011]. Several other methods have also been proposed for combining data from multiple sources to aid in the diagnosis of subjects with MCI [Devanand et al. 2008, Hinrichs et al. 2009, Walhovd et al. 2010, Ewers et al. 2012, Kruczyk et al. 2012]. These are described as decision support tools, although they are perhaps more accurately thought of as data analysis methods that could be applied in decision support. In the end, none of the machine learning-based CDSSs for AD have been evaluated by clinicians to provide evidence of their usefulness.

3.4 Quantification of Alzheimer’s disease state

Traditionally, machine learning has been used for classification with the intention of labelling a patient as having or not having a disease. In situations in which the overall classification accuracy is low (e.g. close to 70% in early prediction of AD), a single label is clinically irrelevant and more knowledge must be extracted from the data. For example, a probability of the given label can be provided, giving additional context to the classification.

With a plethora of measurement values and computational resources available, it has become possible to analyse all patient data comprehensively and provide a rank or a risk score indicating where in the continuum of a disease the patient is, based purely on his/her measurement values. This type of analysis is especially well suited to slowly progressing diseases, in which the subject’s condition gradually deteriorates and is followed over longer periods. AD is such a disease and thus a good candidate for these types of methods. As a result, several methods for quantifying AD progression have been proposed in recent research.

AD is characterized by changes in the brain that can be observed with modern imaging technologies, such as PET and MRI. Converting high-dimensional imaging data into continuous clinical variables is a recurring topic in recent research. Wang et al. [2010] applied a regional feature extraction approach and further feature selection to create a regression model from a training set of MRI data. The goal of the regression is to discover a relationship between brain atrophy patterns and the clinical stage of the disease, the latter being measured by clinical variables. Similar goals were addressed by Fan et al. [2010] who presented a method to estimate clinical variables from brain images by quantitatively evaluating the continuous transition from the normal state to the diseased state. Their research is built on morphological measures derived from structural MRI and a regression method that models several clinical variables that capture the changing disease state.

Chen et al. [2011] introduced a hypometabolic convergence index (HCI) for the assessment of AD and compared it with other biological, cognitive and clinical measures, and demonstrated its capability to predict clinical decline in MCI patients. The HCI is a single measure intended to reflect the extent to which the pattern and magnitude of cerebral hypometabolism in an individual’s PET image corresponds to that in probable AD patients and is generated using a voxel-based image analysis algorithm.

Bioprofiling is an unsupervised machine learning approach for the analysis of biomedical data to support the management and care of patients with AD [Escudero et al. 2012]. Bioprofile analysis derives personal bioindices that indicate how closely a subject’s data resemble the pattern of AD. Bioprofile analysis uses an unsupervised *k*-means technique to cluster measurement variables so that subjects are divisible into those with a bioprofile of AD and those without it. Results show that there is a pattern of AD detectable in the measurement data of patients. The pattern is also in line with the hypothetical model of AD evolution [Jack et al. 2010]. Longitudinal analysis of the changes in bioindices found that having the AD bioprofile at baseline was associated with a risk of progressing from MCI to AD.

A general-purpose statistical methodology for deriving a disease progression score (DPS) using multiple biomarkers from subjects with neurodegenerative disease was proposed by Jedynak et al. [2012]. The methodology yields an Alzheimer's DPS score for each subject and each time point in a data set. In addition, a description of the changes in the biomarkers is produced allowing observation of the temporal ordering of biomarkers. This ordering was noted to follow the hypothetical model of AD evolution [Jack et al. 2010]. In short, the DPS methodology stages individuals according to their state of disease progression and deduces common temporal behaviours of biomarkers in the disease itself.

It is evident that quantification of the disease state and disease progression are emerging as tools to help in the early diagnostics of AD. The latter could also be useful in drug research for monitoring the efficacy of treatments. Though none of the methods above are routinely applied in clinical practice yet, it is expected that some disease state quantification methods will eventually find their way into clinicians' daily workflows.

4. Disease State Index and Disease State Fingerprint

The design and development of the data analysis and visualization methods presented in this thesis began in 2008. The background to the thesis work was in heart disease research, for which Koikkalainen et al. [2008] developed a method for interpreting the results provided by automatic cardiac image processing algorithms. The algorithms processed cardiac images to derive several features of numeric data, containing information useful for assessing a patient's condition, but without additional tools, it was difficult to see which data were truly important and what could be said from all the data put together. To help clinicians interpret the data generated by image quantification, a new supervised learning method and a visualization method were developed.

Parallel to the cardiac imaging research, a European research project called PredictAD[1] was being prepared with the goal to find efficient biomarkers from heterogeneous patient data and integrate them into a clinical decision support tool. The goal of the project was to make early diagnosis and monitoring of the progress of AD more efficient, reliable and objective. It was discovered that the data analysis and visualization concepts developed with cardiac diseases could be developed further to address the new problem at hand. The new goal was to design and develop methods for quantification and visualization of AD patients' disease state based on heterogeneous and sparse measurement data. The new methods were also intended to be used in clinical decision support. The background, requirements and development of these methods are presented in this chapter.

4.1 Background

The foundation for the methods developed in this thesis work was a method called disease state parameter (DSP), which was used for identifying patients with early familial dilated cardiomyopathy (DCM) [Koikkalainen et al. 2008]. In the study, the patient group consisted of 12 subjects who had a genetic mutation that might

[1] www.predictad.eu, verified 29.4.2013

cause DCM but who were all judged to be healthy based on echocardiography. The control group consisted of 14 healthy subjects without the genetic mutation. Volumes, wall thicknesses and wall motions in both the left ventricle and the right ventricle were quantified with automatic MRI processing methods. The DSP method was applied to combine all the MRI parameters into a single global cardiac function index. A visual representation was also created to help assess the individual image parameters, allowing a visual comparison with the disease group (see Figure 3). In this study, it was found that the average DSP of the patient group was significantly higher than that of the control group and that with the DSP method, subclinical familial DCM might be recognized at an early stage.

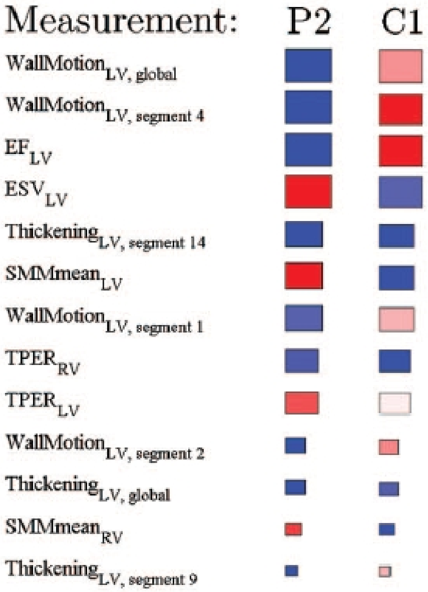


Figure 3. DSP visualizations for a disease case (P2) and a control case (C1). The colours denote larger (red) or smaller (blue) measurement values compared with the mean value of the control group. Nodes are sized according to their statistical significance when comparing control and disease groups. Adapted from [Koikkalainen et al. 2008] with permission from the Radiological Society of North America © 2008 RSNA.

The main limitations of the DSP method are that it expects complete sets of data and that all the features must have similar statistical properties. This was easily achieved in the original study in which image processing algorithms were guaranteed to produce full data sets of appropriately distributed data. In other situations, these requirements can be more difficult to fulfil. Nevertheless, for this thesis work, DSP provided a platform from which a new method could be developed, one that works with more heterogeneous and sparse data sets and that could provide clinical decision support in complex diseases, especially in AD. Three concepts from the DSP method survived the transformation to a more generalized method, even though the implementation details changed considerably. First, the DSP computation introduced a mathematical equation that evaluates similarity between patient measurements and control and positive cases, a so-called *fitness* function. Second, the weighting of fitness values with feature importance was used to derive a global index from all the data. Third, the colours chosen for DSP visualizations (blue, white and red) were used in this thesis work as well.

4.2 Requirements specification process

To reiterate, the reason for developing new data analysis and visualization methods was to support decision-making in the early diagnosis of AD. There was also a need for a method that would allow AD progression to be followed quantitatively and objectively. Specifically, the goal was to develop a method that provides an objective data-driven estimation of the patient's progressing disease state. Discussions with clinicians specialized in dementia were initiated, including evaluations of several successive mock-ups of CDSSs. Based on these sessions, several requirements were identified. Consideration was also given to other medical domains besides AD. The approach to the problem was pragmatic: there was a clear intention to create a machine learning method that could work with realistic clinical data sets and later be deployed as a CDSS at memory clinics. In addition to the requirements specified by the end-users during the design sessions, some requirements were self-imposed. Altogether, requirements for the algorithms were considered in five areas, as categorized by Han and Kamber [2011]:

- **Accuracy:** Basic classifier performance characteristics such as area under the receiver operating characteristic curve (AUC), accuracy, sensitivity and specificity, compared with state-of-the-art classification methods.
- **Speed:** Computational costs of using the given method.
- **Robustness:** Ability of the classifier to make predictions given noisy data or data with missing values.
- **Scalability:** Ability to construct and use the classifier efficiently given large amounts of data.
- **Interpretability:** Level of understanding and insight that is provided by the classifier or predictor.

4.2.1 Accuracy requirements

For accuracy, the requirement was to be comparable to modern classification methods that have been shown to perform well in various settings. The methods commonly used for benchmarking were the naïve Bayesian classifier, SVM, logistic regression (LR) and RF.

4.2.2 Speed requirements

The speed of building the model and evaluating it were both considered to be of great importance. The main reason for this was to support use by clinicians in an interactive CDSS. If new data became available or the clinician wanted to exclude some data from the analysis, recreating and re-evaluating the model should be virtually instantaneous. This would allow exploration of the patient data and hopefully allow clinicians to infer connections between measurements that are otherwise easy to miss.

Another reason for emphasizing speed was the consideration of personalized medicine, which encompasses the use of risk algorithms and biomarkers for improved diagnostics and treatments [Ginsburg & McCarthy 2001]. A computationally low-cost method allows quick construction and evaluation of personalized disease models. With a supervised learning method, personalization could mean, e.g., using as training data only those cases that are of the same age group, gender and/or genotype as the patient being studied. Data analyses could also be personalized by using healthy controls as the reference for building a regression model applied to patient data [Koikkalainen et al. 2012]. For a clinician to be able to apply such methods interactively in a decision support tool, building of the disease models and evaluating them with patient data should be very quick.

Ultimately, the speed requirement was defined such that both building and evaluating the model must be possible without perceivable delay while a clinician is using the CDSS that implements the method.

4.2.3 Robustness requirements

It was mentioned earlier that machine learning methods are currently underused in clinical settings. One reason for this is the nature of clinical data collection, which can be haphazard and sporadic. Clinical data can also have errors and missing data, even in well-controlled clinical trials [Little et al. 2012]. These issues create significant problems for many machine learning methods. Methods may have strict requirements for the inputs they accept and most require some data preprocessing before they can be applied. According to Han and Kamber [2011], preprocessing steps commonly required by machine learning methods include:

- data cleaning – detecting and correcting or removing corrupt or inaccurate records and/or replacing missing data with substituted values,
- data integration – the merging of data from multiple data stores,
- data reduction – deriving a set of values used for machine learning from raw source data or removing features that do not contribute to the results, and
- data transformation – converting values from the data format of a source system into the data format of a destination system or normalizing data to a common space.

For the method developed in this thesis work, minimal pre-processing of patient data was an important goal. If possible, patient data should be accepted as is and any issues normally associated with problematic or missing data handled within the algorithm implementation. The method should thus function even with pre-existing data sets that were collected without consideration for supervised classification. It should not mandate adding (e.g. imputation), removing (e.g. feature selection) or normalizing of data as a pre-processing step. These requirements were also supported by the clinicians who indicated that they preferred to base their decisions on realistic data that was actually measured from patients.

Another requirement related to robustness was determinism, i.e. given the same data model and inputs, the results should always be the same. Determinism would make the validation of the method as a clinical decision support tool and approval as a medical device simpler, since the risks associated with using the results for diagnosis would be related only to how good the model and data for the question were, not the stochastic result of a single evaluation of the model.

Finally, small changes to the inputs were expected to result in small changes to the outputs, making the method well suited to monitoring disease progression. This property of a machine learning method is called stability, indicating how the method is perturbed by changes to its inputs. A stable learning algorithm is one for which the prediction does not change much when the training data are modified slightly. Some classifiers can give a very high probability of having AD even when in reality the data cannot predict AD anywhere near that accuracy. In addition, by changing a single patient value slightly, the results could be reversed and indicate a very low probability of having the disease. Unstable outputs like this would limit interest and also make following disease state longitudinally impossible for clinicians.

4.2.4 Scalability requirements

The number of training cases used for building the disease models was expected to be at most some thousands of previously diagnosed patients. As for the number of features, anything between a few features up to thousands of features was expected. The whole range of possibilities should be supported, with the speed requirements defined earlier also fulfilled. Supporting scaling at this level would allow relatively large amounts of data are to be processed interactively. The scaling should also extend to the reporting of analysis results, enabling the clinicians to absorb all the important information, regardless of the scale of the data.

4.2.5 Interpretability requirements

Interpretability is a subjective measure and therefore more difficult to assess. In the discussions with clinicians it became apparent that their ability to interpret the results is as important as the accuracy of the method. Incorporating new information into the diagnostic process would be challenging if the method provided, for example, only a single number indicating the probability of a patient having a disease. Thus, the method was required to provide a comprehensive and objective estimate of a patient's disease state that also corresponded to his/her clinical status.

Another requirement for interpretability was to keep the algorithm understandable to the level at which clinicians were able to verify the results using pen and paper if they wanted. The reasoning for this 'white box' approach was that if clinicians are able to see and understand how the algorithm arrives at its results, they should be more comfortable using this information in their decisions. Obviously, with enough data, manual verification would become inconvenient, but nevertheless it should be possible. This requirement was in clear contrast to many modern classifiers, which process the data as a 'black box' that cannot be easily inspected by humans, especially if they are not machine learning experts.

The final major interpretability requirement was to indicate clearly the influence of diagnostic tests and any raw measurement values on the results. This would allow clinicians to see how much the different tests affect the classification, possibly determine which tests should be performed next and evaluate the results appropriately. Lastly, related to the speed requirement, the inclusion and exclusion of variables was required to be interactively modifiable, allowing exploration of patient data in search of answers to several clinical questions.

4.2.6 Consideration of existing methods

Several existing methods were considered after the requirements became clear. Not surprisingly, quick quantification of the disease state from heterogeneous and sparse data in a deterministic and understandable manner had not been extensively addressed in previous research at the time the work started. Although several promising approaches were found, they invariably fell short in areas of robustness, speed or interpretability.

Well-known classifiers and regression methods were considered first. Ensemble methods like RF and stochastic methods such as GP were the most promising. Ultimately, none of the existing methods fully satisfied all the requirements set for this work: their outputs (probability estimates of having the disease or not) did not always produce values that reflected disease progression. The algorithms were also often overwhelmingly complex to clinicians. Some research using these methods have since been done and they appear to be reasonably good solutions for assessing disease progression [Young et al. 2012, Chincarini et al. 2011]. The other recently introduced disease state quantification methods mentioned in Section 3.4 were published in parallel with this thesis work and thus were not available for consideration when the work started. The method proposed in this thesis work is compared with the other disease quantification methods in more detail in Section 6.2.

4.3 Disease State Index

Research and development work for this thesis was done using the commercial software package MATLAB[2]. This work resulted in the supervised learning method DSI and an associated visualization method DSF, described later in this chapter. These methods are the main topic of this thesis. Results from evaluating the methods with various data sets are provided in the next chapter in which the thesis publications are summarized.

In short, the DSI is a supervised learning method that processes heterogeneous patient data to derive numeric index values denoting the *disease state* of a patient. Disease state can be considered a condition related to the progression of a disease based on data measured from a patient. DSI is the numeric quantification of disease state, obtaining values between [0, 1]. In practice, the DSI is computed by comparing a patient's measurement values comprehensively with previously diagnosed subjects with and without a disease. Previously diagnosed patients are provided as training data for the method, containing examples of control (healthy) cases and positive (disease) cases. The numeric values resulting from evaluating DSI, i.e. disease indices or DSI values, are defined as the location or rank of the patient between the control and positive cases. They denote the similarity of patient data to the positive cases in the training data. Thus, increasing DSI values indicate greater similarity to patients having the disease, based on the comparison with the training data. The following sections describe in detail how the DSI is computed.

4.3.1 Supervised learning, classification and regression

In machine learning, we can assume that a model of a system is defined with a set of parameters:

$$y = g(x|\theta), \tag{1}$$

where *g(·)* is the model, θ are its parameters, *x* is the input and *y* is the output. In supervised learning, the parameters in θ are optimized by observing training data and minimizing errors in the mapping between training data inputs to the correct outputs. Regression is defined as supervised learning where the correct outputs are numeric values [Alpaydin 2010]. Thus the expected output, *y*, can be a number in the case of regression or a class code (e.g., healthy/disease) in the case of classification. The inputs and outputs can both be multidimensional.

The DSI is a supervised learning method, since it receives pairs of inputs and the correct output classes as training data from a supervisor and from these learns a mapping from the inputs to the output values. It is slightly unorthodox in the

[2] MATLAB Release 7.6 and newer, The MathWorks, Inc., Natick, Massachusetts, USA

sense that it lies somewhere between regression and classification. In the training data, there are no numeric output values that would guide the building of a regression model. However, since DSI results in numeric output – disease indices indicating the progression state of the disease – it implicitly defines a regression model based on the labelled training data. The DSI can be used as a classifier if the numeric disease index is translated into a class label such as ‘healthy’ or ‘disease’. Thus, the DSI is comparable to classification methods that provide a numeric class probability that determines the output class.

4.3.2 Assumptions

The DSI method assumes that values measured for disease diagnostics and used for classification adhere to a distribution that changes in a certain way as the disease progresses. For example, the volume of hippocampus is known to decrease with AD progression and the β-amyloid load in the brain tissue increases with disease progression [Jack et al. 2013]. This allows modelling of the progressing disease state based on the differences between the control (healthy) population and the positive (disease) population in the training data. Figure 4 shows example distributions of data drawn from such populations, representing a situation in which the positive population obtains larger values than the control population. The implication of this assumption is that when the DSI assesses similarity to the disease population, a change in a measured value in a certain direction always changes the output in the same direction, i.e. the DSI is a monotonic function.

The assumption of measures changing monotonically due to a disease does not hold in a situation in which a feature has a normal range for controls but yields increased or decreased values for the positive population (or vice versa). For example, some blood tests and the amount of sleep per day may function like this. The assumption established above requires these features to be split into two distinct features, one modelling the elevation in a positive distribution and another modelling the decrease. Within the data sets used in the thesis work, no such features were encountered. Nevertheless, splitting a feature for analysis is not optimal and finding a better solution is one of the directions for future work, as discussed in Chapter 6.

Situations in which value distributions remain relatively constant with disease progression do not present a problem for the method. These features simply get ignored by design, as described in Section 4.3.4.

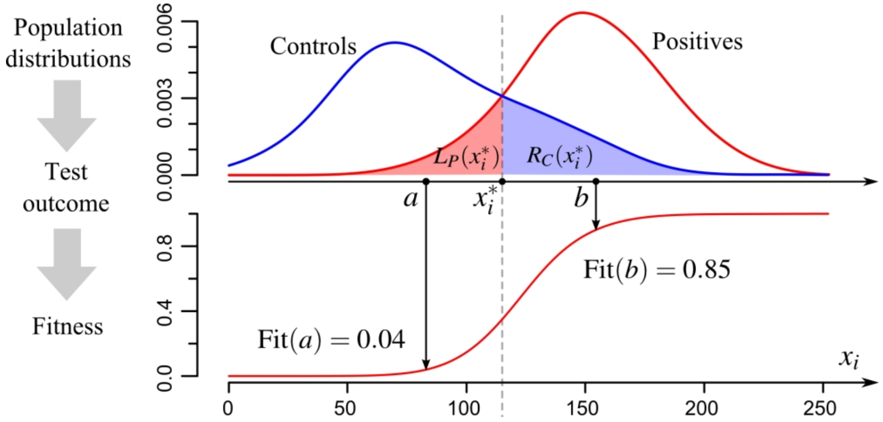


Figure 4. Examples of control and positive populations and the evaluation of the resulting fitness function at locations *a* and *b*. Reproduced from Publication II with permission from the Institute of Electrical and Electronics Engineers © 2011 IEEE.

4.3.3 Fitness

The first step in calculating the DSI is to derive a *fitness* function for each individual feature included in the model. Given patient values x_1, x_2, $\cdots$, x_n, the fitness function provides a non-linear transformation of the value x_i into the fitness space with range [0, 1], based on the differences between the control and positive distributions. For any features that can be represented on a numeric scale, the fitness function $Fit(x_i)$ is defined as:

$$Fit(x_i) := \frac{L_P(x_i)}{L_P(x_i) + R_C(x_i)}, \tag{2}$$

where $L_P(x_i)$ is the left integral of the probability density function (PDF) for positive class values P_i, and $R_C(x_i)$ is the right integral of PDF for control class values C_i, as shown in Figure 4. The derivation of the fitness function can be conducted in an analogous manner if populations are interchanged, i.e. the positive distribution has smaller values than the control distribution. The order of populations is determined by comparing the medians and means of the distributions. In practice, computation of the fitness function in Equation 2 is not done with an estimated PDF but in a discrete manner using the original raw training data values. All values for a feature found in the training set are evaluated sequentially, setting each value in turn as a decision threshold x_i^*, one of which is shown in Figure 4. Now, it is possible to replace the division of integrals in Equation 2 with the fraction of rejection errors (false negatives) from all the errors (both false negative and false positive), written as:

$$Fit(x_i^*) := \frac{L_P(x_i^*)}{L_P(x_i^*) + R_C(x_i^*)} = \frac{FNR(x_i^*)}{FNR(x_i^*) + FPR(x_i^*)}, \quad (3)$$

where FNR and FPR refer to false negative and false positive rates, i.e. the ratios of incorrectly classified instances when x_i^* is used as the decision threshold. As can be inferred from Equation 3, a variance in population sizes does not affect the fitness function: equal PDFs result in the same fitness function regardless of the number of samples in the distributions. After establishing the discrete fitness function using the available training data values, evaluating fitness for any patient value x is possible by interpolating between the discrete instances. Figure 4 shows the result of evaluating fitness at two points, *a* and *b*.

The fitness function in Equation 3 is monotonic (proven in the supplementary material for Publication I) according to the assumption made in Section 4.3.2. The main benefit of monotonicity is illustrated in Figure 5, where the results from evaluating the fitness function are compared with conditional probability. With data sampled from real patient populations, other machine learning methods may indicate that hippocampal atrophy (cell death) could in some cases be a positive change, which does not make clinical sense. This cannot happen with the fitness function and thus the assumption made in Section 4.3.2 is often appropriate for clinical decision support.

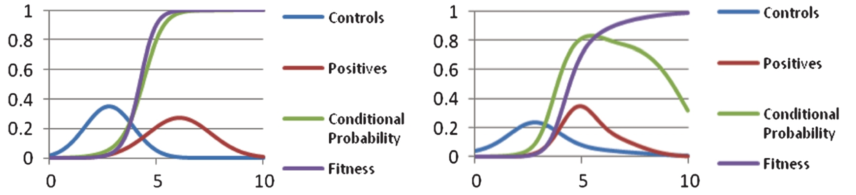


Figure 5. Conditional probability and fitness computed with two distributions of control and positive cases. On the left, well-behaving distributions produce monotonously increasing curves with both methods. On the right, limited amount of data with long tails and possibly some outliers cause drastic changes to the behaviour of conditional probability. Reprinted from Publication I with permission from IOS Press © 2011 IOS Press.

For a purely categorical variable $x_i \in \{\Omega_1, \cdots, \Omega_n\}$ where the categories cannot be represented on a scale, conditional probability of the subject belonging to the positive population when observing $\Omega = x_i$ can be used as the fitness value. In this case the lack of monotonicity does not matter as the categories are assumed to be independent of each other.

Figure 6 shows an example in which we want to evaluate the fitness for a patient measurement value. All controls (blue happy faces) and positives (red neutral faces) in the training data have been positioned along the X-axis according to a measurement of cerebral atrophy. The patient studied is indicated by the black face. Computation of fitness is illustrated in Figure 7. Four (4) cases out of 40 in the control group were found to have more atrophy than our patient, i.e. FPR = 4/40 = 0.10. In addition, there are 24 positive cases with less atrophy than our patient out of 42 positive cases in total, resulting in FNR = 24/42 = 0.57. The fitness value computed for the patient is thus $FNR/(FNR+FPR)$ = 0.85, meaning that the patient measurement fits much better to the positive group than to the control group, which can also be confirmed visually in Figure 6.

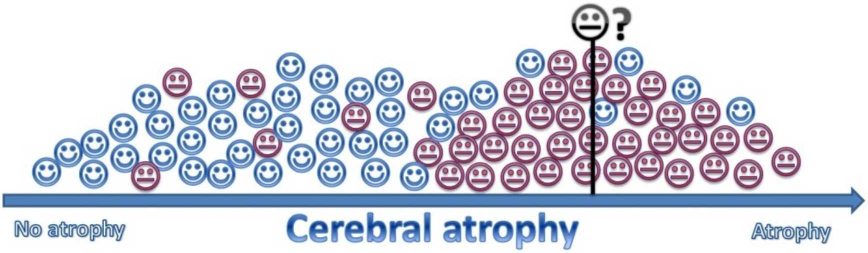


Figure 6. Values for controls (blue) and positives (red) in the training set and a patient whose fitness is to be evaluated.

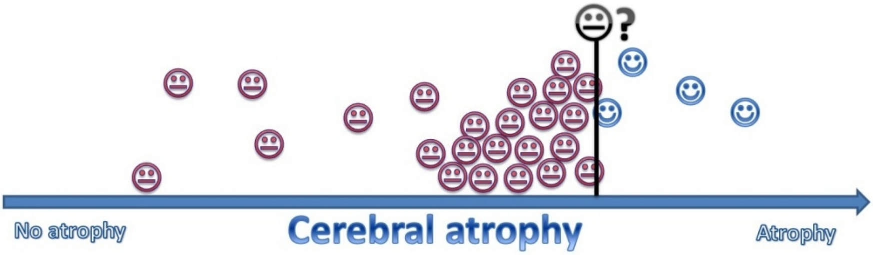


Figure 7. To evaluate fitness, the patient value is chosen as a decision threshold and the ratios of the remaining control and positive cases are computed.

The example above is slightly simplified, since in reality the fitness is evaluated by interpolating between two discretely evaluated fitness values, i.e. interpolating between the fitnesses of the closest values in the training data that are larger and smaller than the patient value.

4.3.4 Relevance

Some features are better for quantifying disease progression than others. This characteristic of a feature is modelled in the DSI algorithm with a measure called *relevance*. Similarly to fitness, the relevance function gives a value between [0, 1]. With relevance, increasing values indicate better discrimination capability. Unlike fitness, the relevance computation does not depend on the patient measurement;

it is based purely on the distributions of control and positive cases in the training data. In short, relevance is defined as a feature's ability to separate the known control and the positive populations. Relevance for scale values that increase with disease progression is defined as:

$$Rel(i) := \max\{0, L_C(x_i^*) + R_P(x_i^*) - 1\}, \tag{4}$$

where $L_C(x_i^*)$ is the left integral of PDF for control values C_i, and $R_P(x_i^*)$ is the right integral of PDF for positive values P_i at the decision threshold x_i^* (shown in Figure 4). The decision threshold x_i^* for the relevance computation is the point at which the fitness function (Equation 2) evaluates to 0.5, but it could also be selected to be the point at which the control and positive PDFs intersect.

To help understand Equation 4, $L_C(x_i^*)$ can be thought of as specificity for training data, i.e. the proportion of patients whose data indicate no disease and who test negative for it. Conversely, $R_P(x_i^*)$ is the sensitivity of training data, i.e. the proportion of positives with data indicating having the disease. As with fitness, relevance can be derived for interchanged population distributions and the actual computation is done using discrete data from the training set. Figure 8 shows examples of relevance computed for three scale features.

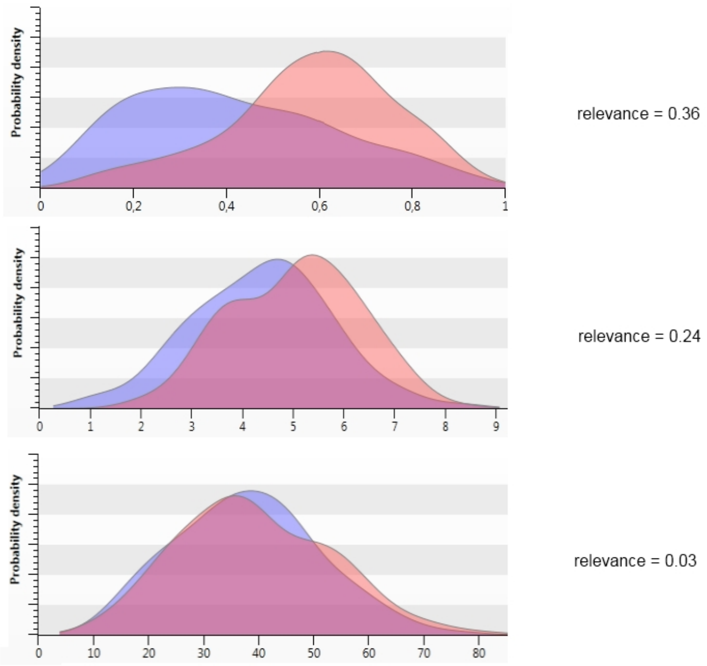


Figure 8. Relevance computed for three variables with different scales (X axis) based on control (blue) and positive (red) distributions in the training data. Increasing relevance indicates better separation of the distributions.

With purely categorical variables for which the categories cannot be ordered on a scale, relevance is computed using only the training cases that are in the same category as the independent variable, i.e. the ones that have the same value as the patient being studied:

$$Rel(i) := \max\{0, Sens(x_i) + Spec(x_i) - 1\}, \tag{5}$$

where $Sens(x_i)$ is the ratio of positive cases with fitness > 0.5 to the total number of positive cases in category $\Omega = x_i$, and $Spec(x_i)$ is the ratio of control cases with fitness < 0.5 to the total number of control cases in category $\Omega = x_i$. Figure 9 shows examples of relevance computed for two categories.

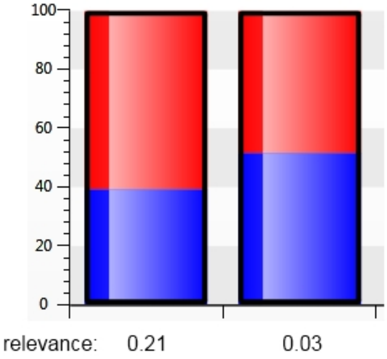


Figure 9. Relevance for two categories of a purely categorical feature using control (blue) and positive (red) distributions from the training data.

Equations 4 and 5 are virtually the same. The only difference is that in Equation 5 training data for only one category within the feature is included.

When the relevance of a feature evaluates to zero, the feature discriminates between the classes as poorly as a random label. A relevance approaching one, i.e. where both sensitivity and specificity for training data are close to one, indicates that the feature is very capable of discriminating between controls and positives and is thus an excellent candidate for estimating the disease state. As with fitness, relevance computation does not depend on the sizes of training populations but is determined purely by the separation of PDFs derived from those populations.

4.3.5 Combining fitness and relevance as the Disease State Index

To study the combination of several variables, the *n*-variable scalar valued Disease State Index function $DSI(x_1, x_2, \ldots, x_n)$ is defined as a weighted mean of fitness and relevance values:

$$DSI(x_1, x_2, \cdots, x_n) := \frac{\sum_{i=1}^{n} Rel(i) Fit(x_i)}{\sum_{i=1}^{n} Rel(i)}. \tag{6}$$

Since both relevance and fitness are in the range [0, 1], it naturally follows that Equation 6 also results in a value in the range [0, 1]. Figure 10 illustrates the evaluation of Equation 6 where the fitness values of three individual patient measurements are weighted by the variable relevancies to obtain a composite DSI value for the three measurements.

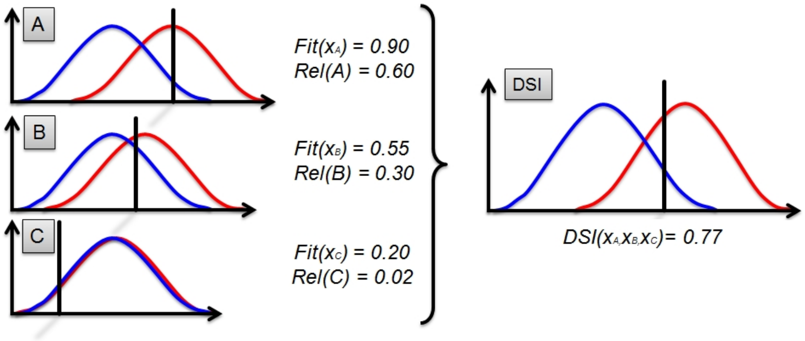


Figure 10. Evaluating DSI as a weighted mean of fitness and relevance values. Black bars denote the location of a measured patient value in relation to the control (blue) and positive (red) distributions.

Conceptually, individual fitness values computed for measurements and their combination as a composite DSI value exist in a common space. This means that the fitness of an individual feature value can be considered as the DSI value for that feature. For example, by computing the fitness for a volume of hippocampus measurement, one can determine the *disease state* of a patient based on that single measurement. Combining this with fitnesses of other measurements produces a composite disease state estimate, which is still in the same space but comprises information from several measurements. The DSI values are assumed to lie on an interval scale, i.e. one unit on the scale represents the same magnitude across the whole range of the scale. As the DSI values are based on fitness, increasing values of the DSI indicate increasing similarity to the positive population and a more severe disease state based on the training data.

4.3.6 Recursion to derive the total Disease State Index

To provide a holistic estimate of the patient disease state, contributions from all the tests and biomarkers used in the analysis must be made available to the domain experts. This is made possible in the DSI method by organizing all the measurement values in a tree hierarchy in which the tests and biomarkers exist in separate branches but combine into a common root (see Figure 11). Currently, organizing features into the trees is an ad-hoc procedure, but a rule of thumb exists: the tree should adhere to a structure that the domain experts find reasonable. This implies that similar tests can be grouped into appropriate categories and also that individual tests can be grouped based on what they measure. For example, all neuropsychological tests could be located under a single node in the tree and a single neuropsychological test could be divided into groups of features, depending on the purpose of each test section producing those features, such as memory, cognition, visuospatial capability, etc. Another example is division of imaging features into categories by the region of the brain they are derived from, an example of which is shown on page 52.

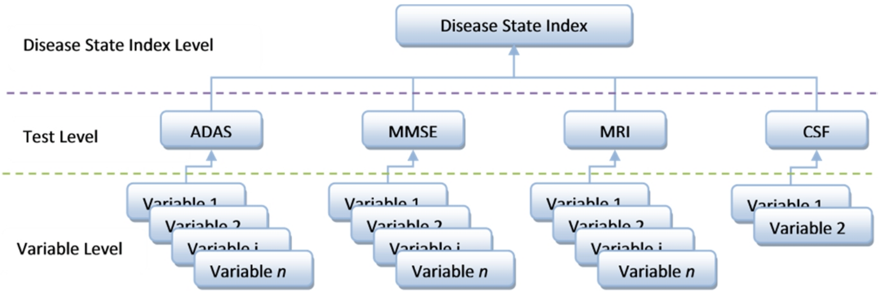


Figure 11. Patient data organized as a tree hierarchy with individual variables (leaf nodes, representing the original raw values measured from the patients), tests (internal nodes) and the total DSI (root node). Additional levels and varying branch depths can be employed to modify the granularity of the tree. Reprinted from Publication I with permission from IOS Press © 2011 IOS Press.

By not combining all patient measurements in one step, the DSI method allows hierarchical study of measurement categories. In other words, Equation 6 is not evaluated once, but instead, recursion is applied to yield a hierarchy of DSI values that the domain experts can read and interpret. The three steps described above – determining fitness (Section 4.3.3), relevance (Section 4.3.4) and combining them as a composite DSI (Section 4.3.5) – are repeated recursively by grouping the data until a single DSI value is arrived at. Recursion starts with the parents of the leaf nodes. DSI values obtained by combining the leaves are then used for evaluating fitness and relevance at the upper levels of the hierarchy. Eventually, the recursion ends at a single DSI value representing the overall quantified disease state of the patient. This is called the total DSI value and is derived from all available patient measurement data. The total DSI is based on the disease model defined by the training data and the tree hierarchy. The computation is fully deterministic, according to the requirement set in Section 4.2.3.

The recursion described above results in a full hierarchy of fitness, relevance and DSI values that represent the disease state of the patient based on particular measurements. The DSI values in the hierarchy indicate how individual measurement values, groups of measurement values and all the patient data, as a whole, match the disease profile as defined from a large number of previously known training cases. Relevance values show how important each piece of data are considered. Together, relevance and DSI values provide a comprehensive and objective quantification of disease state based on patient measurement data.

4.3.7 Summary of the DSI method

The DSI method computes a continuous disease index value by comparing patient measurements with training data that has discrete class labels for control (healthy) and positive (disease) cases. After deciding the tree hierarchy organization, the process of computing the DSI follows these steps:

1) For each patient measurement, compute fitness (Section 4.3.3).
2) For each feature, compute relevance (Section 4.3.4).
3) According to the tree hierarchy organization, combine fitness and relevance values inside each branch as a composite DSI (Section 4.3.5).
4) Using the composite DSIs from step 3 as measurements, continue recursively from step 1 (Section 4.3.6).
5) Stop when all measurements have converged to a single measurement, i.e. the total DSI.

The resulting DSI values can be understood as the percentage of patient measurement data fitting the disease profile, which is modelled by the training data and the tree hierarchy. Relevance can be understood as the importance of individual features and their combinations in measuring disease progression.

The DSI method supports the requirements set in Section 4.2. Every step of the method is simple, both conceptually and computationally, making it quick to evaluate and interpret. It accepts scale and category data and, with additional fitness functions, could be extended to support other types of data. Missing values do not create problems when building the disease model or evaluating it with previously unseen patient data; each feature is first handled alone using the available data and then combined with other data. The result of applying the DSI method is not only a single DSI value characterizing all data but a full hierarchy of fitness, relevance and DSI values that can be read and interpreted by a domain expert.

The correlation between features was considered when developing the DSI method. The tree hierarchy and the recursion resulting from it appear to counter issues normally associated with correlation. For data sets with a large number of features, a step that applies principal component analysis (PCA) to the leaf nodes of the data hierarchy was developed. This did not have a meaningful effect on the method performance with the disease data sets used in this thesis and thus is not considered an integral part of the method. The added benefit of using original raw data as the starting point for the algorithm is that interpretation of the results is simple, since all the raw values used in the computation are ones with which domain experts are already familiar.

4.4 Disease State Fingerprint

In an analogy to human fingerprints and DNA fingerprints, DSF visualizations form unique disease fingerprint patterns, enabling quick visual inspection of the disease state and raw measurement data at several levels of abstraction. In the DSF, patterns emerge from a tree of nodes rendered according to the DSI organization, using shapes and colours to quickly identify the patient's disease state. The DSF is a visual counterpart of the DSI method intended to make reading of the original measurement data and analysis results quick and easy. It allows domain experts to see at a glance how the DSI values were computed and to determine which data are more important than others for the question at hand. The DSF also makes it possible to build a generic data analysis platform for visualizing disease state progression interactively in a CDSS. This section shows – based on Publications I–III – how DSF visualizations are derived from the DSI results.

4.4.1 Colours

The DSF uses a gradient of colours from blue to white to red, indicating increasing DSI values, as shown in Figure 12. The choice of colours produces a heat map, in which cold (blue) colours indicate similarity to healthy controls and hot (red) colours similarity to disease state. Although colours resembling traffic lights were considered, ranging from good (green) to neutral (yellow) to poor (red), their use was avoided due to the difficulties they would create for colour-blind people.

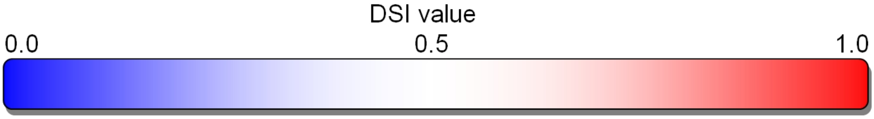
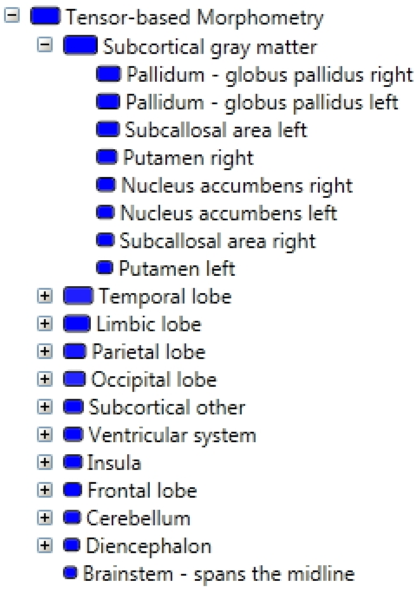


Figure 12. Different DSI values are indicated using colours.

4.4.2 Node sizes

The DSF uses size to indicate relevance. The larger the node, the more relevant it is. To compare relevancies accurately, there needs to be a reference to compare node sizes to, or, more simply, the numeric relevance value can be shown to users when necessary. By default, nodes with a relevance of zero are not shown. A custom threshold could also be selected, hiding nodes that are less relevant than the selected threshold. When observing many features simultaneously, sibling nodes in the tree hierarchy are organized in order of relevance, as shown in Figure 13.

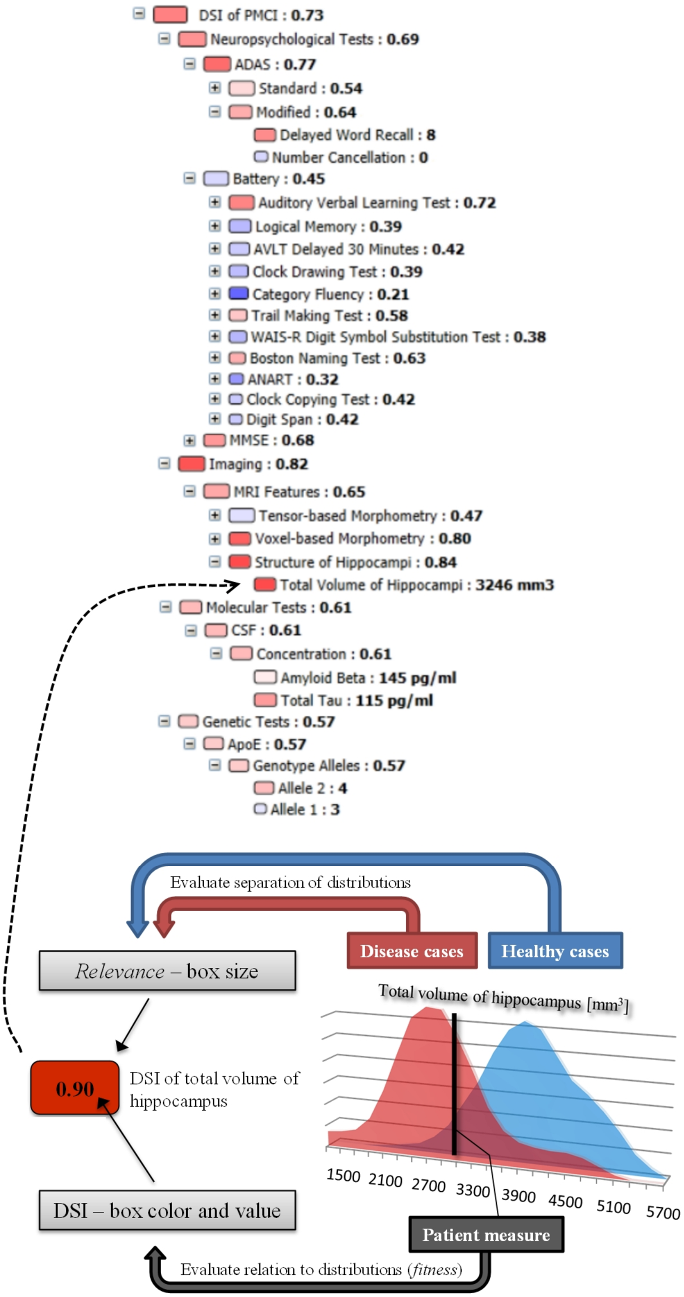


- ⊟ Tensor-based Morphometry
  - ⊟ Subcortical gray matter
    - Pallidum - globus pallidus right
    - Pallidum - globus pallidus left
    - Subcallosal area left
    - Putamen right
    - Nucleus accumbens right
    - Nucleus accumbens left
    - Subcallosal area right
    - Putamen left
  - ⊞ Temporal lobe
  - ⊞ Limbic lobe
  - ⊞ Parietal lobe
  - ⊞ Occipital lobe
  - ⊞ Subcortical other
  - ⊞ Ventricular system
  - ⊞ Insula
  - ⊞ Frontal lobe
  - ⊞ Cerebellum
  - ⊞ Diencephalon
  - Brainstem - spans the midline

Figure 13. Node sizes indicate differences in relevance. Siblings are sorted according to decreasing relevance. The tree shows features from an MRI processing method grouped by the regions of the brain from which they are derived.

4.4.3 Combining nodes in a hierarchy as the DSF

The combination of DSI and relevance values within a tree hierarchy captures the essence of patient data in relation to the studied disease. DSI values rendered as shades of red indicate which patient data are similar to the positive population in the training data, and the size specifies how relevant that information is based on previously diagnosed cases. A large DSI value and high relevance for a neuropsychological test, for example, indicate that the patient performed similarly to a known AD population and that the test discriminates between healthy and AD patients with good accuracy. This is visualized in the DSF as a large red node that is easy to notice. On the other hand, a test with a large DSI value but little or no relevance may often be ignored, since the test is unable to differentiate between the control and positive populations. Accordingly, these kinds of features are very small or even hidden in the DSF visualization. Figure 14 illustrates how individual points of data are combined in the DSF visualization to form a comprehensive picture for evaluating the disease state.



- DSI of PMCI : **0.73**
  - Neuropsychological Tests : **0.69**
    - ADAS : **0.77**
      - Standard : **0.54**
      - Modified : **0.64**
        - Delayed Word Recall : **8**
        - Number Cancellation : **0**
    - Battery : **0.45**
      - Auditory Verbal Learning Test : **0.72**
      - Logical Memory : **0.39**
      - AVLT Delayed 30 Minutes : **0.42**
      - Clock Drawing Test : **0.39**
      - Category Fluency : **0.21**
      - Trail Making Test : **0.58**
      - WAIS-R Digit Symbol Substitution Test : **0.38**
      - Boston Naming Test : **0.63**
      - ANART : **0.32**
      - Clock Copying Test : **0.42**
      - Digit Span : **0.42**
    - MMSE : **0.68**
  - Imaging : **0.82**
    - MRI Features : **0.65**
      - Tensor-based Morphometry : **0.47**
      - Voxel-based Morphometry : **0.80**
      - Structure of Hippocampi : **0.84**
        - Total Volume of Hippocampi : **3246 mm3**
  - Molecular Tests : **0.61**
    - CSF : **0.61**
      - Concentration : **0.61**
        - Amyloid Beta : **145 pg/ml**
        - Total Tau : **115 pg/ml**
  - Genetic Tests : **0.57**
    - ApoE : **0.57**
      - Genotype Alleles : **0.57**
        - Allele 2 : **4**
        - Allele 1 : **3**

Evaluate separation of distributions
Disease cases
Healthy cases
Relevance – box size
0.90
DSI of total volume of hippocampus
DSI – box color and value
Total volume of hippocampus [mm³]
1500
2100
2700
3300
3900
4500
5100
5700
Patient measure
Evaluate relation to distributions (*fitness*)

Figure 14. At the top is a DSF visualization of patient data with a large share of measurement values indicating early AD. The computation of the DSI for an MRI variable is depicted at the bottom. Adapted from Publication I with permission from IOS Press © 2011 IOS Press.

In Figure 14, the names of the tests and their DSI values (or raw measurement values in the case of leaf nodes) are shown next to each node. DSI values are indicated by both colours and numbers, providing an overview of the disease state for the patient from any branch of the tree in relation to the training set. Red coloured nodes with DSI values approaching one indicate similarity to early AD cases. Blue colour and DSI values close to zero indicate similarity to stable MCIs. The relevance of a test is indicated by the size of the node next to the test's name. Not all nodes are fully expanded; collapsed nodes show the overall DSI value from that test section. Here, neuropsychological tests and MRI contribute most to the total disease index, indicated by the largest node sizes and red colour. Nodes in the tree hierarchy can be presented within a software tool such that they are interactively expanded and collapsed. This allows users of the DSF to see an overview of all the data and, when necessary, drill into each individual patient measure. Leaves of the tree show original raw patient data (actual test results), such as 'Delayed Word Recall', which is a task in a neuropsychological test, and 'Total Volume of Hippocampi', derived from structural brain MRI.

4.4.4 Longitudinal DSF visualizations

Data from multiple time points can be rapidly analysed with the DSI method. Feeding the longitudinal results to the DSF produces visualizations with a temporal component. The results of longitudinal DSF visualizations are shown in Figure 15. The left side of the figure shows DSFs in which the DSI values of the individual tests at different time points are shown. The total DSI values (the topmost rows of the DSFs) combine results from all the tests. The size of a box indicates how well a feature discriminates between control and positive cases. Again, colours indicate into which group the data fit best. The right side shows linear regression of the total DSI values (red dashed line with white circles). Black squares present the total DSI values of a patient. A vertical line indicates the age (on the x-axis) of the patient being studied.

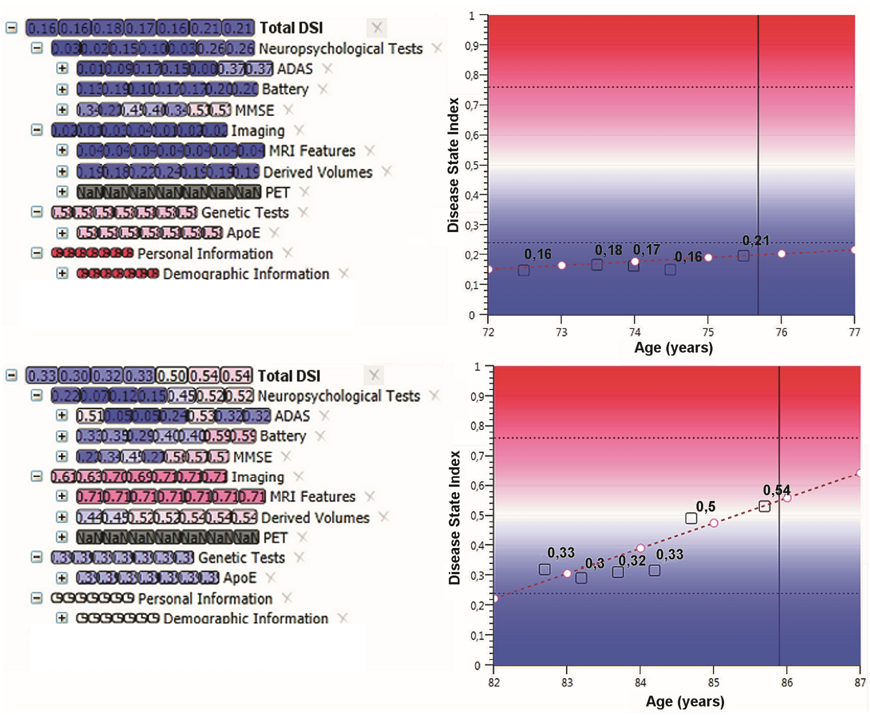


Figure 15. Longitudinal DSF visualizations for two MCI patients. The rows of boxes show the disease state evaluated at approximately 6-month intervals. Reprinted from Publication III with permission from IOS Press © 2014 IOS Press.

4.4.5 Summary of the DSF visualizations

The DSF provides a quickly interpretable visual overview of the patient state, obtained from data-driven and evidence-based analysis of patient data. Using colours and shapes, it draws attention to the data that are most relevant, reducing the need to go over hundreds of data points individually. DSF clearly discloses the factors contributing to the results, highlights the relevant measures and, thus, supports application of clinical judgment. The DSF respects the requirements specified earlier in Section 4.2 and emphasizes interpretability. It supports scalability by allowing a huge number of raw data points to be visualized interactively with only a subset of data visible at any time. It can also provide detailed information of any individual measurement if needed. Longitudinal visualizations allow clinicians to objectively monitor a changing disease state, and they can also be used to visualize the effects of drug treatments on the progression of AD.

4.5 Implementation of the DSI and DSF in the PredictAD tool

A software library implementing the methods and an interactive CDSS using the library, called the PredictAD tool, were developed in parallel with the DSI and DSF methods. The goal of the PredictAD tool was to provide clinical decision support in the early diagnosis of AD using the DSI and DSF methods.

The development of all the software was done in the C# language using Microsoft .NET Framework 3.5 or later [Hejlsberg et al. 2006]. User interface components were implemented with Windows Presentation Foundation (WPF) 3.5 or later. The PredictAD tool and the interactive implementations of the DSI and DSF methods were evaluated with clinical partners several times in an iterative development process. The following sections describe the PredictAD decision support tool, based on Publications II, V and VI.

4.5.1 Software library implementing the DSI and DSF methods

A proprietary software library implementing the DSI and DSF methods provides an application programming interface (API) for managing data, computing DSI values and visualizing the results interactively with the DSF (see Figure 16). It is a standalone library applicable to several domains in addition to early diagnosis of AD.

The library provides an abstraction for data repositories as persistence stores (see (a) in Figure 16) that allow receiving data from multiple data sources. The underlying data source can be virtually anything, a database, a web service or simply a set of data files on a disk. A data definition layer(b) is used for describing entries (the types of tests done on a patient) and features (the types of raw measurement values within entries). Definitions are application-specific meta-data and are configured in source code or by XML (Extensible Markup Language) when initializing the library for use. The organization of the DSI tree hierarchy is also described within this layer. The actual data are read from the persistence stores into another layer(c), where all the entities (e.g. patients), entries and feature values are represented by object instances.

To perform DSI computations, the library needs to know how to select control and positive cases from the training data. For this, a rule-based grouping system(d) was developed. A CDSS using the API is responsible for defining the grouping rules, e.g., “if diagnosis equals AD, assign patient to positive group AD”. After applying grouping rules, control and positive cases in the training data are known to the library(e). Next, using a configurable sampling system(f), the library selects particular data from training cases as the training data for the DSI method. Software tools were created for interactive creation and modification of the grouping and sampling rules systems.

Finally, having sampled the training data(g), the library uses them together with the patient measurements(c) and the feature hierarchy(b) to evaluate the DSI(h). All data are first organized according to the tree hierarchy, and a disease model is trained. Fitnesses, relevancies and DSI values are then computed recursively to obtain a total DSI value from all the available patient data.

To visualize the data and results from applying the DSI method to the user, the library provides graphical user interface components for interactively displaying and manipulating DSI trees[i], data distributions[j], entry timelines[k] and entry details[l]. If a user wishes to examine the DSI or relevance values of any tree node in more detail, clicking on the node provides more information in the form of data distributions. Any test or measurement node can also be omitted from the DSI tree interactively. This can be a useful feature if the user considers certain results unreliable or wants to test different hypotheses.

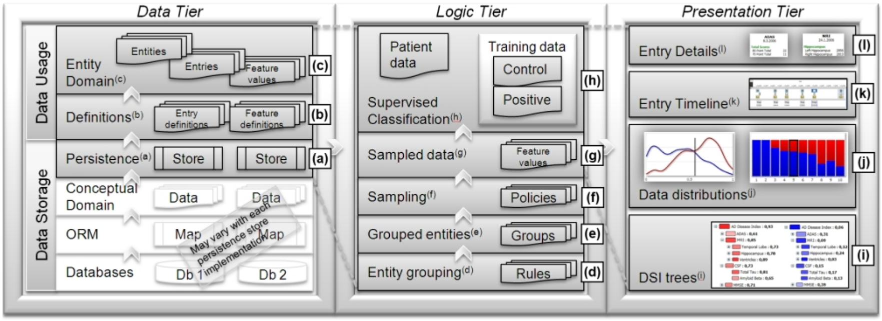


Figure 16. Overview of the architecture of a library implementing the DSI and DSF methods showing the main directions of data flow. Reproduced from Publication II with permission from Institute of Electrical and Electronics Engineers © 2011 IEEE.

4.5.2 PredictAD tool – a CDSS for early diagnosis of AD

The PredictAD tool integrates heterogeneous data such as imaging biomarkers, CSF biomarkers and results from neuropsychological tests for compact visualization within an interactive user interface. The reason for building the PredictAD tool was to investigate whether it – by using the DSI and DSF methods – can assist physicians in the early diagnosis of AD. The hypothesis was that physicians interacting with the software could predict conversion from MCI to AD better than without using the tool. This would allow some patients to be diagnosed earlier than they are currently, making possible earlier delivery of treatments and better selection of subjects in pharmacological trials.

The tool was developed iteratively with clinicians. In the first prototype, basic design and architecture was put in place. In the second prototype, the user interface was improved and new features implemented. The third prototype was used for validation with clinicians. There also exists a more recent prototype version of the PredictAD tool, which will be used in future studies at several memory clinics. See Figure 17 for an overview of the tool evolution over these prototypes.

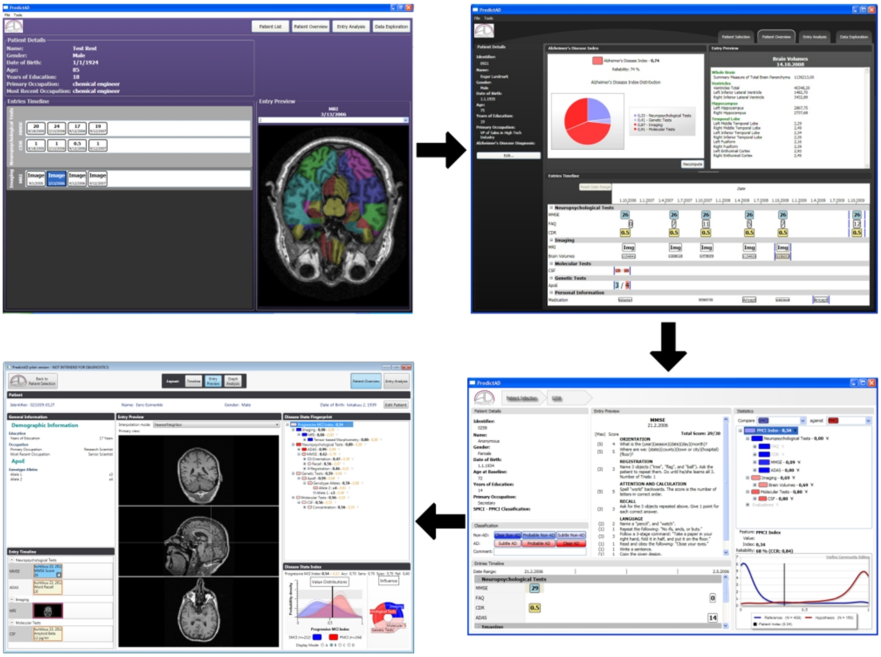


Figure 17. Evolution of the PredictAD tool research prototypes. Starting from top left in clockwise order: prototypes 1, 2 and 3, and the current prototype.

The PredictAD tool provides an overview screen from which all patient data can be easily accessed. The overview screen contains basic demographic information and a timeline of the tests and measurements performed on the patient. In the latest versions, an interactive implementation of the DSF is also visible on the overview screen, showing data analysis results from the DSI. These are provided by the software library implementation described in Section 4.5.1.

4.5.3 Summary of the PredictAD tool

The PredictAD tool was developed as a means to validate the DSI and DSF methods clinically. The implementation of the application was also a central deliverable for the EU-funded PredictAD project. The design and development work to build the PredictAD tool was a software engineering project whose detailed description is outside the scope of this thesis. Nevertheless, the work resulted in a CDSS that could be installed in end-user environments for validation by clinicians, forming a crucial part of this thesis work.

5. Summary of publications

This chapter describes the six original publications on which this thesis is based. The chapter is organized according to the goals of this thesis.

5.1 Design and implementation of the DSI and DSF

The design and implementation of the DSI and DSF methods and the PredictAD tool were published mainly in two articles: Publications I and II. Since the design and implementation have already been covered by the previous chapter, they are not discussed further here.

5.2 Evaluation of the supervised learning method

The evaluation of the DSI method consisted of several studies in which the method was applied to large AD cohorts and publicly available data sets of other diseases. The goal was to validate the DSI as a disease state quantification method and a classifier that performs at a level similar to the current state-of-the-art classification methods. Considering the intended use of the method as a platform for CDSS, these evaluations would also address other requirements listed in Section 4.2, including speed, robustness and interpretability of the results.

5.2.1 Classification performance

The goal of Publications I and II was to describe the methods and validate the DSI method as a supervised classifier using a large AD cohort and several publicly available disease data sets. In these studies, the primary goal was to evaluate the DSI method's ability to discriminate patients with MCI between those who will develop AD and those who will not. In other words, the DSI method was used for predicting conversion from MCI to AD. In addition, Publication II evaluated the performance of the DSI with publicly available hepatitis, heart disease, and diabetes data sets. In all of these studies, the results were benchmarked against reference classifiers.

The studies used cross-validation for the method evaluation. Cross-validation is a technique for assessing how the results of a statistical analysis generalize to an independent data set. One round of cross-validation involves partitioning a sample of data into two complementary subsets. A disease model is built using one subset (training set), and this model is validated on the other subset (test set). To reduce variability, several iterations of cross-validation are performed using different partitions, and the validation results are averaged over the rounds. Both studies used ten iterations of stratified 10-fold cross-validation. Stratified selection means that the ratio of control to positive cases remains the same over all iterations. 10-fold validation means that the training data is divided into ten subsets and each set is used once for testing and nine times for training. The final results of these studies were computed by averaging ten iterations of 10-fold cross-validation, i.e. classifiers were trained and tested 100 times to obtain robust performance estimates.

The data used in these studies were from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [Mueller et al. 2005]. ADNI is a longitudinal five-year study of Alzheimer's disease conducted in the USA and Canada, with the aim of developing and validating surrogate markers for early detection and monitoring of AD progression. ADNI measured the progression of MCI and early AD using biomarkers and clinical and neuropsychological assessment. ADNI recruited approximately 400 people with MCI to be followed for three years, in addition to recruiting 200 normal elderly individuals and 200 AD patients. From the MCI patients recruited to ADNI, the studies in Publications I and II included those whose last clinical diagnosis remained MCI, forming a classification group of SMCIs (n=190), and those whose last clinical diagnosis was AD, forming a group of PMCIs (n=154, average time to make AD diagnosis: 19 months). Using the sparse and heterogeneous baseline measurement data alone, the DSI method's ability to predict conversion to AD was evaluated. The baseline data included neuropsychological tests, magnetic resonance imaging data, molecular test data and genetic test data (see Table 1).

The DSI model of the progressing disease state was shown to discriminate well between different diagnostic classes. The main result in Publication I regarding classification performance was the comparison with the LR, SVM and Bayes classifier when predicting conversion from MCI to AD. The results showed that the DSI method performed at a level similar to that of the reference classifiers. Prediction accuracy for all classifiers was close to 70% and the AUCs using all the data were approximately 75% (see Figure 18).

Table 1. Demographic and clinical data of SMCI and PMCI (average conversion time 19 months from baseline) groups. The data are expressed as counts and percentages of available data except for age and education, which are expressed as the mean (± standard deviation).

	SMCI	**PMCI**
Subjects	190	154
Gender		
Male	125 (66%)	93 (60%)
Female	65 (34%)	61 (40%)
Demographics, years		
Age	74.8 (± 7.6)	74.2 (± 6.9)
Education	15.8 (± 3.1)	15.6 (± 2.9)
Available baseline data		
MMSE	190 (100%)	154 (100%)
ADAS	189 (99%)	152 (99%)
TMT	186 (98%)	153 (99%)
MRI	171 (90%)	135 (88%)
CSF	94 (49%)	83 (54%)
APOE	190 (100%)	154 (100%)

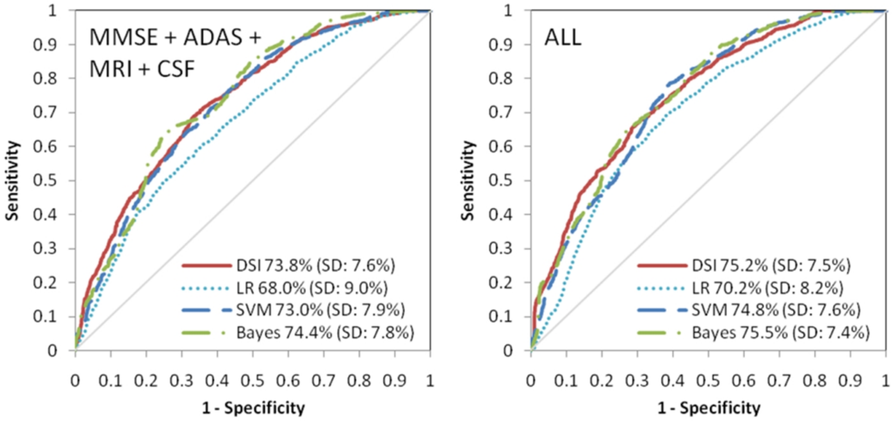


Figure 18. The area under the curve from the receiver operating characteristic curve (AUC) of DSI compared with the LR, SVM and Bayes classifiers when predicting conversion from MCI to AD using a subset of the data and all the data. Reprinted from Publication I with permission from IOS Press © 2011 IOS Press.

Publication II used the same cohort of ADNI patients but slightly different data to study prediction of the conversion from MCI to AD once more. Again, the results shown in Table 2 indicate that the DSI method is able to discriminate between patients as well as the other classifiers.

Table 2. Comparison of DSI classification performance with the reference methods when predicting conversion from MCI to AD. The means and standard deviations (SD) are listed over ten iterations of 10-fold cross-validation.

Method	AUC	Accuracy	Sensitivity	Specificity
DSI	0.75 ± 0.08	0.68 ± 0.08	0.70 ± 0.12	0.66 ± 0.10
SVM	0.75 ± 0.08	0.67 ± 0.07	0.64 ± 0.11	0.69 ± 0.11
Bayes	0.76 ± 0.08	0.67 ± 0.07	0.65 ± 0.12	0.69 ± 0.11
LR	0.69 ± 0.09	0.62 ± 0.07	0.73 ± 0.10	0.53 ± 0.11

In addition to the ADNI MCI data set, the DSI method was evaluated in Publication II using three other medical data sets (Pima Indian Diabetes, Cleveland Heart Disease and Hepatitis) available online[3]. The performance with these data sets was compared with publicly available benchmark results[4]. The DSI method was shown to perform slightly worse than the best method in each case but better than the average (see Table 3). The best benchmark methods differed in all cases and they were optimized individually for each problem, unlike the DSI method, which was simply given the raw data without any pre-processing.

Table 3. Classification accuracy comparison using publicly available data sets and benchmark results. The table shows number of subjects and means and standard deviations (SD) of classification accuracies from ten iterations of 10-fold cross-validation[a], 10-fold cross-validation[b] and averages of methods beating the majority class classifier[c].

Data set	Controls / Positives	DSI[a]	Benchmark maximum[b]	Benchmark average[c]
Diabetes	500/268	0.75±0.04	0.78±0.04	0.74±0.03
Heart disease	164/139	0.81±0.06	0.85±0.01	0.79±0.06
Hepatitis	123/32	0.84±0.08	0.90±0.01	0.85±0.04

[3] Frank, A. and Asuncion, A. (2010). *UCI Machine Learning Repository.* http://archive.ics.uci.edu/ml, verified available 25.4.2013

[4] Duch, W. (2011). Comparison of classification results. http://www.is.umk.pl/projects/datasets.html, verified available 25.4.2013

In summary, the classification accuracies and AUCs achieved by the DSI method compare well with other methods. With regard to predicting MCI to AD conversion, the results correspond with several other studies using similar data sets. Cui et al. [2011] combined MRI, CSF and neuropsychological tests at the baseline to obtain a prediction accuracy of 67.1% and an AUC of 79.6%. Davatzikos et al. [2011] utilized MRI images to construct a Spatial Pattern of Abnormalities for the Recognition of Early AD (SPARE-AD) index that classified correctly 52.3% of 239 MCI patients (AUC 66.0%). A logistic regression model presented by Ewers et al. [2012] achieved an accuracy of 76.3%. Classifying SMCI and PMCI patients with full sets of MRI, PET and CSF measurements using SVM reached an accuracy of 76.4% and an AUC of 80.9% [Zhang et al. 2011].

5.2.2 Computational performance

In Section 4.3, the design of the DSI method was shown to be rather simple. Intuitively, it can be accepted that the computational requirements for analysing patient data with this method are conservative. Publication II used the initial unoptimized implementation of the DSI to evaluate its performance within a CDSS. The training of the DSI model using 344 training cases and computation of DSI values for a patient took on average 860 ms (standard deviation 74 ms). Re-evaluation of DSI values after interactively excluding or including a feature was virtually instantaneous, consistently taking less than one millisecond on a common laptop PC. These computational times show that the method can be applied as an interactive CDSS for clinicians who wish to find answers to particular questions. Although the speed of comprehensively analysing all patient data is already quite quick, it should be noted that there is plenty of room for optimizing the performance of the implementation.

5.2.3 Quantification of disease progression

In Publication III, the objective was to study disease progression quantitatively using heterogeneous longitudinal data from the ADNI MCI cohort. The study evaluated whether it is possible to discern significant trends in the severity of AD as reflected by the DSI and whether subjects who convert from MCI to AD have a different longitudinal DSI behaviour to subjects who do not. To evaluate this, regression parameters were derived from DSI values computed at several time points. The classification of subjects into converters (PMCIs) and non-converters (SMCIs) on the basis of regression parameters was then studied. The cohort used in this study was a more recent and updated version of the ADNI MCI population. See Table 4 for demographic information of the study cohort. Analyses were done using stratified 10-fold cross-validation.

Table 4. Demographics of the Publication III cohort at baseline. Data presented as number of subjects (percentage of subjects %) or mean ± standard deviation. There are fewer subjects than in Publications I and II since only those subjects with enough longitudinal measurements could be included.

	SMCI	**PMCI**
Subjects	149 (51.6%)	140 (48.4%)
Gender		
Female	51 (34.2%)	55 (39.3%)
Male	98 (65.8%)	85 (60.7%)
Age (years)	75.1 ± 7.4	75.4 ± 6.7
Education (years)	15.9 ± 3.0	15.6 ± 3.0

The results in Table 5 show that the change in DSI values over time, as reflected by the slope of the regression equation from longitudinal DSI values, clearly differs between the SMCI and PMCI groups. The slope of the PMCI cases was five times higher than the slope of the SMCI cases. When the slopes of the SMCI cases were studied further, it was noticed that there were two subgroups in the SMCI group: a group with lower slopes and another group with higher slopes that overlap the slopes of the PMCI cases (see Figure 19). It was proposed that the SMCIs with higher slopes represent patients that would progress into AD or other dementia later if the follow-ups were continued. This finding was similar to previous studies in which some SMCI cases had data similar to early AD, suggesting that these subjects may progress into AD in the future [Cui et al. 2011, Davatzikos et al. 2011].

Table 5. Regression parameters of longitudinal DSI values for SMCI and PMCI groups. Values are median (25th percentile, 75th percentile). Disease state index values were derived using all available variables. * statistically significant difference between the groups (Mann-Whitney U test, $p < 0.0005$), + significantly different from zero (one-sample Wilcoxon Signed Rank test, $p < 0.0005$).

	SMCI	**PMCI**
Slope*	0.002 (0.000, 0.006) $^{+}$	0.010 (0.005, 0.015) $^{+}$
Intercept*	0.295 (0.139, 0.621) $^{+}$	0.754 (0.626, 0.860) $^{+}$

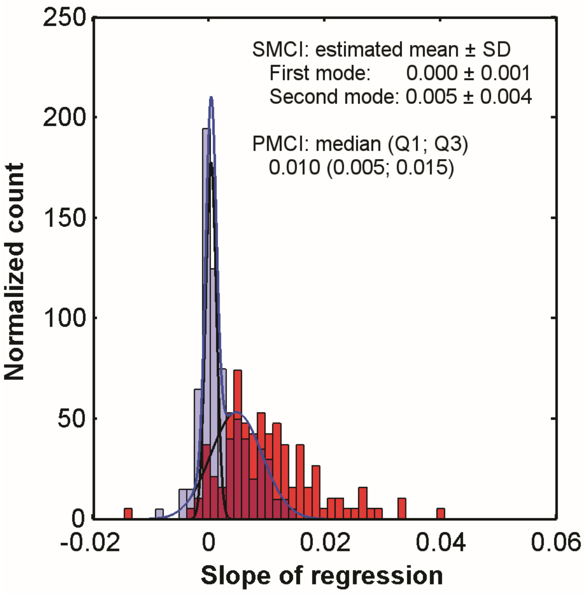


Figure 19. Histograms of the slopes for the SMCI (blue) and PMCI (red) cases. There are two separate subgroups in the SMCI group. A mixture distribution of two normal curves fitted to the slopes of the SMCIs is shown. The areas of the histograms are scaled to one (SD = standard deviation, Q1 = 25th quartile, Q3 = 75th quartile). Reprinted from Publication III with permission from IOS Press © 2014 IOS Press.

The classification using regression parameters obtained from longitudinal DSI values achieved a performance comparable to that of other studies using similar longitudinal ADNI data sets. Regression parameters combining all the data achieved the best classification accuracies and AUCs. The classification accuracy and AUC for the slopes were 76.9% and 82.3%, respectively. For the constants, the prediction accuracy was 74.6% and the AUC was 80.8%.

5.2.4 Optimizing the diagnosis of early AD in MCI

When advising patients and families on the likelihood of transition from MCI to AD, a predictor model with sensitivity and specificity over 80% is essential because false positive and negative rates of over 20% are clinically unacceptable [Ronald and Nancy Reagan Research Institute & National Institute on Aging 1998]. Recent research on large AD cohorts has shown that predicting AD at an early stage using the commonly available biomarkers cannot generally achieve that level, as prediction accuracies remain closer to 70% [Devanand et al. 2008, Hinrichs et al. 2009, Walhovd et al. 2010, Cui et al. 2011, Davatzikos et al. 2011, Ewers et al. 2012, Kruczyk et al. 2012].

As described in Chapter 4, the DSI quantifies the disease state of a patient in a continuous variable in the range [0, 1]. The disease indices from applying the DSI method were found to differ from disease probabilities computed with other classification methods. The DSI values were evenly distributed across the output range, whereas other classifiers maximize class separation to achieve optimal classification performance. In Figure 20, the DSI values are shown to relate to the clinical status of the patient more linearly than the disease probabilities. This linearity was also verified statistically using Kruskal-Wallis and Pearson tests in Publication I.

The goal of Publication IV was to implement and evaluate a novel clinical decision support strategy that makes use of this feature of the DSI method. The premise was to exploit the linearity of the DSI for selecting patients who could be discriminated with improved accuracy, i.e. with sensitivities and specificities closer to 90% than 70%. In practice, the challenge of prediction was approached from a reverse angle, which may better address the clinical need. First, a target prediction accuracy was found at 87.7% by modelling the amount of evidence available in the data when clinical AD diagnoses are made. Then, by selecting subjects with the most or least evidence of early AD – i.e., the ones with the largest and smallest DSI values – subgroups of patients were formed. These subgroups with strong evidence of the disease were such that when considering only them, the prediction accuracy for the selection reached the predetermined target accuracy. The cohort used for these analyses was the same as in Publication III and described in Table 4, and all the analyses were done using ten iterations of 10-fold cross-validation.

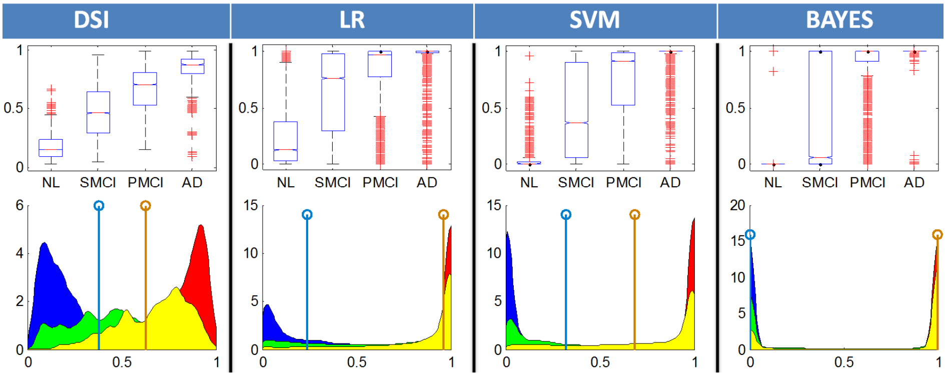


Figure 20. Index and probability distributions of the ADNI MCI data set using DSI, LR, SVM and Bayes, displayed as box plots and probability density estimates of patient classes: healthy controls (blue), SMCI (green), PMCI (yellow) and AD (red). In the box plots, a line in the middle is the median, the upper and lower ends of the box are the 75% and 25% percentiles, and the whiskers show the range. Index/probability values of two arbitrarily chosen SMCI (light blue) and PMCI (brown) patients with similar clinical test results and biomarker data are visualized on top of each distribution graph. The locations of the stems illustrate differences between the methods. Reprinted from Publication I with permission from IOS Press © 2011 IOS Press.

The results obtained with this strategy were promising, considering the DSI's intended use case of clinical decision support in the early diagnosis of AD. Two years before the study subjects received clinical AD diagnoses, approximately one in four had strong evidence of early AD in their measurement data, allowing classification at the target accuracy. One year before AD diagnoses, approximately half of the subjects were included in the group of decisive cases with strong evidence of early AD. Thus, it appears that half of the patients who waited for their AD diagnoses for one or more years could have been considered eligible for diagnosis at least a year earlier, if identified correctly at that time. This is because the early signs of AD were evident in their measurement data and being included in the group of decisive cases implies a prediction accuracy close to 90%, a level similar to clinical diagnoses themselves. In addition to potential AD converters, the strategy revealed with similar accuracy subjects who were likely to remain stable based on their data. These results can be easily understood by examining the green (SMCI) and yellow (PMCI) areas in Figure 20 within a situation in which we only consider subjects whose DSI < 0.3 or DSI > 0.7. Hardly any PMCIs have DSI < 0.3 and, although not visible in the figure, hardly any SMCIs have DSI > 0.7. Thus, the accuracy for discriminating between SMCI and PMCI increases when ambiguous cases between 0.3 and 0.7 are dismissed from consideration.

A feature of the DSI method that helps this optimization strategy is that the relevance function (Equation 4) drives sensitivity and specificity evenly. Relevance does not maximize accuracy; instead it maximizes the sum of sensitivity and specificity. Since the sensitivities and specificities of the cases selected with the DSI method are more or less evenly distributed, the clinical requirement of having both high sensitivity and specificity is fulfilled without additional effort.

In summary, the approach presented in Publication IV provides an additional tool for applying the DSI method in clinical decision support. This data analysis strategy allows clinicians to determine a target sensitivity and specificity and obtain in response threshold DSI values that indicate how high (or low) the total DSI value must be to predict future decline to AD (or stability of MCI) at the target accuracy.

5.3 Evaluation of the clinical decision support system

The DSI method has been shown to perform well as a classifier, providing comprehensive analyses of patient measurement data and a platform for visualizing the results with the DSF. Unfortunately, having these properties does not guarantee that actual benefits would be gained by having the methods available in a CDSS. In other words, until the performance of clinicians making decisions using decision support tools incorporating these methods is evaluated, one does not known whether the assumed benefits are real.

This section provides a summary of the results from Publications V and VI, in which the DSI and DSF methods were evaluated by clinicians in clinical decision support scenarios as part of the PredictAD tool. In addition to the clinical validation studies, all the original publications had a component in which the DSF trees –

disease fingerprints – were studied visually to evaluate the viability of the DSF approach. The section starts with the most important findings from qualitative assessments of DSF visualizations and concludes by presenting results from the clinical validation studies.

5.3.1 Single subject DSF visualizations

Publication I presented the case for using DSI and DSF to obtain a quick overview of patient data in clinical decision-making. One of the main objectives in making this publication was to visually inspect patient DSFs to evaluate their clinical practicality. Over the course of this initial study, DSFs of countless SMCI and PMCI patients were inspected to confirm that they expressed the state of the patient data in relation to control and positive populations and highlighted the tests and variables contributing to the results. Figure 21 shows example DSFs for four subjects: clear SMCI, subtle SMCI, subtle PMCI and clear PMCI. With the clear cases, nearly all the variables point towards AD (shades of red) or against it (shades of blue). With the more ambiguous cases (DSI closer to 0.5), there is a mix of colours that show which patient data indicate AD and which do not.

From Figure 21, it can be seen that the DSF provides a quickly interpretable visual overview of patient state, obtained from statistical analysis of patient data. The colours and shapes of the DSF draw attention to the data that are most relevant, bypassing the need to go over all data points individually. Nevertheless, all data are still available for inspection, if the need to study them arises. DSF also clearly discloses the factors contributing to the results and thus supports application of clinical judgment. To the author’s knowledge, the DSF is unique as a supervised learning data and visualization method that was developed with a philosophy emphasizing both prediction accuracy and clinical interpretability equally.

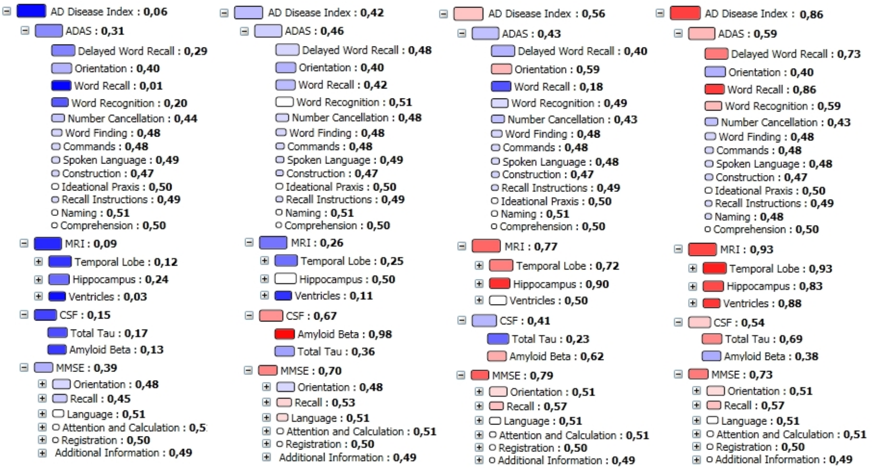


Figure 21. Data of four patients at baseline visualized using the DSF. Starting from the left, two SMCI patients and two PMCI patients are shown. The box sizes (denoting relevance) indicate the capability of a variable or test to discriminate between SMCI and PMCI cases. Sibling nodes are ordered top to bottom according to relevance. Colours indicate into which group the patient data fit better: blue equals SMCI and red equals PMCI. A unique disease fingerprint emerges from the node sizes and colours for each patient, allowing quick evaluation of the patient state and reviewing of individual tests and variables contributing to the results. Reprinted from Publication I with permission from IOS Press © 2011 IOS Press.

5.3.2 Longitudinal DSF visualizations

Since AD is a slowly progressing disease, it is often monitored for some period of time before a diagnosis is arrived at. To facilitate decision support when following the progression of AD and to allow better analysis of longitudinal data, the DSF visualization was extended in Publication III to support temporal data.

A visualization of the progression of AD in MCI subjects with the DSF is demonstrated in Figure 15 on page 55. Most nodes in the longitudinal DSF of a clear SMCI case are blue, indicating that the patient data remained constantly similar to the data of previously seen stable cases (top part of Figure 15). The other SMCI case with increasing DSI values and the DSF changing from blue towards red is shown at the bottom of Figure 15. The data on this subject appear to be progressing slowly towards AD and the subject could be an early AD case yet to be diagnosed, as hypothesized in Section 5.2.3. A regression line showing the trend of changing DSI values projects the trajectory of the disease state into the future.

5.3.3 Comparison with current diagnostic guidelines

In the study described in Publication V, the baseline data of 391 MCI cases in the ADNI cohort were analysed with the objective to predict final clinical diagnoses after three to five years of follow-ups. The baseline data from all the MCIs were evaluated by a single clinician using the PredictAD tool and current guidelines of prodromal AD as identified by combinations of cognitive scores, visual assessment of middle temporal lobe atrophy on MRI, and CSF biomarkers [Dubois et al. 2007, Dubois et al. 2010, Jack et al. 2011, McKhann et al. 2011, Sperling et al. 2011]. The working hypothesis was that computer-assisted analysis of patient data could improve the accuracy of the diagnostic predictions.

The results show that the PredictAD tool alone and the clinician with the assistance of the PredictAD tool were more accurate in predicting three-year MCI outcomes than current research criteria for diagnosis of prodromal AD. Guideline-based predictions using different combinations of examinations achieved accuracies between 57–65%. The accuracy with the PredictAD tool was slightly above 70%, which is not very high with regard to clinical utility but is comparable with the current state of the art. Nevertheless, the clinician was able to select one-third of patients with a clear indication of either early AD or stable MCI for whom the accuracy was 84%, which is at a level that could influence clinical reasoning.

5.3.4 Predicting conversion from MCI to AD with the PredictAD tool

In Publication VI, a cohort of 140 MCI subjects was selected from ADNI. Three clinicians specializing in neurodegenerative diseases used a prototype version of the PredictAD tool to predict which MCI patients would later convert to AD. They rated each subject on a scale of six categories, ranging from 'Clear non-AD' to 'Clear AD'. Non-AD categories were defined to be used for subjects with any other condition than early phase AD. Classifications by clinical raters were compared with the raters' own classifications when deprived of the tool, i.e. only having the patient data on paper. The golden truth for the diagnostic predictions was the clinical diagnoses made by ADNI investigators. In other words, clinical raters were asked to predict three-year conversion outcomes (SMCI or PMCI) using only baseline data from MCIs, including cognitive tests, MRI and CSF biomarkers. The hypothesis was that clinicians would perform better with the tool than without it.

Prototype three of the PredictAD tool, described in Section 4.5.2, provided patient information of subjects one by one to the clinical raters. In the tool, a timeline panel showed tests that had been administered to the patient. Selecting a test from the timeline displayed it in a preview panel, providing detailed results from the selected test. The DSF visualizations showed how patient data relate to data from previously diagnosed SMCI and PMCI cases.

There were two major findings from this study. First, it was shown that inter-rater agreement was greater when clinicians had the tool than when they did not have the tool. Second, the prediction accuracy of clinical raters was superior when using the tool.

When the three raters were using the tool, the inter-rater agreement between them was good, evaluated with quadratic weighted Cohen's kappa (see [Bowers 2008]) as 0.64, 0.76 and 0.80. When deprived of the tool, the agreement between the raters was only moderate (Cohen's kappa: 0.41, 0.43 and 0.71). The agreement between classifications made by a single rater using either the tool or paper charts was relatively good (Cohen's kappa: 0.58, 0.70 and 0.77). In summary, inter-observer differences between ratings were minimized when they used the tool to make categorizations.

When clinical raters were deprived of the tool, there was a decrease in every rater's classification performance. Overall, there was a statistically significant decrease in classification accuracy from 70.0% to 63.2% from when the tool was used to when it was not used. In addition to the decreased classification accuracy, clinical raters were less confident in categorizing patients as 'clear' cases without the tool. With the tool, clinicians selected a third of the patients as clear cases and achieved a prediction accuracy of 85.6% for them. Again, this number could be considered high enough to affect clinical reasoning. With paper charts, only a quarter of patients were selected as clear cases, achieving an accuracy of 82.2% for them. In conclusion, the study found evidence that the PredictAD tool with the DSI and DSF methods allows clinicians to interpret patient data better and predict future decline to AD more accurately.

6. Discussion

6.1 Accomplishment of objectives

The objective of this thesis work was to design and evaluate a disease state quantification method and apply it to clinical decision support in Alzheimer's disease. The work was accomplished by developing three components in parallel and iteratively with clinicians who were considered potential end-users of the methods. The first of them is a supervised machine learning method (DSI) that compares patient data with previously diagnosed cases and estimates the patient's disease state quantitatively. The second component is a method for visualizing the DSI results to allow quick, interactive study of raw patient measures and their relation to a disease. This visualization uses colours and shapes to distinguish the differences between patients, and it was named DSF in an analogy to unique human and DNA fingerprints. The third component is the PredictAD tool, a software tool implementing the DSI and DSF methods in an interactive user interface that was created for performing evaluations with clinicians.

The DSI method, together with its visual counterpart DSF, had several requirements to fulfil, as listed in Section 4.2. Optimizing the classification accuracy was crucial, but it was not the only goal. It had to be balanced against the interpretability of the results, the robustness when using heterogeneous and incomplete data sets and the processing speed requirements. The PredictAD tool needed to provide the methods to clinicians in a package that benefits diagnostics, while also minding usability issues. Based on the results from the studies presented in Chapter 5, this thesis work was able to achieve its objectives.

In Publications I and II, the DSI method was shown to be as accurate as state-of-the-art classifiers when predicting conversion from MCI to AD. The method also performed well with several other publicly available disease data sets. The characteristics of DSI and DSF make them convenient for representing disease state. As discussed in Publications I and III–VI, the linear and slowly changing response to changes in inputs corresponds to the clinical status. This allows quantifying of disease progression over time and, perhaps more importantly, allows for the easy selection of patients with early evidence of the disease for more accurate diagnostics. Other machine learning methods may be capable of extracting the same information from the data, but often they cannot provide it to human readers in a way that it is easily taken into use in clinical decision-making.

The DSI algorithm is computationally lightweight, allowing interactive use and quick generation of personalized disease models. The speed of training and testing multiple disease models is quick enough to allow several hypotheses to be evaluated, which will be important for differential diagnostics. Scalability was achieved by keeping algorithmic complexity low and providing the DSF for visualizing large amounts of data without the need to show everything at once.

For robustness, the design of the DSI takes into account the heterogeneity and sparseness of clinically collected data by allowing the disease models to use whichever values are available. None of the studies employed any pre-processing of data beyond specifying the DSI tree hierarchy. Quite simply, raw patient data were fed into the DSI and DSF methods without data cleaning, feature selection, normalization or any other pre-processing steps.

The interpretability of the results was guaranteed by keeping the DSI method, the DSF visualizations and their interactive implementation within the PredictAD tool relatively simple. The DSF trees allow quick reading of patient data and results from the DSI method. In current practice, clinicians are required to browse test results individually, often in several systems, possibly losing track of the big picture. Clinicians, particularly those with less experience, may be more confident diagnosing AD at an early stage if they are able to see all the data at once and also how patient data relate to previously diagnosed disease populations. While the DSI and DSF increase the amount of information available to a clinician, they also allow clinicians to concentrate on what is important and ignore less relevant information, making the most of existing data.

The clinical objectives of this thesis were to create methods and tools that allow end-users to objectively and comprehensively assess a patient’s disease state in order to diagnose AD earlier and more accurately. The results from Publications V and VI show that the best classification accuracies and agreement between clinicians were achieved when they used the PredictAD tool for decision support. In other words, the most accurate and consistent results were achieved when clinicians combined their clinical expertise in AD with the additional information and context provided by the DSI, DSF and PredictAD tool. Clinicians were also able to select clear cases in which data contained strong evidence of early AD with good accuracy. Thus, it appears that some patients who currently wait several years for an AD diagnosis could be diagnosed earlier if the collected data were interpreted correctly. Earlier diagnoses would bring benefits to treatments, which could start earlier, and to drug trials, for which improved patient selection would reduce the number of subjects needed. Eventually, better profiling of patients could also improve targeting of drugs to the correct patients. Longitudinally quantifying disease state allows objective monitoring of disease progression for diagnostics and for evaluation of drug treatment efficacy.

All the publications in this thesis reflect the reality that current prodromal AD guidelines and combinations of biomarkers are not perfect for predicting AD in the early phase. Prediction accuracies reported in recent literature are commonly between 60–80%. Nevertheless, with the DSI, DSF and PredictAD tool, it was possible to select subgroups of patients where the prediction accuracy is approaching 90%. This capability rises from the objective and evidence-based information on the state of the patient, integrated from large amounts of imperfect heterogeneous measurement data and reported in a manner that can be incorporated to decision making processes. To maximize the value of data that are already being collected in investigations with the patients, clinicians should have tools that allow them to better assess disease severity and detect sub-populations for which diagnostic accuracies are high enough to affect clinical reasoning.

It should be noted that not every step of the design and development work needed to create the DSI and DSF is fully explored in this thesis. For example, several ways to evaluate fitness and relevance were considered in addition to the ones presented in Sections 4.3.3 and 4.3.4. Effects of different approaches on method performance were often negligible or even negative. As an example, more accurate models of relevance caused overfitting of data, sometimes making the final predictions less accurate. The recursive building of the DSI hierarchies also saw several iterations before arriving at the current one, which produces intuitive, stable and consistent results at several levels of abstraction. When considering the impact of different approaches on interpretability and performance, the author believes that the current implementation strikes a good balance between them.

6.2 Impact of the research in its field

Clinical decision support systems have a long history, but there are only a few stand-out success stories. The approach in this thesis work was to produce methods and tools that are viable for the research of diseases but that could also be transferred to clinics in the real world. The intention was to minimize issues restricting deployment, so that the methods have the potential to make an impact in the research community and later in the medical field.

The DSI and DSF methods presented in this thesis are ideologically related to two recently developed disease state quantification methods: bioprofile and disease progression score [Escudero et al. 2012, Jedynak et al. 2012]. These methods quantify heterogeneous patient data to provide an objective and continuous measure for neurodegenerative disease progression over the course of AD. These methods are thus designed for a similar purpose to that of DSI and DSF.

The core concept of the bioprofile method introduced by Escudero et al. [2012] is that it applies unsupervised clustering to a limited set of features, which is selected based on the hypothetical model of AD progression by Jack et al. [2010]. Of the features selected from four different tests, the training data are divided into two clusters each. A priori information about the tests is used for labelling the clusters as ‘having disease’ and ‘not having disease’. When previously unseen patient data are analysed, normalized distances to the cluster centroids are compared to produce a continuous variable called a bioindex between [0, 1]. Classification accuracies achieved with bioprofiling are slightly lower than those commonly published with similar data sets and the method appears to provide little additional value when combining data from multiple tests. Bioprofiling as a method is very simple to grasp, however, allowing intuitive understanding of the results.

The DPS method introduced by Jedynak et al. [2012] is mathematically much more involved than either of the DSI and bioprofile methods. It aims to model the temporal dynamics of multiple biomarkers in a data-driven manner and validates the results in AD by comparing the DPS model to the hypothetical model of AD progression by Jack et al. [2010]. The DPS method analyses all longitudinal patient data simultaneously, fitting the measures to sigmoidal functions over time. From a clinical decision support point of view, it is not totally clear how the DPS would best be used since the method requires longitudinal measurements. This is a challenge for early diagnostics since the first estimates of disease state and progression become available only after waiting for the disease to progress for several months. Nevertheless, as a data-driven disease state quantification method, the DPS is a very valuable addition to the evidence supporting the recently introduced and updated hypothetical models of AD disease progression [Jack et al. 2010, Jack et al. 2013].

To the author’s best knowledge, machine learning methods with goals and requirements similar to those of this thesis work do not exist. The DSI and DSF methods are in a unique niche, and because of the encouraging results they have been taken into regular use by a small group of people in the research community. The methods presented in this thesis seem to be a good fit when supervised learning is applied and interpreted by laymen or domain experts. In addition to Publications I–VI, there are already several other publications using the DSI and DSF as the data analysis methods. For example, the method has been applied to discriminating frontotemporal dementia from MCI and AD [Muñoz-Ruiz et al. 2013] and discriminating AD from several other dementias [Simonsen et al. 2013]. The methods and the PredictAD tool have also been chosen as the clinical decision support platform for two large EU projects targeting differential diagnostics of dementia, VPH-DARE@IT[5] and PredictND[6], which should make the methods well-known in this field. In the machine learning community, the DSI and DSF methods are not yet well known, most probably due to the lack of publications about them published in conferences or journals in the field.

There has also been interest in applying the DSI and DSF methods to contexts other than dementia. They are already being used in TBIcare[7], an EU-funded project that provides objective and evidence-based solutions for the management of traumatic brain injuries. There are also initial concepts for using the algorithm to produce ‘wellness index’ estimates in lifestyle management. In general, there seems to be potential for wider exploitation of the methods, but the extent of their realization remains to be seen.

[5] http://www.vph-dare.eu/, verified available 15.2.2014
[6] http://www.predictnd.eu/, verified available 15.2.2014
[7] http://www.tbicare.eu/, verified available 18.9.2013

6.3 Limitations of the studies

The majority of the research work for this thesis was done with arguably the most comprehensive AD data set available to the wider research community, ADNI. Although it is by all accounts an excellent data set, having used only one is a severe limitation. To counter this limitation, method generalization was evaluated to some extent in Publication II. The studies by Muñoz-Ruiz et al. [2013] and Simonsen et al. [2013] have since provided additional evidence of method generalization. Some of the results in this thesis have also been repeated with another large AD data set, DESCRIPA (Development of screening guidelines and criteria for predementia Alzheimer's disease), described originally by Visser et al. [2008]. This work is pending publication in a journal in the near future. Yet another study that applies the DSI method to four large AD data sets and provides information about inter-data set generalization performance is also under way. Subsequently, the PredictAD tool will be applied to unselected prospective clinical data sets to assess its value in the context of a memory clinic, providing realistic performance measures in clinical use. The current version of the PredictAD tool is being applied to identification of early AD in a pilot study at two sites in Finland. In the future, a version supporting differential diagnostics of dementia will be evaluated at multiple sites in Europe with prospective unselected patients.

The main disadvantage of the DSI and DSF methods and the PredictAD tool is that they require a properly validated training data set of control and disease cases. Building large training data sets is challenging since the data must be collected from actual patients within legal and ethical boundaries. As training data become available, the risks associated with using particular sets of data for modelling a disease and basing decisions on them must be controlled. When using clinical diagnoses as the golden standard, it should also be remembered that an accuracy of 100% is not a realistic target. Clinical AD diagnoses are not always confirmed pathologically, and when they are, the agreement between clinical and neuropathological diagnoses is only 70–90% [Lim et al. 1999, Petrovitch et al. 2001, Kazee et al. 1993]. To achieve the best performance, training data sets with clinical diagnoses and pathological confirmation should be used.

Data obtained in research studies are currently the best starting point for training data, but populations in those represent a selected group of patients as opposed to a general mixed memory clinic population. As such, research data may not fully represent the situation at clinics. For example, affective disorders or measures of depression and anxiety are not included in the AD models in this thesis, although abnormal mood and anxiety of mild severity are associated with MCI and may be confounding factors in the diagnostic process. This omission is simply due to not having those particular data available for analysis. Nevertheless, any data that can influence diagnostic decisions should be included in the DSI model of AD progression or, at a minimum, be available to the clinician when making decisions.

A big limitation in almost all of the current research cohorts is that some of the stable MCI patients may have converted to AD after the follow-up period ended. These patients remain as SMCI in the data but are, in fact, PMCIs, skewing the results. It is possible that the effects of this issue are seen in Publication III, in which a group of SMCI patients was found to have disease progression associated with AD.

In this thesis, all data analyses were performed at cohort level without personalization based on individual patient demographics or genotype. The effects of personalizing the analyses are expected to be rather small, but they should be taken into account when diagnosing real patients. Personalized estimates of the patient's condition would provide as accurate a picture of the situation as possible.

Currently, the methods presented in this thesis support two-class problems only. Extending to multi-class problems is very important since differential diagnostics between multiple possible dementias and mixed dementias are clinically relevant questions. Supporting differential diagnostics requires a strategy for comparing patient data with several disease groups in parallel and an extension to the DSF visualization for quickly comparing patient data with multiple possible disease classes. The performance of the methods in differential diagnostics must also be clarified in upcoming research projects.

6.4 Future work

The work presented in this thesis is an initial platform for the DSI, DSF and PredictAD tool. There are many avenues of research that can be explored to take them further.

The most pressing consideration for future research is a limitation mentioned above: addressing how these methods are best applied when multiple diseases are under consideration. Since the DSI algorithm is computationally inexpensive, several hypotheses can be evaluated quickly. Thus, reducing the multi-class problem into multiple binary classification problems is a valid strategy. Building binary classifiers allows distinguishing between one disease and the rest (one versus all) or between every pair of diseases (one versus one). In one versus all, the disease getting the highest DSI value is the one that patient data resembles most closely. In the one versus one approach, every classifier assigns the patient to one of the two diseases and a composite classification is produced using voting or some other ensemble method. In addition to extensions to the data analysis with the DSI, the DSF visualization requires further development to allow quick interpretation of data produced in multi-class classification problems.

In the future, the DSI should provide support also for features for which values can both increase and decrease in the case of pathology. For example, sleep of eight hours per day can be considered normal but both four hours and twelve hours may be indicative of dementia. As described in Section 4.3.2, the DSI currently requires such features to be split into two. The DSI method would benefit from automatic detection of such features and from a fitness function that produces increasing DSI values when deviating from the normal range, irrespective of the direction of the deviation.

The DSI was designed to be used with unprocessed raw patient data collected at clinics in routine investigations. Accordingly, in the original publications comprising this thesis, using raw heterogeneous patient data without any pre-processing did not adversely affect accuracy or performance. The hierarchical evaluation of the DSI appears to alleviate issues in data correlation as described in Section 4.3.7, but correlations between features and their impact on the optimal organization of the DSI tree hierarchy should be studied more carefully. Generally, developing methods for constructing, optimizing and validating the DSI tree hierarchies is considered an important future research path. As new data sets arrive, there should be tools that propose a hierarchy suitable for data analysis with the DSI method instead of the current ad-hoc approach. Similarly, feature selection methods and the need to apply them should be studied more closely.

With the extensions to the DSI and DSF methods described above, the methods should be able to handle most clinical decision support problems for which they are intended. To verify this, the methods must be tested extensively, using as many data sets as possible. It is expected that in terms of accuracy, the DSI method will not be the best possible classifier for every problem. For decision support and data visualization, it is nevertheless important that the method constantly achieves good accuracies compared with other classifiers, so that it is known to perform robustly in a wide range of problems.

To really see whether the PredictAD tool is able to improve the diagnostics of AD, it must be evaluated with unselected prospective patients at several memory clinics. As was already mentioned, there are several studies in the planning phase that have this agenda. The goal of these studies is to verify that when clinicians analyse patient data with the help of the DSI and the DSF, they are able to make AD diagnoses earlier and more accurately. When support for differential diagnostics is added to the DSI and DSF, the clinical evaluations will also include consideration of multiple possible dementias. In clinical diagnostics, connecting the DSI to the updated hypothetical model of the AD progression should also be taken into account [Jack 2013]. Since the DSI produces results normalized to a range between zero and one, it should be relatively easy to provide additional visualizations of the data overlaid on the hypothetical model. In the clinical evaluations, it is also important to make sure that clinicians can use the CDSS easily, so that it is not dismissed because of usability issues.

Lastly, to simplify deployment of the PredictAD tool to various clinics, tighter integration with HISs should be explored. Although the current implementation can receive brain MRIs from the hospital's picture archiving and communication system (PACS), it is still considered a stand-alone system (category 1) as defined in Section 2.2. The PredictAD tool could be a better fit to clinics as an integrated system (category 2) or as a service model (category 4). Making the switch is mostly a technical and financial issue, due to the fact that HISs are complex software systems and integrations with them are expensive to implement. Standards, like the ones presented in Section 2.2 should help in integration work, but as Brooks

[1987] has said, there is no silver bullet. One option to reduce the need for extensive integrations is to consider business models for bringing a subset of the tools available to clinicians. There may be ways to provide some of the DSI, DSF, and PredictAD tool functionalities to clinicians in a way that would still provide benefits while keeping integrations to existing systems narrow in scope.

7. Conclusions

In this thesis, a disease state quantification method was presented and applied to clinical decision support in AD. The work comprised design, development and evaluation of three components: a supervised learning method (DSI), an accompanying data visualization method (DSF) and an end-user software tool implementing these methods called the PredictAD tool.

The performance of the methods and the tool were evaluated computationally and by clinicians specialized in neurodegenerative diseases. The DSI method performed as well as state-of-the-art reference classifiers and the classification results were consistent throughout all of the studies. In general, the accuracies for predicting conversion from MCI to AD achieved similar levels to those found in recent literature. In addition, the accuracy for predicting conversion from MCI to AD from baseline measurements reached a clinically relevant level of 85% for at least one third of patients in every study. Even though other classification methods may achieve similar accuracies, clinicians always need to consider several aspects in parallel, making the DSI with its visual counterpart DSF good candidates for use in a decision support tool.

In summary, this thesis shows that the supervised learning method DSI, the visualization method DSF and the PredictAD tool could be valuable additions for memory clinics. They could provide assistance in diagnosing AD at an early phase of the disease, selecting patients in pharmacological trials and following the progression of the disease. The linearity of the DSI and its response to changes in the inputs correspond well with a patient's clinical status. The DSF provides quickly interpretable visualizations of patient data and reveals their relation to disease progression. And finally, the PredictAD tool allows clinicians to objectively evaluate all available patient data and select cases in which early AD can be identified accurately enough to be clinically relevant.

References

Albert, M. S., DeKosky, S. T., Dickson, D., Dubois, B., Feldman, H. H., Fox, N. C., Gamst, A., Holtzman, D. M., Jagust, W. J., Petersen, R. C., Snyder, P. J., Carrillo, M. C., Thies, B. & Phelps, C. H. (2011). The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, **7**(3), 270–279.

Alpaydin, E. (2010) *Introduction to Machine Learning, 2nd ed*. Cambridge, MA: MIT Press.

Andersson, S., Heijl, A., Bizios, D. & Bengtsson, B. (2012). Comparison of clinicians and an artificial neural network regarding accuracy and certainty in performance of visual field assessment for the diagnosis of glaucoma. *Acta Ophthalmologica*. Online publication ahead of print, doi: 10.1111/j.1755–3768.2012.02435.x.

Barnett, G. O., Cimino, J. J., Hupp, J. A. & Hoffer, E. P. (1987). DXplain: an evolving diagnostic decision-support system. *The Journal of the American Medical Association*, **258**(1), 67–74.

Berner, E. S. (Ed.). (2007). *Clinical decision support systems: theory and practice*. Dordrecht, Netherlands: Springer Science+ Business Media.

Bhat, G., Biradar, V. G., Sarojadevi, H. & Nalini, N. (2012). Artificial neural network based cancer cell classification. *Computer Engineering and Intelligent Systems*, **3**(2), 7–16.

Bleich, H. L. (1969). Computer evaluation of acid-base disorders. *Journal of Clinical Investigation*, **48**(9), 1689.

Bond, J., Stave, C., Sganga, A., Vincenzino, O., O'connell, B. & Stanley, R. L. (2005). Inequalities in dementia care across Europe: key findings of the facing dementia survey. *International Journal of Clinical Practice*, **59**(146), 8–14.

Bowers, D. (2008). Medical statistics from scratch: an introduction for health professionals. Hoboken, NJ: Wiley-Interscience.

Braak, H. & Braak, E. (2012). Evolution of the neuropathology of Alzheimer's disease. *Acta Neurologica Scandinavica*, **94**(165), 3–12.

Breiman, L. (2001). Random forests. *Machine learning*, **45**(1), 5–32.

Brookmeyer, R., Johnson, E., Ziegler-Graham, K. & Arrighi, H. M. (2007). Forecasting the global burden of Alzheimer's disease. *Alzheimer's and Dementia*, **3**(3), 186–191.

Brooks, F. P. (1987). No silver bullet – essence and accident in software engineering. *IEEE Computer* **20**(4): 10–19*.*

Cattel, C., Gambassi, G., Sgadari, A., Zuccala, G., Carbonin, P. & Bernabei, R. (2000). Correlates of delayed referral for the diagnosis of dementia in an outpatient population. *The Journals of Gerontology Series A: Medical Sciences*, **55**(2), 98–102.

Chen, K., Ayutyanont, N., Langbaum, J., Fleisher, A. S., Reschke, C., Lee, W., Liu, X., Bandy, D., Alexander, G. E., Thompson, P. M., Shaw, L., Trojanowski, J. Q., Jack, C. R., Landau, S. M., Foster, N. L., Harvey D. J., Weiner, M. W., Koeppe, R. A., Jagust, W. J. & Reiman, E. M. (2011). Characterizing Alzheimer's disease using a hypometabolic convergence index. *NeuroImage*, **56**(1), 52–60.

Chincarini, A., Bosco, P., Calvini, P., Gemme, G., Esposito, M., Olivieri, C., Rei, L., Squarcia, S., Rodriguez, G., Bellotti, R., Cerello, P., De Mitri, I., Retico, A. & Nobili, F. (2011). Local MRI analysis approach in the diagnosis of early and prodromal Alzheimer's disease. *NeuroImage*, **58**(2), 469–480.

Collen, M. F., Rubin, L., Neyman, J., Dantzig, G. B., Baer, R. M. & Siegelaub, A. B. (1964). Automated multiphasic screening and diagnosis. *American Journal of Public Health and the Nation's Health*, **54**(5), 741–750.

Cortes, C. & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**(3), 273–297.

Cui, Y., Liu, B., Luo, S., Zhen, X., Fan, M., Liu, T., Zhu, W., Park, M., Jiang, T. & Jin, J. S. (2011). Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PLOS ONE*, **6**(7), e21896.

Davatzikos, C., Bhatt, P., Shaw, L. M., Batmanghelich, K. N. & Trojanowski, J. Q. (2011). Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, **32**(12), 2322–e19.

De Dombal, F. T., Leaper, D. J., Staniland, J. R., McCann, A. P. & Horrocks, J. C. (1972). Computer-aided diagnosis of acute abdominal pain. *British medical journal*, **2**(5804), 9.

Devanand, D. P., Liu, X., Tabert, M. H., Pradhaban, G., Cuasay, K., Bell, K., de Leon, M. J., Doty, R. L., Stern, Y. & Pelton, G. H. (2008). Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biological psychiatry*, **64**(10), 871–879.

Diniz, B. S., Pinto Jr, J. A. & Forlenza, O. V. (2008). Do CSF total tau, phosphorylated tau, and β-amyloid 42 help to predict progression of mild cognitive impairment to Alzheimer's disease? A systematic review and meta-analysis of the literature. *World Journal of Biological Psychiatry*, **9**(3), 172–182.

Donald, A. & Greenhalgh, T. (2000). *A hands-on guide to evidence based health care: Practice and Implementation*. Oxford, UK: Blackwell Science.

Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., Delacourte, A., Galasko, D., Gauthier, S., Jicha, G., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Stern, Y., Visser, P. J. & Scheltens, P. (2007). Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology*, **6**(8), 734–746.

Dubois, B., Feldman, H. H., Jacova, C., Cummings, J. L., DeKosky, S. T., Barberger-Gateau, P., Delacourte, A., Frisoni, G., Fox, N. C., Galasko, D., Gauthier, S., Hampel, H., Jicha, G. A., Meguro, K., O'Brien, J., Pasquier, F., Robert, P., Rossor, M., Salloway, S., Sarazin, M., de Souza, L. C., Stern, Y., Visser, P. J. & Scheltens, P. (2010). Revising the definition of Alzheimer's disease: a new lexicon. *The Lancet Neurology*, **9**(11), 1118–1127.

Duchesne, S., Crépeault, B. & Hudon, C. (2010). Knowledge-based discrimination in Alzheimer's disease. *Medical Content-Based Retrieval for Clinical Decision Support, Lecture Notes in Computer Science*, **5853**(1), 89–96.

Escudero, J., Ifeachor, E. & Zajicek, J. P. (2012). Bioprofile analysis: a new approach for the analysis of biomedical data in Alzheimer's disease. *Journal of Alzheimer's Disease*, **32**(4), 997–1010.

Ewers, M., Walsh, C., Trojanowski, J. Q., Shaw, L. M., Petersen, R. C., Jack, C. R., Feldman, H. H., Bokde, A. L. W., Alexander, G. E., Scheltens, P., Vellas, B., Dubois, B., Weinder, M. & Hampel, H. (2012). Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiology of Aging*, **33**(7), 1203–1214.

Fan, Y., Kaufer, D., & Shen, D. (2010, April). Joint estimation of multiple clinical variables of neurological diseases from imaging patterns. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 852–855.

Farlow, M. R., He, Y., Tekin, S., Xu, J., Lane, R., & Charles, H. C. (2004). Impact of APOE in mild cognitive impairment. *Neurology*, **63**(10), 1898–1901.

Fasano, P. (2013). *Transforming Health Care: The Financial Impact of Technology, Electronic Tools and Data Mining*. Hoboken, NJ: Wiley.

Floares, A., Floares, C., Vermesan, O., Popa, T., Williams, M., Ajibode, S., Chang-Gong, L., Lixia, D., Jing, W., Nicola, T., Jackson, D, Dinney, C. & Adam, L. (2011). Intelligent clinical decision support systems for non-invasive bladder cancer diagnosis. *Computational Intelligence Methods for Bioinformatics and Biostatistics*, **6685**(1), 253–262.

Garg, A. X., Adhikari, N. K., McDonald, H., Rosas-Arellano, M. P., Devereaux, P. J., Beyene, J., Sam, J. & Haynes, R. B. (2005). Effects of computerized clinical decision support systems on practitioner performance and patient outcomes. *The Journal of the American Medical Association*, **293**(10), 1223–1238.

Geldmacher, D. S. & Whitehouse, P. J. (1996). Evaluation of dementia. *New England Journal of Medicine*, **335**(5), 330–336.

Ginsburg, G. S. & McCarthy, J. J. (2001). Personalized medicine: revolutionizing drug discovery and patient care. *Trends in Biotechnology*, **19**(12), 491–496.

Greenes, R. A. (Ed.). (2011). *Clinical decision support: the road ahead*. Waltham, MA: Academic Press.

Gultepe, E., Nguyen, H., Albertson, T. & Tagkopoulos, I. (2012). A Bayesian network for early diagnosis of sepsis patients: a basis for a clinical decision support system. In *2nd International Conference on Computational Advances in Bio and Medical Sciences*, 1–5.

Han, J. & Kamber, M. (2011). *Data Mining: Concepts and Techniques, 3rd Revised Edition*. Burlington, MA: Morgan Kaufmann.

Hinrichs, C., Singh, V., Xu, G. & Johnson, S. (2009). MKL for robust multi-modality AD classification. In *2009 International Conference of Medical Image Computing and Computer-Assisted Intervention*, 786–794.

Hejlsberg, A., Wiltamuth, S. & Golde, P. (2006). *The C# programming language*. Boston, MA: Addison-Wesley Professional.

Hripcsak, G., Ludemann, P., Pryor, T. A., Wigertz, O. B. & Clayton, P. D. (1994). Rationale for the Arden syntax. *Computers and Biomedical Research*, **27**(4), 291–324.

Huang, S., Shen, Q. & Duong, T. Q. (2011). Quantitative prediction of acute ischemic tissue fate using support vector machine. *Brain research*, **1405**(1), 77–84.

Hughes, L., Mthembu, M. & Adams, L. (2011). Diagnostic work-up and treatment of dementia. *Geriatric Medicine*, **41**(11), 595–600.

Institute of Medicine (US) & Committee on Quality of Health Care in America. (2001). *Crossing the quality chasm: A new health system for the 21st century*. Washington, DC: National Academies Press.

Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C. & Trojanowski, J. Q. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer’s pathological cascade. *The Lancet Neurology*, **9**(1), 119.

Jack, C. R., Albert, M. S., Knopman, D. S., McKhann, G. M., Sperling, R. A., Carrillo, M. C., Thies, W. & Phelps, C. H. (2011). Introduction to the recommendations from the National Institute on Aging-Alzheimer’s Association workgroups on diagnostic guidelines for Alzheimer’s disease. *Alzheimer’s and Dementia*, **7**(3), 257–262.

Jack, C. R., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G. Lowe, V., Kantarci, K., Bernstein, M. A., Senjem, M. L., Gunter, J. L., Boeve, B. F., Trojanowski, J. Q., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C. & Knopman, D. S., for the Alzheimer’s Disease Neuroimaging Initiative (2012). Shapes of the trajectories of 5 major biomarkers of Alzheimer disease. *Archives of neurology*, **69**(7), 856–867.

Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., Shaw, L. M., Vemuri, P., Wiste, H. J., Weigand, S. D., Lesnick, T. G., Pankratz, V. S., Donohue, M. C. & Trojanowski, J. Q.

(2013). Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *The Lancet Neurology*, **12**(2), 207–216.

Jaspers, M. W., Smeulers, M., Vermeulen, H. & Peute, L. W. (2011). Effects of clinical decision-support systems on practitioner performance and patient outcomes: a synthesis of high-quality systematic review findings. *Journal of the American Medical Informatics Association*, **18**(3), 327–334.

Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., Raunig, D., Jedynak, C. P., Caffo, B. & Prince, J. L. (2012). A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease neuroimaging initiative cohort. *NeuroImage*, **63**(3), 1478–1486.

Johnson, D. K., Storandt, M., Morris, J. C. & Galvin, J. E. (2009). Longitudinal study of the transition from healthy aging to Alzheimer disease. *Archives of neurology*, **66**(10), 1254.

Kaplan, B. (2001). Evaluating informatics applications – clinical decision support systems literature review. *International Journal of Medical Informatics*, **64**(1), 15–37.

Kawamoto, K., Houlihan, C. A., Balas, E. A., & Lobach, D. F. (2005). Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. *British Medical Journal*, **330**(7494), 765.

Kawamoto, K., & Lobach, D. F. (2005). Design, implementation, use, and preliminary evaluation of SEBASTIAN, a standards-based Web service for clinical decision support. In *American Medical Informatics Association Annual Symposium Proceedings*, 380–384.

Kazee, A. M., Eskin, T. A., Lapham, L. W., Gabriel, K. R., McDaniel, K. D. & Hamill, R. W. (1993). Clinicopathologic correlates in Alzheimer disease: assessment of clinical and pathologic diagnostic criteria. *Alzheimer Disease & Associated Disorders*, **7**(3), 152–164.

Koikkalainen, J. R., Antila, M., Lötjönen, J. M., Heliö, T., Lauerma, K., Kivistö, S. M., Sipola, P., Karrtinen, M. A., Kärkkäinen, S. T. J., Reissell, E., Kuusisto, J., Laakso, M., Orešič, M., Nieminen, M. S. & Peuhkurinen, K. J. (2008). Early familial dilated cardiomyopathy: identification with determination of disease state parameter from cine MR image data. *Radiology*, **249**(1), 88–96.

Koikkalainen, J., Pölönen, H., Mattila, J., van Gils, M., Soininen, H., & Lötjönen, J. (2012). Improved classification of Alzheimer's disease data via removal of nuisance variability. *PLOS ONE*, **7**(2), e31112.

Kruczyk, M., Zetterberg, H., Hansson, O., Rolstad, S., Minthon, L., Wallin, A., Blennow, K., Komorowski, J. & Andersson, M. G. (2012). Monte Carlo feature selection and rule-based models to predict Alzheimer's disease in mild cognitive impairment. *Journal of Neural Transmission*, **119**(7), 821–831.

Landau, S. M., Harvey, D., Madison, C. M., Reiman, E. M., Foster, N. L., Aisen, P. S., Petersen, R. C., Shaw, L. M., Trojanowski, J. Q., Jack, C. R., Weinder, M. W. & Jagust, W. J. (2010). Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, **75**(3), 230–238.

Ledley, R. S., & Lusted, L. B. (1959). Reasoning foundations of medical diagnosis. *Science*, **130**(3366), 9–21.

Legido-Quigley, H., McKee, M., Walshe, K., Suñol, R., Nolte, E. & Klazinga, N. (2008). How can quality of health care be safeguarded across the European Union? *British Medical Journal*, **336**(7650), 920.

Lim, A., Tsuang, D., Kukull, W., Nochlin, D., Leverenz, J., McCormick, W., Bowen, J., Teri, L., Thompson, J., Peskind, E. R., Raskind, M. & Larson, E. B. (1999). Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series. Journal of the American Geriatrics Society, 47(5), 564.

Lindgren, H. (2008). Decision support system supporting clinical reasoning process-an evaluation study in dementia care. *Studies in Health Technology and Informatics*, **136**(1), 315–320.

Lindgren, H. (2011a). Towards personalized decision support in the dementia domain based on clinical practice guidelines. *User Modeling and User-Adapted Interaction*, **21**(4), 377–406.

Lindgren, H. (2011b). Limitations in physicians' knowledge when assessing dementia diseases-an evaluation study of a decision-support system. *Studies in Health Technology and Informatics*, **169**(1), 120–124.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P. & Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *New England Journal of Medicine*, **367**(14), 1355–1360.

Liu, C., Shen, J., Pan, C., Yang, L., Mou, S., Wang, H., & Liang, Y. (2010). MALDI-TOF MS combined with magnetic beads for detecting serum protein biomarkers and establishment of boosting decision tree model for diagnosis of hepatocellular carcinoma. *American Journal of Clinical Pathology*, **134**(2), 235–241.

Lötjönen, J., Wolz, R., Koikkalainen, J., Julkunen, V., Thurfjell, L., Lundqvist, R., Waldemar, G., Soininen, H. & Rueckert, D. (2011). Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. *NeuroImage*, **56**(1), 185–196.

Madureira, S., Verdelho, A., Moleiro, C., Ferro, J. M., Erkinjuntti, T., Jokinen, H., Pantoni, L., Fazekas, F., Van der Flier, W., Visser, M., Waldemar, G., Wallin, A., Hennerici, M. & Inzitari, D. (2010). Neuropsychological predictors of dementia in a three-year follow-up period: data from the LADIS study. *Dementia and Geriatric Cognitive Disorders*, **29**(4), 325–34.

Marling, C. & Whitehouse, P. (2001). Case-based reasoning in the care of Alzheimer's disease patients. *Case-Based Reasoning Research and Development, Lecture Notes in Computer Science*, **2080**(1), 702–715.

McDonald, C. J. (1976). Protocol-based computer reminders, the quality of care and the non-perfectibility of man. *New England Journal of Medicine*, **295**(24), 1351–1355.

McDonald, C. J., Murray, R., Jeris, D., Bhargava, B., Seeger, J. & Blevins, L. (1977). A computer-based record and clinical monitoring system for ambulatory care. *American Journal of Public Health*, **67**(3), 240–245.

McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D. & Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease report of the NINCDS-ADRDA work group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, **34**(7), 939–939.

McKhann, G. M., Knopman, D. S., Chertkow, H., Hyman, B. T., Jack, C. R., Kawas, C. H., Klunkk, W. E., Koroshetzl, W. J., Manlym, J. J., Mayeuxm, R., Mohsp, R. C., Morris, J. C., Rossor, M. N., Scheltens, P., Carrillot, M. C., Thiest, B., Weintraubu, S. & Phelps, C. H. (2011). The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, **7**(3), 263–269.

Miller, R. A., Pople Jr, H. E. & Myers, J. D. (1982). Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *The New England journal of medicine*, **307**(8), 468.

Miller, P. L. (1986). Extending computer-based critiquing to a new domain: attending, essential-attending, and vq-attending. *International Journal of Clinical Monitoring and Computing*, **2**(3), 135–142.

Morris, J. C. (2005). Early-stage and preclinical Alzheimer disease. *Alzheimer Disease and Associated Disorders*, **19**(3), 163.

Mourão-Miranda, J., Oliveira, L., Ladouceur, C. D., Marquand, A., Brammer, M., Birmaher, B., Axelson, D. & Phillips, M. L. (2012). Pattern recognition and functional neuroimaging help to discriminate healthy adolescents at risk for mood disorders from low risk adolescents. *PLOS ONE*, **7**(2), e29482.

Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., Trojanowski, J. Q., Toga, A. W. & Beckett, L. (2005). Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimer's and Dementia*, **1**(1), 55–66.

Muñoz-Ruiz, M. Á., Hartikainen, P., Hall, A., Mattila, J., Koikkalainen, J., Herukka, S. K., Julkunen, V., Vanninen, R., Liu, Y., Lötjönen, J. & Soininen, H. (2013). Disease state fingerprint in frontotemporal degeneration with reference to Alzheimer's disease and mild cognitive impairment. *Journal of Alzheimer's Disease*, **35**(4), 727–739.

Osborn, G. G. & Saunders, A. V. (2010). Current treatments for patients with Alzheimer disease. *Journal of the American Osteopathic Association*, **110**(9), 16–26.

Osheroff, J. A., Pifer, E. A., Teich, J. M., Sittig, D. F. & Jenders, R. A. (2005). *Improving outcomes with clinical decision support: an implementer's guide.* Chicago, IL: Productivity Press.

Oteniya, L., Coles, R. & Cowie, J. (2005). DemNet: A clinical decision support system to aid the diagnosis of dementia. In *Proceedings of the 22nd HealthCare Computing Conference*, 289–297.

Oteniya, L. (2008). *Bayesian belief networks for dementia diagnosis and other applications: a comparison of hand-crafting and construction using a novel data driven technique.* Stirling, Scotland: Department of Computing Science, University of Stirling.

Palmer, K., Fratiglioni, L. & Winblad, B. (2003). What is mild cognitive impairment? Variations in definitions and evolution of nondemented persons with cognitive impairment. *Acta Neurologica Scandinavica*, **107**(179), 14–20.

Pauker, S.G., Gorry, G.A., Kassirer J.P. & Schwartz, W.B. (1976). Toward the simulation of clinical cognition: taking a present illness by computer. *The American Journal of Medicine*, **60**(7), 981–995.

Peck, C. C., Sheiner, L. B., Martin, C. M., Combs, D. T. & Melmon, K. L. (1973). Computer-assisted digoxin therapy. *New England Journal of Medicine*, **289**(9), 441–446.

Petersen, R. C., Smith, G. E., Waring, S. C., Ivnik, R. J., Tangalos, E. G. & Kokmen, E. (1999). Mild cognitive impairment: clinical characterization and outcome. *Archives of Neurology*, **56**(3), 303.

Petersen, R. C. (2004). Mild cognitive impairment as a diagnostic entity. *Journal of Internal Medicine*, **256**(3), 183–194.

Petersen, R. C., & Negash, S. (2008). Mild cognitive impairment: an overview. *CNS Spectrums*, **13**(1), 45.

Petrovitch, H., White, L. R., Ross, G. W., Steinhorn, S. C., Li, C. Y., Masaki, K. H., Davis, D. G., Nelson, J., Hardman, J., Curb, J. D., Blanchette, P. L., Launer, L. J., Yano, K. & Markesbery, W. R. (2001). Accuracy of clinical criteria for AD in the Honolulu–Asia Aging Study, a population-based study. *Neurology*, **57**(2), 226–234.

Pillai, L. (2011). SVM model for amino acid composition based prediction of mycobacterium tuberculosis. *Journal of Computer Science & Systems Biology*, **4**(3), 47–49.

Pople, H. E., Myers, J. D., & Miller, R. A. (1975). DIALOG: A model of diagnostic logic for internal medicine. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence*, 848–855.

Pryor, T. A., Gardner, R. M., Clayton, P. D. & Warner, H. R. (1983). The HELP system. *Journal of Medical Systems*, **7**(2), 87–102.

Rajaratnam, J. K., Marcus, J. R., Flaxman, A. D., Wang, H., Levin-Rector, A., Dwyer, L., Costa, M., Lopez, A. & Murray, C. J. (2010). Neonatal, post-neonatal, childhood, and under-5 mortality for 187 countries, 1970–2010: a systematic analysis of progress towards Millennium Development Goal 4. *The Lancet*, **375**(9730), 1988–2008.

Ram, P., Berg, D., Tu, S., Mansfield, G., Ye, Q., Abarbanel, R. & Beard, N. (2003). Executing clinical practice guidelines using the SAGE execution engine. *Studies in Health Technology and Informatics*, **107**(1), 251–255.

Ronald and Nancy Reagan Research Institute & National Institute on Aging (1998). Consensus Report of the Working Group on: Molecular and Biochemical Markers of Alzheimer's Disease. *Neurobiology of Aging*, **19**(2), 109–116.

Rosenblatt, F. (1958). The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. *Psychological Review*, **65**(6), 386–408.

Sackett, D. L., Rosenberg, W. M., Gray, J. A., Haynes, R. B. & Richardson, W. S. (1996). Evidence based medicine: what it is and what it isn't. *British Medical Journal*, **312**(7023), 71.

Sadeghi, S., Barzi, A., Sadeghi, N. & King, B. (2006). A Bayesian model for triage decision support. *International Journal of Medical Informatics*, **75**(5), 403–411.

Săftoiu, A., Vilmann, P., Gorunescu, F., Janssen, J., Hocke, M., Larsen, M., Iglesias-Garcia, J. & Ciurea, T. (2012). Efficacy of an Artificial Neural Network–Based Approach to Endoscopic Ultrasound Elastography in Diagnosis of Focal Pancreatic Masses. *Clinical Gastroenterology and Hepatology*, **10**(1), 84–90.

Sanchez, E., Toro, C., Carrasco, E., Bonachela, P., Parra, C., Bueno, G., & Guijarro, F. (2011). A Knowledge-based Clinical Decision Support System for the diagnosis of Alzheimer Disease. In *13th IEEE International Conference on e-Health Networking Applications and Services*, 351–357.

Sanchez, E., Toro, C., Artetxe, A., Graña, M., Sanin, C., Szczerbicki, E., Carrasco, E. & Guijarro, F. (2013). Bridging challenges of Clinical Decision Support Systems with a semantic approach: a case study on breast cancer. *Pattern Recognition Letters*, Online publication ahead of print, doi: 10.1016/j.bbr.2011.03.031.

Shortliffe, E. H., Davis, R., Axline, S. G., Buchanan, B. G., Green, C. C. & Cohen, S. N. (1975). Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. *Computers and Biomedical Research*, **8**(4), 303–320.

Sim, I., Gorman, P., Greenes, R. A., Haynes, R. B., Kaplan, B., Lehmann, H. & Tang, P. C. (2001). Clinical decision support systems for the practice of evidence-based medicine. *Journal of the American Medical Informatics Association*, **8**(6), 527–534.

Simonsen, A. H., Mattila, J., Hejl, A. M., Garde, E., van Gils, M., Thomsen, C., Lötjönen, J., Soininen, H. & Waldemar, G. (2013). Application of the PredictAD Decision Support Tool to a Danish Cohort of Patients with Alzheimer's Disease and Other Dementias. Dementia and geriatric cognitive disorders, **37**(3–4), 207–213.

Sitar-Taut, V. A., Zdrenghea, D., Pop, D. & Sitar-Taut, D. A. (2009). Using machine learning algorithms in cardiovascular disease risk evaluation. *Journal of Applied Computer Science & Mathematics*, Issue 5.

Sittig, D. F., Gardner, R. M., Pace, N. L., Morris, A. H. & Beck, E. (1989). Computerized management of patient care in a complex, controlled clinical trial in the intensive care unit. *Computer Methods and Programs in Biomedicine*, **30**(2), 77–84.

Sordo, M., Boxwala, A. A., Ogunyemi, O. & Greenes, R. A. (2004). Description and status update on GELLO: a proposed standardized object-oriented expression language for clinical decision support. Studies in Health Technology and Informatics, **107**(1), 164–168.

Speechly, C. M., Bridges-Webb, C. & Passmore, E. (2008). The pathway to dementia diagnosis. Med Journal of Australia, **189**(9), 487–489.

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., Iwatsubo, T., Jack, C. R., Kaye, J., Montine, T. J., Park, D. C., Reiman, E. M., Rowe, C. C., Siemers, E., Stern, Y., Yaffe, K., Carrillo, M. C., Thies, B., Morrison-Bogorad, M., Wagster, M. V. & Phelps, C. H. (2011). Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's and Dementia*, **7**(3), 280–292.

Spitzer, R. L., Gibbon, M., Skodol, A. E., Williams, J. W. & First, M. (2002). *Dsm-IV-Tr Casebook: A Learning Companion to the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, Text Revision*. Arlington, VA: American Psychiatric Publishing.

Steinfort, D. P., Liew, D., Conron, M., Hutchinson, A. F. & Irving, L. B. (2010). Cost-benefit of minimally invasive staging of non-small cell lung cancer: a decision tree sensitivity analysis. *Journal of Thoracic Oncology*, **5**(10), 1564–1570.

Streba, C. T., Ionescu, M., Gheonea, D. I., Sandulescu, L., Ciurea, T., Saftoiu, A., Vere, C. C. & Rogoveanu, I. (2012). Contrast-enhanced ultrasonography parameters in neural network diagnosis of liver tumors. *World Journal of Gastroenterology*, **18**(32), 4427.

Toro, C., Sanchez, E., Carrasco, E., Mancilla-Amaya, L., Sanín, C., Szczerbicki, E., Graña, M., Bonachela, P., Parra, C., Bueno, G. & Guijarro, F. (2012). Using set of experience knowledge structure to extend a rule set of clinical decision support system for Alzheimer's disease diagnosis. *Cybernetics and Systems*, **43**(2), 81–95.

Visser, P. J., Verhey, F. R. J., Boada, M., Bullock, R., De Deyn, P. P., Frisoni, G. B., Frölich, L., Hampel, H., Jolles, J., Jones, R., Minthon, L., Nobili, F., Olde Rikkert, M., Ousset, P., Rigaud, A., Scheltens, P., Soininen, H., Spiru, L., Touchon, J., Tsolaki, M., Vellas, B., Wahlund, L.-O., Wilcock, G. & Winblad, B. (2008). Development of screening guidelines and clinical criteria for predementia Alzheimer's disease. Neuroepidemiology, 30(4), 254–265.

Walhovd, K. B., Fjell, A. M., Brewer, J., McEvoy, L. K., Fennema-Notestine, C., Hagler, D. J., Jennings, R. G., Karow, D. & Dale, A. M. (2010). Combining MR imaging, positron-emission tomography, and CSF biomarkers in the diagnosis and prognosis of Alzheimer disease. *American Journal of Neuroradiology*, **31**(2), 347–354.

Wang, Y., Fan, Y., Bhatt, P., & Davatzikos, C. (2010). High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *NeuroImage*, **50**(4), 1519–1535.

Warner, H. R., Toronto, A. F., Veasey, L. G., & Stephenson, R. (1961). A mathematical approach to medical diagnosis. *The Journal of the American Medical Association*, **177**(3), 177–183.

Wimo, A. & Prince, M. J. (2010). *World Alzheimer Report 2010: the global economic impact of dementia*. London, UK: Alzheimer's Disease International.

Wright, A. & Sittig, D. F. (2008). A four-phase model of the evolution of clinical decision support architectures. *International Journal of Medical Informatics*, **77**(10), 641.

Wu, S., Chaudhry, B., Wang, J., Maglione, M., Mojica, W., Roth, E., Morton, S. C. & Shekelle, P. G. (2006). Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Annals of Internal Medicine*, **144**(10), 742–752.

Yao, J., Dwyer, A., Summers, R. M. & Mollura, D. J. (2011). Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification. *Academic Radiology*, **18**(3), 306–314.

Yeh, D. Y., Cheng, C. H. & Chen, Y. W. (2011). A predictive model for cerebrovascular disease using data mining. *Expert Systems with Applications*, **38**(7), 8970–8977.

Young, J., Modat, M., Cardoso, M. J., Ashburner, J. & Ourselin, S. (2012). Classification of Alzheimer's disease patients and controls with Gaussian processes. In *9th IEEE International Symposium on Biomedical Imaging*, 1523–1526.

Zhang, D., Wang, Y., Zhou, L., Yuan, H. & Shen, D. (2011). Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage*, **55**(3), 856–867.

PUBLICATION I

A disease state fingerprint for evaluation of Alzheimer's disease

In: Journal of Alzheimer's Disease 27(1),
pp. 163–176.

A Disease State Fingerprint for Evaluation of Alzheimer's Disease

Jussi Mattila[a,*], Juha Koikkalainen[a], Arho Virkki[a], Anja Simonsen[b], Mark van Gils[a], Gunhild Waldemar[b], Hilkka Soininen[c], Jyrki Lötjönen[a] and for The Alzheimer's Disease Neuroimaging Initiative**

[a] *VTT Technical Research Centre of Finland, Tampere, Finland*

[b] *Department of Neurology, Section 2082, The Copenhagen Memory Clinic & The Memory Disorders Research Group, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark*

[c] *Department of Neurology, Kuopio University Hospital, Kuopio, Finland*

Accepted 26 May 2011

Abstract. Diagnostic processes of Alzheimer's disease (AD) are evolving. Knowledge about disease-specific biomarkers is constantly increasing and larger volumes of data are being measured from patients. To gain additional benefits from the collected data, a novel statistical modeling and data visualization system is proposed for supporting clinical diagnosis of AD. The proposed system computes an evidence-based estimate of a patient's AD state by comparing his or her heterogeneous neuropsychological, clinical, and biomarker data to previously diagnosed cases. The AD state in this context denotes a patient's degree of similarity to a previously diagnosed disease population. A summary of patient data and results of the computation are displayed in a succinct Disease State Fingerprint (DSF) visualization. The visualization clearly discloses how patient data contributes to the AD state, facilitating rapid interpretation of the information. To model the AD state from complex and heterogeneous patient data, a statistical Disease State Index (DSI) method underlying the DSF has been developed. Using baseline data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), the ability of the DSI to model disease progression from elderly healthy controls to AD and its ability to predict conversion from mild cognitive impairment (MCI) to AD were assessed. It was found that the DSI provides well-behaving AD state estimates, corresponding well with the actual diagnoses. For predicting conversion from MCI to AD, the DSI attains performance similar to state-of-the-art reference classifiers. The results suggest that the DSF establishes an effective decision support and data visualization framework for improving AD diagnostics, allowing clinicians to rapidly analyze large quantities of diverse patient data.

Keywords: Alzheimer's disease, automatic, biomarkers, computer-assisted, decision making, information processing, projections and predictions

Supplementary data available online: http://www.j-alz.com/issues/27/vol27-1.html#supplementarydata07

INTRODUCTION

Diagnosing Alzheimer's disease (AD) is a nonspecific, subjective, and error-prone process, especially in the early phases of the disease [1]. Because of their inherent difficulty, diagnoses often come late, taking up to two years after initial memory problems occur [2]. Current criteria for AD require early and dominating decline in episodic memory supported by abnormal biomarkers [3, 4]. If a patient with objective evidence of cognitive impairment does not yet meet the criteria for AD or for other dementia, he or she

** Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf.

*Correspondence to: Jussi Mattila, VTT Technical Research Center of Finland, Tekniikankatu 1, FIN-33101, Tampere, Finland. Tel.: +358 40 592 7979; Fax: +358 20 722 3499; E-mail: jussi.mattila@vtt.fi.

is usually diagnosed as having mild cognitive impairment (MCI) [5]. MCI is a heterogeneous state with several possible outcomes and is associated with an increased risk of developing AD, particularly when memory impairment is the predominant symptom [6]. For early diagnosis of AD, a key issue is finding tests and biomarkers that determine which subjects with MCI will develop AD. Here, the term 'biomarker' is used in a broad sense, encompassing biologic features obtained by any and all detection modalities providing information about the disease.

Diverse sets of neuropsychological tests and biomarkers have been investigated for their efficacy to predict conversion from MCI to AD [7–11], and several studies have shown that combining results can yield even better predictions [12–14]. Increased knowledge about cognitive tests and biomarkers has influenced a recent proposal for a new lexicon, where the term AD encompasses the whole spectrum of the disease from predementia to dementia phases and further emphasizes the combination of clinical and biomarker data [15]. However, results from many of the studies are not easily applied in daily diagnostic work. They may require a specific test pattern that is not available or is incompatible due to local or national differences in execution. Occasionally, the statistical analysis methods lack transparency, making them hard to incorporate into local decision making processes. Ultimately, despite all attempts, there have not yet been findings that would comprehensively differentiate MCI subjects who develop AD (progressive MCIs, PMCI) from those who do not (stable MCIs, SMCI).

New approaches for improving the diagnostic process in AD are needed. Computer-based analyses of patient data can quantify information with good diagnostic accuracy, in some cases comparable to experienced clinicians [16]. Tools that help manage the constantly increasing amounts of complex patient data can increase the quantity of information clinicians can examine, and can reveal subtle aspects of information that are buried under a wealth of clinical data [17, 18]. Clinical decision support systems (CDSS) have shown their potential in reducing medical errors and increasing health care quality and efficiency [19–21]. Visualization techniques for analyzing biomedical and temporal data are already commonplace [22, 23], and novel clinical information visualization solutions are constantly being developed [24–26]. Consequently, a statistical Disease State Index (DSI) method is proposed for deriving a scalar value denoting the AD state or progression of AD in suspected AD patients. In this context, AD state measures similarity of patient data to previously-diagnosed healthy and AD populations. While the DSI provides yet another piece of information to clinicians, its goal is to distill existing patient data to a few parameters at a high abstraction level, allowing them to quickly find relevant information and disregard irrelevant information. A Disease State Fingerprint (DSF) visualization technique is also proposed for displaying patient data and DSI values in a concise and interpretable format, extended from earlier research in another biomedical domain [27]. Together, they offer a decision support system that allows clinicians to rapidly extract knowledge from large quantities of heterogeneous patient data and combine them with personal expertise for making the diagnosis.

The main contributions of this work are the proposal of a novel patient data visualization technique (DSF) and the definition of an underlying statistical method for modeling progressing disease state (DSI). The DSI is evaluated against state-of-the-art classifiers using baseline data from the Alzheimer's Disease Neuroimaging Initiative (ADNI); its ability to discriminate healthy elderly controls, SMCIs, PMCIs, and ADs and its capability to predict conversion from MCI to AD are considered. Interpretation of the resulting DSF visualizations and characteristics of the proposed system are reviewed to assess their clinical applicability.

MATERIALS AND METHODS

Alzheimer's disease neuroimaging initiative (ADNI)

ADNI is a longitudinal 5-year study of AD conducted in the USA and Canada, with the goal of developing and validating surrogate markers for early detection and monitoring of AD progression. After launching in late 2004, approximately 800 participants, ranging in age from 55 to 90 years, were recruited for the study: 200 healthy elderly controls, 400 patients with diagnosed MCI, and 200 with early diagnosed AD. Follow-ups of ADNI participants were done by telephone or in person every 6 to 12 months for a period of two to three years. All participants underwent repeated cognitive and neuropsychological testing and magnetic resonance imaging (MRI) scanning. Other tests, including positron emission tomography (PET) and lumbar puncture providing cerebrospinal fluid (CSF) samples, were done more infrequently and not necessarily for all participants. Data from the study are freely available to researchers in an online database at the UCLA Laboratory of Neuroimaging (LONI)

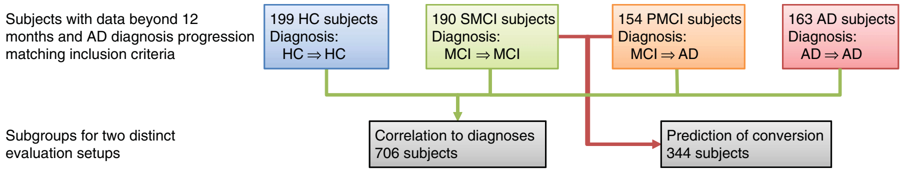


Fig. 1. Patients were divided into two overlapping subgroups. Correlation to actual diagnoses was evaluated using all subjects, capability to predict conversion from MCI to AD using only SMCI and PMCI subjects.

website (http://www.loni.ucla.edu/ADNI/). The site also provides exact information regarding ADNI neuroimaging instrumentation, procedures, and data processing.

Study cohorts and data selection

The analyses in this paper included ADNI subjects who *a)* had follow-up data available beyond the 12-month visit period and *b)* belonged to one of four diagnostic groups based on the baseline diagnosis and the latest diagnosis available in the database (accessed on September 2, 2010). The first three diagnostic groups included subjects whose latest diagnosis was the same as the baseline diagnosis, particularly elderly healthy controls (HC, $n = 199$), stable MCIs (SMCI, $n = 190$), and Alzheimer's disease (AD, $n = 163$). The fourth group was a progressive MCI group (PMCI, $n = 154$), whose diagnosis at the baseline of the ADNI study was MCI, but had converted to AD (on average after 19 months) over the course of the study. Patients whose diagnosis had changed otherwise, such as from MCI or AD to healthy subjects, were excluded from this study. Study cohort selection is illustrated in Fig. 1 and demographic data for the diagnostic groups are presented in Table 1.

Table 1
Demographic and clinical data of the healthy control (HC), stable mild cognitive impairment (SMCI), progressive mild cognitive impairment (PMCI, average conversion time 19 months from baseline), and Alzheimer's disease (AD) groups

	HC	SMCI	PMCI	AD
Subjects	199	190	154	163
Diagnosis				
Baseline	HC	MCI	MCI	AD
Latest	HC	MCI	AD	AD
Gender				
Male	104 (52%)	125 (66%)	93 (60%)	87 (53%)
Female	95 (48%)	65 (34%)	61 (40%)	76 (47%)
Demographics, years				
Age	75.5 (±5.1)	74.8 (±7.6)	74.2 (±6.9)	74.7 (±7.5)
Education	16.1 (±2.8)	15.8 (±3.1)	15.6 (±2.9)	14.9 (±3.1)
Available baseline data				
MMSE	199 (100%)	190 (100%)	154 (100%)	163 (100%)
ADAS	199 (100%)	189 (99%)	152 (99%)	160 (98%)
TMT	199 (100%)	186 (98%)	153 (99%)	156 (96%)
MRI	190 (95%)	171 (90%)	135 (88%)	137 (84%)
CSF	102 (52%)	94 (49%)	83 (54%)	90 (55%)
APOE	199 (100%)	190 (100%)	154 (100%)	163 (100%)

The data are expressed as counts and (percentages) of available data except for age and education, which are expressed as mean (±standard deviation).

All analyses were made using baseline measurement data readily available from the ADNI database. Specifically, patient data obtained from six baseline tests were used; Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS), Mini-Mental State Examination (MMSE), Trail making test from Neuropsychological Battery (TMT), MRI derived volumes (MRI), amyloid-β and total tau from CSF, and apolipoprotein E (APOE). Altogether, analyses were run with 66 unique patient variables distributed into the ten distinct datasets, illustrated in Fig. 2. Sparse and incomplete data were intentionally included to parallel a realistic clinical setting where not all tests are administered to all patients. In ADNI, automated volumetric segmentation of MRI was performed with the Freesurfer image analysis suite [28], which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). Composite variables and summaries of test patterns, e.g., total MMSE score and ADAS 13 point total, were excluded from the datasets, since the same information was contained within the individual variables.

Disease state index

To improve interpretability of heterogeneous patient data, a statistical DSI method has been developed, deriving a scalar index value indicating the state of AD in a patient. The rationale of the DSI is to provide additional evidence-based information by comparing patient data as a whole to a high number of other cases with or without the disease. It is principally intended to

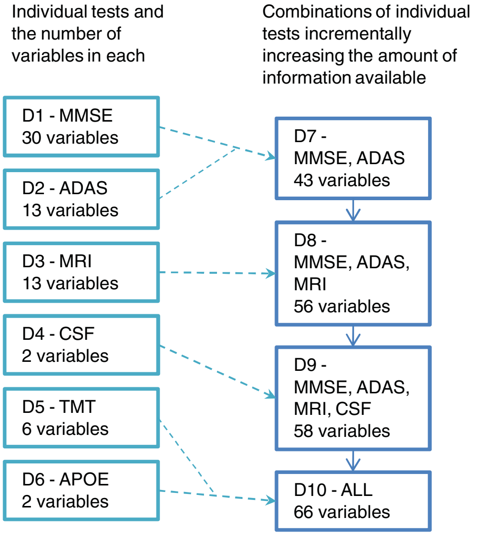


Fig. 2. Analyses were run with ten distinct datasets (D1–D10), formed using variables from six individual tests. The combinations of tests emulate the effect of having incremental tests done, gradually increasing knowledge about the patient. The tests and variable counts included in each dataset are presented in the diagram.

be used with quantitative patient data, such as standardized neuropsychological tests, laboratory test results, and computer-based analyses of medical imaging data. Applying the DSI to patient data results in a value between zero and one, indicating the patient's disease state or progression of the disease. The DSI values are assumed to lie on an interval scale, i.e., one unit on the scale represents the same magnitude across the whole range of the scale. Increasing values of DSI indicate an increasing similarity to AD population, based on the available data. More specifically, DSI measures how individual measurement values and patient data as a whole match the disease profile as defined from a large number of known disease cases.

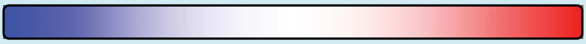
DSI is data agnostic and can be used with any data available. It can determine the disease state between healthy and typical AD, healthy and atypical AD, MCI and AD, and potentially between other dementias and diseases, as long as the training data are available. DSI is also designed to be highly dynamic, not requiring particular tests but using any data acquired and available for the patient being studied. Together, these properties facilitate application of the method at various clinics and re-evaluation of patient data as more test results become available. Choosing a decision boundary allows the DSI to be considered a supervised classifier, discriminating between healthy and diseased patients. Several design requirements were imposed on the DSI according to five classifier performance categories defined by Han and Kamber [29], listed in Table 2.

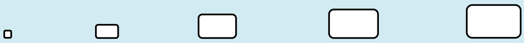
DSI values are computed from patient data in three relatively simple steps. First, each individual patient measurement value, e.g., a single answer in ADAS or the volume of a brain structure derived from MRI, is compared to previously known training data using a *fitness* function. A *fitness* function computes the DSI value for a single patient measurement revealing which population, healthy or diseased, the value fits best. Second, observing only values from known control and disease populations, the *relevance* of each variable is computed, independent of the patient measurement. *Relevance* indicates how well a variable is able to discriminate between the known healthy and diseased populations. Evaluation of *relevance* results in a value between zero and one, obtaining larger values as the separation between control and disease populations increases. Interpretations for different values of DSI and *relevance* are listed in Table 3 and they are derived in full detail in supplementary material (available online: http://www.j-alz.com/issues/27/vol27-1.html#supplementarydata07). Third, DSI and *relevance* values are combined as a weighted arithmetic mean, where DSI values of individual patient measurements are weighted by the variable *relevancies*, to obtain composite DSI values for tests done with the

Table 2
Design goals for the Disease State Index method

Category	Goals for Disease State Index
Interpretability	Provides well-behaving index values that concur with severity of disease state Uses original measurement values in analysis and for reporting the results Facilitates development of interpretable visualizations for expert analyses Accommodates varying clinical and research questions
Prediction accuracy	Classification performance should be comparable to state-of-the-art classifiers
Robustness	Not all patients need to have the same set of tests performed Must be able to use any quantifiable data and all types of variables Missing data should not impose problems for using the method
Computational speed	Allow refinement of parameters and updating of results at interactive rates
Scalability	Enable computation of the model on the fly or beforehand as necessary

Table 3
Interpretations and visualizations of DSI and *relevance*. DSI is computed by comparing the patient values to training data, *relevance* is computed from the known control and diseases population values alone

DSI	0.0	0.5	1.0
Interpretation	Patient value matches the healthy controls perfectly	Patient value falls between control and disease populations, matching both equally well	Patient value matches the AD population perfectly
Visualization	Blue color	White color	Red color
			

Relevance	0.0	0.5	1.0
Interpretation	Not relevant for estimating disease state; variable does not differentiate between known control and disease populations	Relevant for estimating disease state; discrimination capability is halfway between random and perfect discrimination	Very relevant for estimating disease state; variable discriminates perfectly between control and disease populations
Visualization	Excluded from visualization	Intermediate box size	Large box size
			

patient, such as for ADAS and MRI imaging. Correlations between variables can be accounted for at this step, e.g., by applying principal component analysis (PCA) [30].

To obtain a total DSI value representing the combination of all data from multiple tests, the three steps described above are repeated recursively. In lieu of raw measurement values, the DSI values from the previous step are now used for evaluating *relevance* and *fitness*, and merged into a total DSI value (see Fig. 3).

The combination of DSI and *relevance*, schematically depicted in Fig. 4, capture the essence of patient data in relation to the disease. DSI values indicate which patient data are similar to the AD population and *relevance* specifies how important that information should be considered based on previously diagnosed cases. A large DSI value and large *relevance* for a neuropsychological test, for example, indicate that the patient performed similarly to known AD population and that the test has previously been able to discriminate between healthy and AD patients with high accuracy. On the other hand, a test with a large DSI value but little or no *relevance* may usually be ignored, since the test is unable to differentiate between the populations.

Disease state fingerprint

In an analogy to the unique human fingerprints and DNA fingerprints, DSF visualization forms patterns, enabling quick visual inspection of unique disease and patient data at multiple levels of abstraction. In DSF, the patterns emerge from a tree of nodes rendered according to the DSI organization, using shapes and colors to quickly identify the patient's disease state. Specifically, shades of colors indicate DSI values while *relevance* is indicated by node sizes (see Table 3).

The DSF tree allows rapid but detailed reviewing of raw patient measurement data, DSI values, *relevance* values, and the study of their relationship to the disease profile (see Fig. 4). Measures that have zero *relevance* are by default hidden from the DSF visualization. Interactive implementation of the DSF allows visualizations of data distributions (see Fig. 5) and 'drill-down'/'roll-up' operations common to data mining and visual analytics [29]. These operations can be

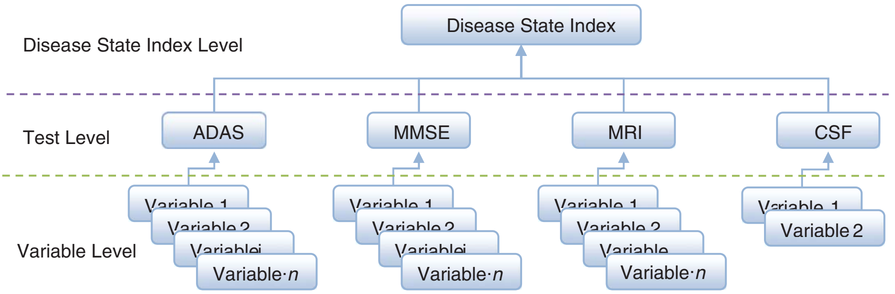


Fig. 3. Organization of a DSI / DSF visualization tree. The tree structure follows organization of patient data consisting of individual variable values (leaf nodes at *Variable Level*), performed tests (internal nodes at *Test Level*), and the resulting total Disease State Index (root node at *Disease State Index Level*). Additional levels can be employed to modify the granularity of the tree.

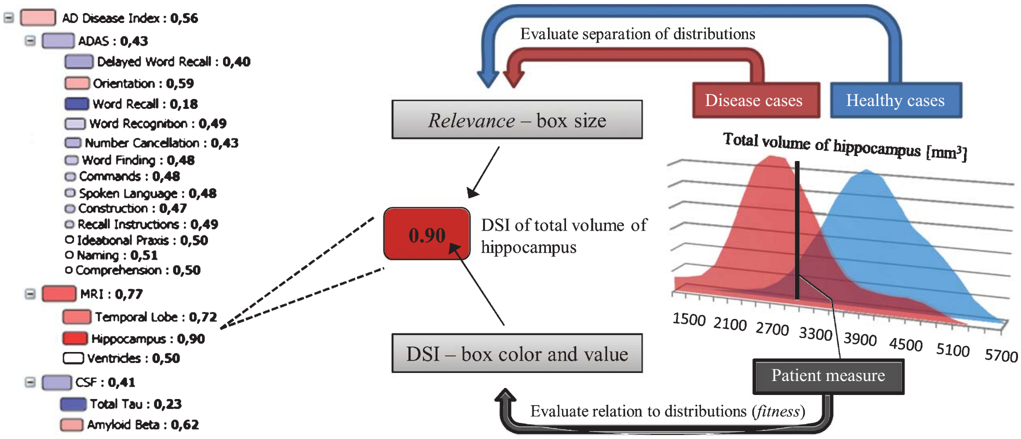


Fig. 4. DSI values of a patient with subtle indication of AD (total DSI value = 0.56). Name of the test and DSI value is shown next to each node. Larger nodes discriminate better between healthy and diseased patients (visualization of *relevance*). 'Hot', i.e., red, nodes highlight patient data that fits AD profile (visualization of DSI). Here, ADAS and MRI contribute the most to the AD Disease State Index, indicated by largest node size. MRI variables, especially volume of hippocampus, whose computation is schematically depicted on the right hand side, push the total DSI value towards AD population.

used for hiding or revealing extra details and for inclusion or exclusion of variables. User initiated changes to DSI model selection can give more control over the study of the patient's disease state, making possible personalized comparison of patient data to previous cases that are of the same gender, age group, ethnicity, and educational degree.

Evaluation

Objectives of the evaluations were to

1. compare the performance of the DSI to state-of-the-art classifiers,
2. evaluate the relationship between the index values and the actual diagnoses,
3. investigate the DSI's capability to predict conversion from MCI to AD, and
4. visually inspect patient DSFs to evaluate their clinical practicality.

In all of the analyses, index values from DSI were compared to the probability of having AD obtained with three reference classifiers: logistic regression (LR) [31], probability estimates from support vector machines (SVM) [32] and Naïve Bayes classifier [33]. These classifiers were chosen as being representative of commonly used classification methods in

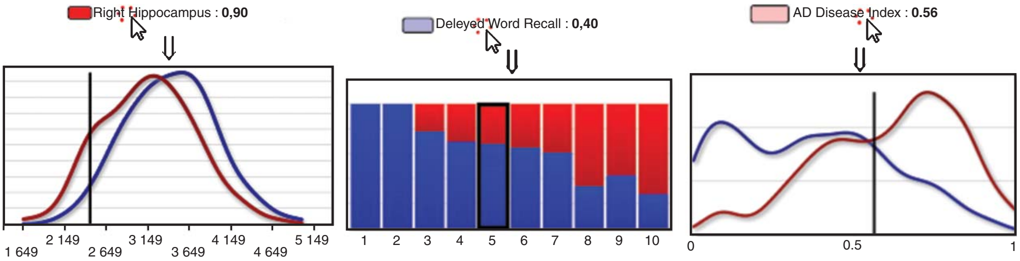


Fig. 5. If a node in the DSF tree is clicked, a comparison of patient data in relation to control and disease population distributions is displayed. Values of the AD population are rendered in red. The healthy population is rendered in blue. Black bars denote the values of the patient being studied. On the left, right hippocampal volume of the patient (2455 mm^3) is overlaid on the distributions. In the center, an ordinal variable (Delayed Word Recall from ADAS) is displayed as bar chart. On the right, total AD Disease State Index of the patient (0.56) is overlaid on the control and AD population DSI values.

practical applications. Like the reference methods, the DSI is a predictor of having AD, and they are all in congruence, with increasing disease probabilities generally resulting in increasing DSI values. Comparisons between values resulting from the DSI and the reference methods are appropriate if one considers the values being used by human readers for decision support. All methods were evaluated using the same training and test data. For LR and SVM, variables not significant between the control and disease populations (Student's t-test result of $p > 0.05$) were excluded. For all reference methods, missing values were handled appropriately. For DSI, this type of pre-processing of the data was not required due to its design.

Comparison to actual diagnoses was performed by training the methods with HC and AD subjects and testing with all patients. The methods' ability to assign values that have a relation with interval-level diagnoses (HC = $\frac{0}{3}$, SMCI = $\frac{1}{3}$, PMCI = $\frac{2}{3}$, and AD = $\frac{3}{3}$) was evaluated using Kruskal-Wallis non-parametric test, Pearson's linear correlation test, and visual inspection. Capability to predict conversion of MCI patients to AD was evaluated by determining area under curve (AUC) measures from receiver-operator curves (ROC) using SMCI-PMCI datasets. MCI patients who obtained index/probability values within the upper or lower ranges of the scale were pooled together to determine classification accuracy for these subsets separately. The patients included in each subset were selected from both ends of the index/probability value range [0, 1], extending to a distance of 0.02, 0.05, 0.1, 0.2, 0.3, and finally 0.4 from either end.

In all analyses, ten iterations of stratified (with same proportions of class labels) 10-fold cross validation were performed to produce robust estimates of performance metrics associated with the methods. Using such a large number of iterations is especially important for data where the differences between classes are subtle and results can vary considerably over consecutive iterations. All analyses were implemented and executed within Matlab version R2010a, using libsvm [32] implementation of SVM and MathWorks® Statistics toolbox implementations of LR and Bayes classifier.

RESULTS

Correlation between disease state index and diagnosis

DSI, LR, SVM, and Bayes classifier were evaluated using baseline data from the ADNI database to determine how they relate to the diagnostic classes of 199 healthy controls, 190 SMCIs, 154 PMCIs, and 163 ADs. Figure 6 shows the box-plots and distributions of values assigned to the patients using the best performing dataset, best individual test dataset, and the worst dataset (ALL, ADAS, and TMT respectively).

The graphs clearly illustrate that DSI is different in nature from the reference methods, distributing index values evenly over the whole scale. The significance of DSI's evident linearity can be appreciated by comparing results from two example patients whose total scores from ADAS differ only slightly (17 vs. 19). With data from ADAS alone, DSI gave to these patients indices of 0.36 and 0.57 (a moderate difference of 0.21), respectively. For the same patients, the probability of having AD estimated by LR were 0.39 and 0.73 (difference of 0.34), by SVM 0.23 and 0.95 (difference of 0.72), and by Bayes 0.0 and 0.75 (difference of 0.75). Especially with SVM and Bayes, the inflated proba-

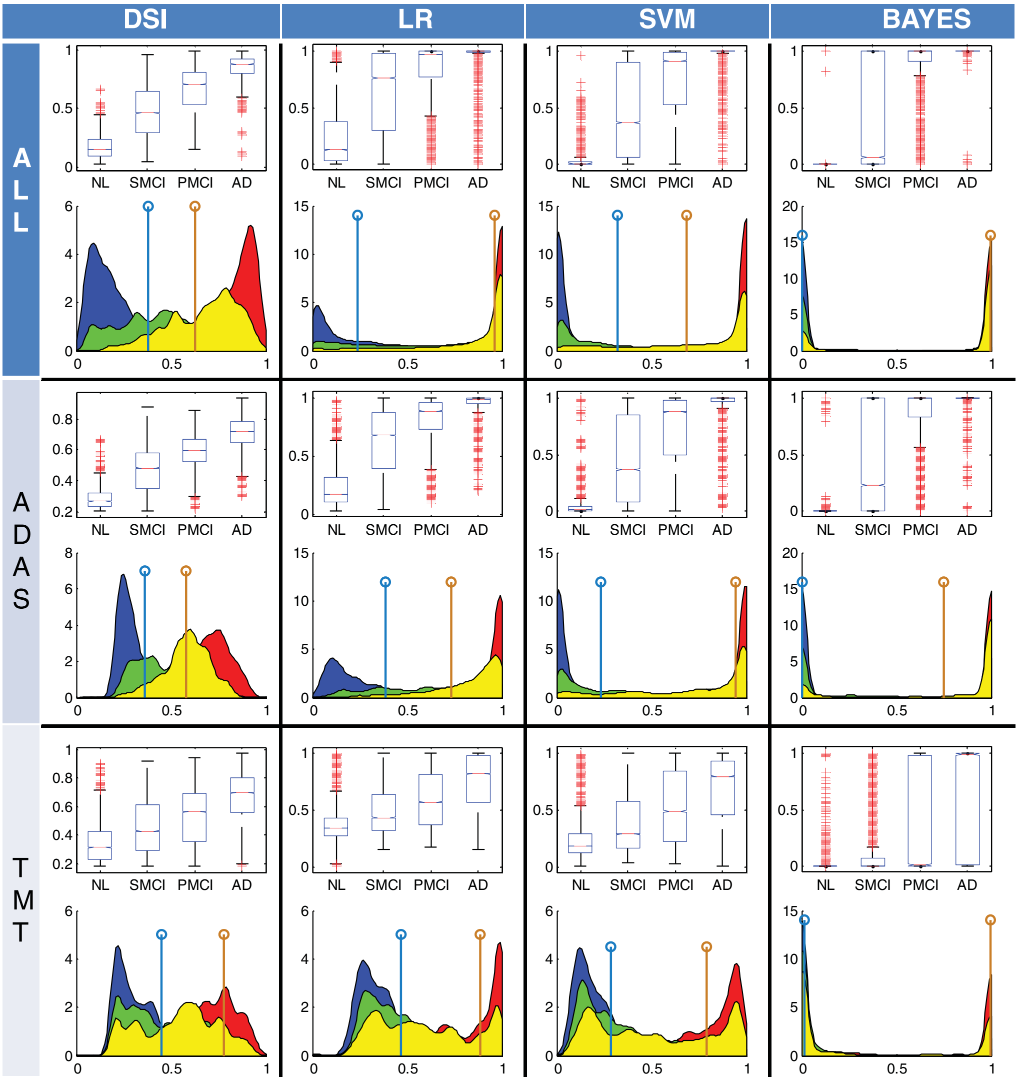


Fig. 6. Index and probability values obtained from evaluating datasets ALL, ADAS, and TMT with DSI, LR, SVM, and Bayes. Results are displayed as box plots and probability density estimates of patient classes NL (blue), SMCI (green), PMCI (yellow), and AD (red) according to index/probability values assigned to them by the methods. In box plots, the line in the middle is the median, the upper and lower ends of the box are the 75% and 25% percentiles, and the whiskers give an indication of the range. Values of two arbitrarily chosen SMCI (light blue) and PMCI (brown) patients with relatively similar clinical test results and biomarker discoveries are visualized on top of each distribution graph. Locations of the stems demonstrate the differences between the methods when assessing individual patients.

bilities obscure what in reality is a small difference in cognitive performance between the patients.

All methods distinguished between the diagnostic categories with high significance ($p < 0.001$ in Kruskal-Wallis test) using all datasets. Linear correlation with interval-level diagnoses also attained high significance ($p < 0.001$ in Pearson) using all datasets. Table 4 shows the eight best and eight poorest performing method/dataset combinations from both statistical tests.

Table 4
Results from the Kruskal-Wallis and Pearson tests using DSI, LR, SVM, and Bayes for discriminating between the diagnostic classes of NL, SMCI, PMCI, and AD and for linear correlation with the interval-level diagnoses, respectively

Kruskal-Wallis					Pearson				
Rank	Method	Dataset	χ^2	p	Rank	Method	Dataset	r	p
1	DSI	ALL	117.6 (5.8)	8.67E–24	1	DSI	ALL	0.56 (0.01)	8.53E–30
2	Bayes	MMSE + ADAS + MRI + CSF	116.4 (6.0)	1.39E–23	2	DSI	MMSE + ADAS + MRI + CSF	0.54 (0.01)	2.70E–26
3	Bayes	MMSE + ADAS + MRI	113.1 (6.4)	1.05E–22	3	DSI	ADAS	0.53 (0.02)	2.88E–26
4	SVM	ALL	113.3 (6.8)	1.45E–22	4	DSI	MMSE + ADASvMRI	0.52 (0.01)	6.10E–25
5	Bayes	ALL	118.2 (7.1)	1.51E–22	5	SVM	ALL	0.50 (0.02)	2.23E–22
6	Bayes	MMSE + ADAS	114.5 (6.6)	2.50E–22	6	LR	ADAS	0.51 (0.02)	2.52E–22
7	DSI	MMSE + ADAS + MRI + CSF	108.4 (5.6)	1.02E–21	7	SVM	MMSE + ADAS + MRI	0.49 (0.01)	7.75E–22
8	SVM	MMSE + ADAS + MRI	107.7 (5.7)	1.31E–21	8	DSI	MMSE + ADAS	0.49 (0.01)	9.92E–22
:	:	:	:	:	:	:	:	:	
33	LR	TMT	40.3 (6.5)	1.25E–06	33	DSI	APOE	0.28 (0.03)	1.72E–06
34	SVM	TMT	39.7 (6.4)	2.49E–06	34	LR	APOE	0.28 (0.03)	1.78E–06
35	Bayes	TMT	35.7 (6.6)	4.49E–06	35	DSI	CSF	0.38 (0.03)	2.28E–06
36	DSI	CSF	29.5 (3.0)	5.01E–06	36	Bayes	APOE	0.28 (0.03)	2.39E–06
37	DSI	TMT	37.3 (6.5)	5.59E–06	37	Bayes	CSF	0.37 (0.03)	2.43E–06
38	SVM	CSF	26.9 (3.5)	2.42E–05	38	Bayes	TMT	0.28 (0.03)	4.16E–06
39	Bayes	CSF	26.4 (3.2)	2.45E–05	39	SVM	CSF	0.35 (0.03)	4.63E–06
40	SVM	APOE	27.0 (5.7)	6.23E–05	40	SVM	APOE	0.27 (0.03)	8.71E–06

The table shows method/dataset performance ordered by the mean of p-values over 10 × 10-fold cross-validation iterations. The Kruskal-Wallis test statistic χ^2 and Pearson test statistic r shown are the mean and standard deviation over 10 × 10-fold cross-validation iterations.

Prediction of MCI to AD conversion

Capability to predict conversion from MCI to AD was evaluated with 190 SMCI and 154 PMCI cases from the ADNI database. Figure 7 shows results from two of the best individual tests and from the four increasingly complete combinations of tests. In general, AUC improves and standard deviation decreases through having better or more patient data available.

Relevance parameters obtained from DSI indicate that within the ADNI database, ADAS is the most relevant single test for predicting conversion from MCI to AD, followed by MRI, APOE, CSF, MMSE, and finally TMT (see Table 5). Within ADAS, relevance values are very similar to weights of a recently introduced ADAS composite [34]. Between all individual variables from all tests, DSI considers the most relevant to be Delayed Word Recall from ADAS (relevance of 0.294), Left Middle Temporal Lobe from MRI (0.262), and Total Tau from CSF (0.258).

Levels of confidence for predicting conversion from MCI to AD

Based on data alone, there are no machine learning methods that can predict conversion from MCI to AD for all cases reliably. Therefore, clinicians always need to consider all available evidence. Nevertheless, index/probability values obtained with the complete dataset (ALL) were examined to determine if the methods studied here could provide more confidence for diagnosing certain subsets of patients. From Table 6 it can be seen that extreme value ranges provide considerably better prediction accuracies and there is a small subset of patients where the classification methods attain perfect prediction accuracy.

Visual inspection of disease state fingerprints

The DSFs of several SMCI and PMCI patients were inspected to confirm that they quickly reveal the state of the patient data in relation to AD population and highlight the tests and variables contributing to the results. Fig. 8 shows example DSFs for clear SMCI, subtle SMCI, subtle PMCI, and clear PMCI cases. With the clear cases, nearly all variables point towards AD (shades of red) or against it (shades of blue). With the subtle cases, there is a mix of colors that show which patient data indicate AD and which do not.

DISCUSSION

The DSF provides a quickly interpretable visual overview of patient state, obtained from evidence-based statistical analysis of patient data. It draws the clinician to data that are the most relevant, omitting the need to go over tens or hundreds of data points individually. DSF clearly discloses the factors contributing to the results, highlights the important measures, and thus supports application of clinical judgment. In its design, equal emphasis was given to prediction accuracy and

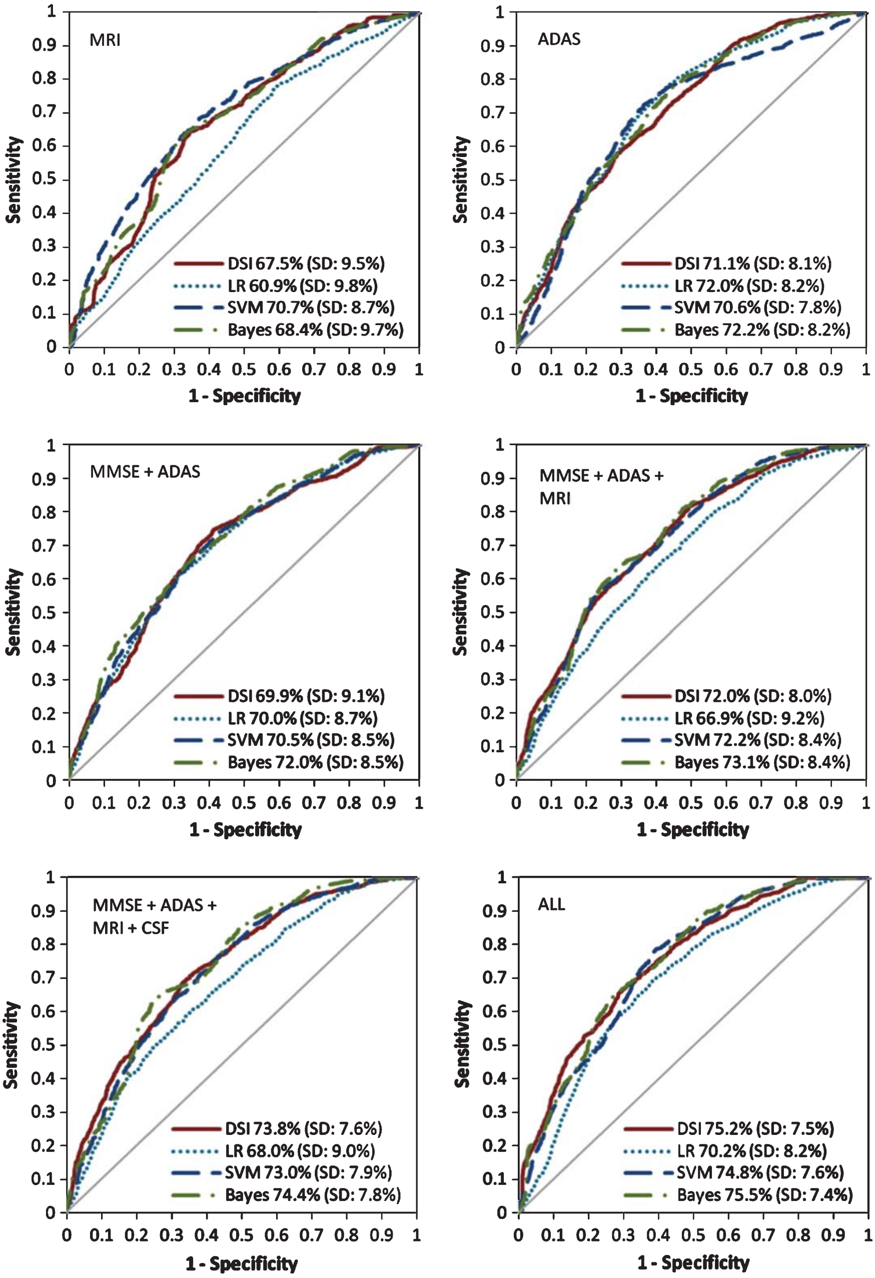


Fig. 7. ROC curves for individual tests and combinations of tests for predicting conversion from MCI to AD. Numbers denote AUC and standard deviations of AUC with the respective datasets over the 10 × 10-fold cross-validation iterations.

Table 5
Relevance values for all data, individual tests, and best individual features (where *relevance* >0.200) for predicting conversion from MCI to AD based on ADNI data

Test	Relevance (SD)	
Disease State Index	0.420	(0.019)
ADAS	0.333	(0.022)
Delayed Word Recall	0.294	(0.018)
Orientation	0.256	(0.017)
Word Recall	0.256	(0.017)
Word Recognition	0.203	(0.017)
MRI	0.300	(0.021)
Left Middle Temporal Lobe	0.262	(0.018)
Right Middle Temporal Lobe	0.246	(0.020)
Left Inferior Temporal Lobe	0.221	(0.021)
Left Hippocampus	0.207	(0.022)
Right Hippocampus	0.201	(0.020)
Right Enthorinal Cortex	0.201	(0.020)
APOE	0.256	(0.016)
Allele B of genotype A/B	0.256	(0.016)
CSF	0.249	(0.029)
Total Tau	0.258	(0.025)
Amyloid-β	0.221	(0.025)
MMSE	0.249	(0.024)
no individual features with relevance >0.200		
TMT	0.207	(0.022)
Time to Complete Trail B	0.202	(0.022)

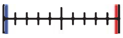
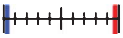
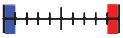
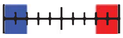
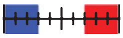
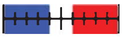
The table shows mean *relevance* values and their standard deviation over 10 × 10-fold cross-validation.

to clinical practicality. To the authors' knowledge there exists no other evidence-based data visualization methods developed with a similar philosophy. Several established machine learning methods were considered for the foundation of DSF, but none were found satisfactory. For example, regression analysis cannot be capitalized fully when working with existing discrete AD diagnoses that do not include much information about the stage or severity of the disease. SVM, with its high dimensional decision boundary, is too abstract for human interpretation. Naïve Bayes works well as a classifier, but results in very unrealistic and unpractical disease probabilities. Thus, DSI was developed to provide a good foundation for visual expert analysis of progressing disease state.

The DSI model of progressing disease state was able to discriminate well between the diagnostic classes of healthy, SMCI, PMCI, and AD and attained good levels of linear correlation, superior to the reference classifiers. Improved linearity is clearly evident with visual inspection of the value distribution graphs in Fig. 6, in which reference methods lean heavily on the head and tail values of the scale even when source data differs only slightly. Thus, DSI is truly indicative of patient state between healthy and AD and appears to correspond well with clinical practice. Even though maximizing classification accuracy was not the only goal, DSI's capability to predict conversion from MCI to AD was similar to the reference classifiers. Analysis of the relevance values reinforced the view that combinations of tests are required for reliable early diagnoses. Interestingly, the relatively simple and computationally low-cost method for computing *relevance* produced almost the same weighting factors as a novel method employed for prediction of 12 months conversion from MCI to AD [34].

Currently, clinicians are forced to browse test results one by one, possibly losing track of the big picture. Analysis of extreme DSI values indicates that there are MCI cases where data leaves little doubt as to whether a patient has AD or not. Particularly those clinicians with less experience might be more confident to diagnose AD at an early stage if they were

Table 6
Classification accuracies of DSI, LR, SVM, and Bayes when observing subgroups of SMCI and PMCI patients based on index/probability values assigned to them

Distance from end of scale [0, 1]	**< 0.02**	**< 0.05**	**< 0.1**	**< 0.2**	**< 0.3**	**< 0.4**
Allowed value ranges						
DSI	*None assigned*	*None assigned*	100% (0.7%)	93.6% (9.1%)	84.0% (30.2%)	75.6% (63.7%)
LR	52.5% (1.2%)	73.6% (5.1%)	71.9% (13.2%)	72.9% (31.7%)	71.4% (52.3%)	67.5% (75.1%)
SVM	100% (1.0%)	95.1% (4.2%)	90.7% (13.1%)	84.3 % (32.3%)	77.3% (53.0%)	72.0% (76.0%)
BAYES	77.6% (57.0%)	74.2% (66.1%)	72.1% (74.2%)	71.1 % (82.7%)	70.7% (88.3%)	70.0% (94.2%)

In parentheses is the percentage of patients assigned to the subgroup over 10 × 10-fold cross-validation iterations. For example, DSI assigned an index value <0.2 or >0.8 to 9.2% of the patients, which was a correct prediction for 93.7% of cases, i.e., classification accuracy for the subgroup was 93.7%.

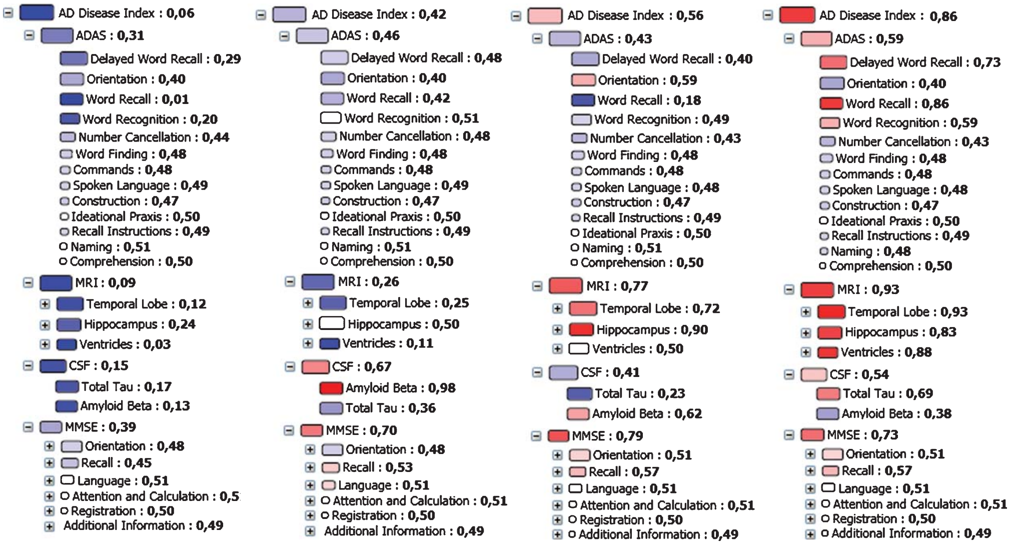


Fig. 8. Four patients visualized using the DSF. Starting from the left, the figure shows two stable MCI (SMCI) patients and two progressive MCI (PMCI) patients. Box sizes (denoting *relevance*) indicate capability of a variable or test to discriminate between SMCI and PMCI cases. The nodes are reordered top to bottom according to this measure. Colors indicate which group the patient data fits better; blue color equals SMCI, red color equals PMCI. A unique disease state fingerprint emerges from the node sizes and color codes for each patient, allowing quick evaluation of patient state and reviewing of individual tests and variables contributing to the results.

able to see all data at once, and also see how patient data relate with previously diagnosed disease population at their clinic. While the DSI and DSF increase the amount of information available to a clinician, they also allow clinicians to concentrate on what is important and ignore irrelevant information, making the most of existing data.

When compared to many other machine learning methods, the benefits of DSI and DSF are numerous. Due to linearity, small changes in patient data cause only small changes in DSI, making interpretation of DSF easier and longitudinal follow-ups consistent. The methods are data agnostic, able to work with any tests or variables in use at a particular clinic. They work with raw test and measurement values, increasing familiarity and requiring no pre-processing of data, feature selection, or data cleanup. All data acquisition modalities are quantified both in isolation and as a part of the whole, providing additional context to the results. It is very easy to support different types of tests and variables (scalar, nominal, ordinal, even textual with text mining methods) with suitable fitness functions. If desired, the probability of a patient having AD can be computed using the DSI values obtained during its evaluation. Unlike as is the case with many other machine learning methods, sparse data creates no problems. Each variable is initially treated individually, and only used if the data exists. Additionally, as long as the training set patients are representative of the control and disease populations, there is no need to have very large quantities of data.

Further potential is anticipated from interactive implementation of DSF, which can provide a quick path to personalized healthcare. Limiting comparison of patient data to cases that are of same gender, age, ethnicity, or educational degree provides personalized results for that patient. A clinical application could also employ *relevance* measures to suggest additional tests to be done, based on their ability to discriminate between healthy and diseased cases. Interactive visualizations of disease population distributions with patient values overlaid on them are an expressive way of comparing patient data to previously diagnosed cases.

Further studies are being planned to cover aspects of DSI and DSF not reported here. Non-linear dependencies between the variables, e.g., differences in cognitive tests due to varying levels of education, were not considered in this work. Stratification of training data will be studied to see if regression of variables based on demographics, such as age and education, further improves the results. Bootstrapping would allow better relevance estimates and, more importantly, would also provide statistical measures that could improve analysis of patient data [35]. Robustness of the DSI will be evaluated with data from several longitudinal studies of AD and other neurodegenerative diseases. Performance of the proposed system will also be examined when there are heavy correlations and other adverse conditions within the data. Utility of an interactive DSF tool is being evaluated with clinicians using existing data from ADNI. There are also plans to take part in upcoming longitudinal studies where an implementation of DSF is provided to the clinicians.

Diagnostic guidelines for AD emphasize the congruence of neuropsychological test results and biomarkers. DSF was designed to enable quick visual analysis of all patient data as a whole. It is a versatile decision support system that uses locally available patient data, presents a synthesis of the information in an understandable manner, and allows an expert to interpret and report the results within the diagnostic process. The proposition is that the DSF can be a clinically relevant tool which enables clinicians to make better and more consistent decisions in daily practice.

ACKNOWLEDGMENTS

The authors thank all participants for their contribution to this project.

This work was supported by the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist); EU-Grant-224328-PredictAD; Name: From Patient Data to Personalised Healthcare in Alzheimer's Disease). Additional grants have been awarded to the medical doctors involved in this study.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration.Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

The funding partners had no active role in this research.

Authors' disclosures available online (http://www.j-alz.com/disclosures/view.php?id=899).

REFERENCES

[1] Waldemar G, Phungh KTT, Burns A, Georges J, Hansen FR, Iliffe S, Marking C, Olde-Rikkert M, Selmes J, Stoppe G, Sartorius N (2007) Access to diagnostic evaluation and treatment for dementia in Europe. *Int J Geriatr Psych* **22**, 47-54.

[2] Bond J, Stave C, Sganga A, Vincenzino O, O'connell B, Stanley R (2005) Inequalities in dementia care across Europe: key findings of the Facing Dementia Survey. *Int J Clin Pract* **59**, 8-14.

[3] Waldemar G, Dubois B, Emre M, Georges J, McKeith IG, Rossor M, Scheltens P, Tariska P, Winblad B (2007) Recommendations for the diagnosis and management of Alzheimer's disease and other disorders associated with dementia: EFNS guideline. *Eur J Neurol* **14**, e1-e26.

[4] Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P (2007) Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS-ADRDA criteria. *Lancet Neurol* **6**, 734-746.

[5] Petersen RC (2004) Mild cognitive impairment as a diagnostic entity. *J Int Med* **256**, 183-194.

[6] Petersen RC, Roberts RO, Knopman DS, Boeve BF, Geda YE, Ivnik RJ, Smith GE, Jack CR Jr (2009) Mild cognitive impairment: Ten years later. *Arch Neurol* **66**, 1447-1455.

[7] Visser PJ, Verhey F, Knol DL, Scheltens P, Wahlund LO, Freund-Levi Y, Tsolaki M, Minthon L, Wallin AK, Hampel H, Bürger K, Pirttila T, Soininen H, Rikkert MO, Verbeek MM, Spiru L, Blennow K (2009) Prevalence and prognostic value of CSF markers of Alzheimer's disease pathology in patients with subjective cognitive impairment or mild cognitive impairment in the DESCRIPA study: A prospective cohort study. *Lancet Neurol* **8**, 619-627.

[8] Morris JC, Roe CM, Grant EA, Head D, Storandt M, Goate AM, Fagan AM, Holtzman DM, Mintun MA (2009) Pittsburgh compound B imaging and prediction of progression from cognitive normality to symptomatic Alzheimer's disease. *Arch Neurol* **66**, 1469-1475.

[9] Petersen RC, Smith GE, Ivnik RJ, Tangalos EG, Schaid DJ, Thibodeau SN, Kokmen E, Waring SC, Kurland LT (1995) Apolipoprotein E status as a predictor of the development of Alzheimer's disease in memory-impaired individuals. *JAMA* **273**, 1274-1278.

[10] Devanand DP, Pradhaban G, Liu X, Khandji A, De Santi S, Segal S, Rusinek H, Pelton GH, Honig LS, Mayeux R, Stern Y, Tabert MH, de Leon MJ (2007) Hippocampal and entorhinal atrophy in mild cognitive impairment: prediction of Alzheimer's disease. *Neurology* **68**, 828-836.

[11] Querbes O, Aubry F, Pariente J, Lotterie JA, eacute D, monet JF, Duret V, Puel M, Berry I, Fort JC, Celsis P (2009) Alzheimer's Disease Neuroimaging Initiative early diagnosis of Alzheimer's disease using cortical thickness: Impact of cognitive reserve. *Brain* **132**, 2036-2047.

[12] Devanand DP, Liu X, Tabert MH, Pradhaban G, Cuasay K, Bell K, de Leon MJ, Doty RL, Stern Y, Pelton GH (2008) Combining early markers strongly predicts conversion from mild cognitive impairment to Alzheimer's disease. *Biol Psychiat* **64**, 871-879.

[13] Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, Weiner MW, Knopman DS, Petersen RC, Jack CR, Jr. (2009) Alzheimer's Disease Neuroimaging, Initiative MRI and CSF biomarkers in normal, MCI, and AD subjects: Diagnostic discrimination and cognitive correlations. *Neurology* **73**, 287-293.

[14] Risacher SL, Saykin AJ, West JD, Shen L, Firpi HA, McDonald BC (2009) Alzheimer's Disease Neuroimaging, Initiative Baseline MRI predictors of conversion from MCI to probable AD in the ADNI Cohort. *Curr Alzheimer Res* **6**, 347-361.

[15] Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, Delacourte A, Frisoni G, Fox NC, Galasko D, Gauthier S, Hampel H, Jicha GA, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Sarazin M, de Souza LC, Stern Y, Visser PJ, Scheltens P (2010) Revising the definition of Alzheimer's disease: a new lexicon. *Lancet Neurol* **9**, 1118-1127.

[16] Kloppel S, Stonnington CM, Barnes J, Chen F, Chu C, Good CD, Mader I, Mitchell LA, Patel AC, Roberts CC (2008) Accuracy of dementia diagnosis – a direct comparison between radiologists and a computerized method. *Brain* **131**, 2969-2974.

[17] Huang QR, Qin Z, Zhang S, Chow CM (2008) Clinical patterns of obstructive sleep apnea and its comorbid conditions: A data mining approach. *J Clin Sleep Med* **4**, 543-550.

[18] Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, Muller R, Robson B, Apte C, Weiss S, Rigoutsos I, Platt D, Cohen S, Knaus WA (2006) Data mining and clinical data repositories: Insights from a 667,000 patient data set. *Comput Biol Med* **36**, 1351-1377.

[19] Bates DW, Cohen M, Leape LL, Overhage JM, Shabot MM, Sheridan T (2001) Reducing the frequency of errors in medicine using information technology. *J Am Med Inform Assoc* **8**, 299-308.

[20] Sim I, Gorman P, Greenes RA, Haynes RB, Kaplan B, Lehmann H, Tang PC (2001) Clinical decision support systems for the practice of evidence-based medicine. *J Am Med Inform Assn* **8**, 527-534.

[21] Kawamoto K, Houlihan CA, Balas EA, Lobach DF (2005) Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. *Evid Based Cardiovasc Med* **9**, 208-210.

[22] Smith NB, Webb A (2010) Introduction to Medical Imaging: Physics, Engineering and Clinical Applications, Cambridge University Press, Cambridge.

[23] Zimetbaum PJ, Josephson ME (2008) *Practical clinical electrophysiology*, Lippincott Williams & Wilkins, Philadelphia.

[24] Aigner W, Kaiser K, Miksch S (2008) Visualization Methods to Support Guideline-Based Care Management. *Stud Health Technol Inform* **139**, 140-159.

[25] Bade R, Schlechtweg S, Miksch S (2004) Connecting time-oriented data and information to a coherent interactive visualization. *Proc CHI* **2004**, 105-112.

[26] Wang TD, Plaisant C, Shneiderman B (2010) Visual Information Seeking in Multiple Electronic Health Records: Design Recommendations and a Process Model. *Proc IHI* **2010**, 46-55.

[27] Koikkalainen JR, Antila M, Lötjönen JM, Heliö T, Lauerma K, Kivistö SM, Sipola P, Kaartinen MA, Kärkkäinen ST, Reissell E, Kuusisto J, Laakso M, Oresic M, Nieminen MS, Peuhkurinen KJ (2008) Early familial dilated cardiomyopathy: Identification with determination of disease state parameter from cine MR image data. *Radiology* **249**, 88-96.

[28] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM (2002) Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341-355.

[29] Han J, Kamber M (2006) Data Mining: Concepts and Techniques, 2nd ed, Morgan Kaufmann Publishers, Burlington.

[30] Jolliffe IT (2002) Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed, Springer, New York.

[31] Hosmer DW, Lemeshow S (2000) Applied logistic regression. Wiley-Interscience, New Jersey.

[32] Chang C, Lin C LIBSVM (2001) A Library for Support Vector Machines, http://www.csie.ntu.edu.tw/cjlin/, Accessed February 17 2011 libsvm.

[33] Mitchell TM (1997) Machine Learning. McGraw-Hill Series in Computer Science, McGraw-Hill, New York.

[34] Llano DA, Laforet G, Devanarayan V (2011) Derivation of a New ADAS-cog Composite Using Tree-based Multivariate Analysis: Prediction of Conversion From Mild Cognitive Impairment to Alzheimer's Disease. *Alz Dis Assoc Dis* **25**, 73-84.

[35] Mooney CZ, Duval RD, Duval R (1993) Bootstrapping: A nonparametric approach to statistical inference. Sage Publications, Inc, London.

Journal of Alzheimer's Disease 27 (2011) 1–3
IOS Press

Supplementary Data

A Disease State Fingerprint for Evaluation of Alzheimer's Disease

Jussi Mattila[a,*], Juha Koikkalainen[a], Arho Virkki[a], Anja Simonsen[b], Mark van Gils[a], Gunhild Waldemar[b], Hilkka Soininen[c], Jyrki Lötjönen[a] and for The Alzheimer's Disease Neuroimaging Initiative**

[a] *VTT Technical Research Centre of Finland, Tampere, Finland*
[b] *Department of Neurology, Section 2082, The Copenhagen Memory Clinic & The Memory Disorders Research Group, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark*
[c] *Department of Neurology, Kuopio University Hospital, Kuopio, Finland*

Accepted 26 May 2011

SUPPLEMENTARY DATA

Derivation of the fitness function for scalar variables

Suppose first that x is a random variable from a distribution combining both control (e.g. healthy) and positive (e.g., disease) subjects, marked with C and P. In addition, assume that the progression of disease increases the observed values of x, making the conditional expected value $E(x|P)$ higher for positives than the corresponding value $E(x|C)$ for the controls (see Fig. 1).

Let us divide the probability density f into the components f_C and f_P, such that

$$\begin{aligned} f(x) &= f_C(x) + f_P(x) \\ &= f(x|C)p(C) + f(x|P)p(P) \end{aligned} \tag{1}$$

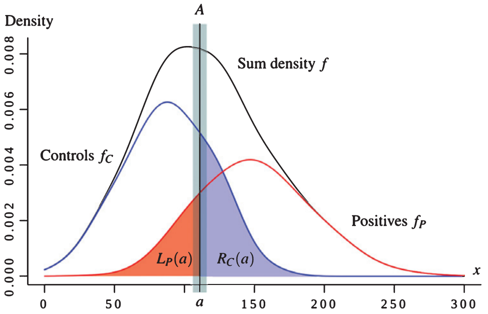


Fig. 1. Probability density function $f(x)$ and its components f_C and f_P for the control and positive groups and, respectively.

where $f(x|C)$ and $f(x|P)$ are the marginal distributions of C and P, and the probabilities $p(C)$ and $p(P)$ correspond to the overall fraction of controls and positives in the study population, respectively. These values are related by the equation $\int_R f(x)dx = \int_R \left[f_C(x) + f_P(X)\right] dx = p(C) + p(P) = 1$, obtained by integrating (1).

Bayes' theorem states that the conditional probability of a subject belonging to the group P after observing $x \in A = (a - \varepsilon; a + \varepsilon)$, where ε is the radius of a small region around the actual observation a, can be written as

** Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf.

*Correspondence to: Jussi Mattila, VTT Technical Research Center of Finland, Tekniikankatu 1, FIN-33101, Tampere, Finland. Tel.: +358 40 592 7979; Fax: +358 20 722 3499; E-mail: jussi.mattila@vtt.fi.

ISSN 1387-2877/11/$27.50 © 2011 – IOS Press and the authors. All rights reserved

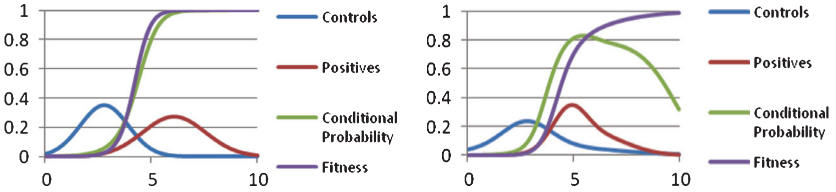


Fig. 2. Conditional probability and *fitness* computed with two distributions of control and positive cases. On the left, synthetic data with no outlier observations produces monotonously increasing curves with both methods. On the right, data with wide distribution tails causes drastic change to conditional probability behavior, rendering it sub-optimal for human interpretation.

$$\begin{aligned} p(P|x \in A) &= \frac{p(x \in A|P)p(P)}{p(x \in A)} \\ &= \frac{p(x \in A|P)p(P)}{p(x \in A|P)p(P) + p(x \in A|C)p(C)} \\ &\rightarrow \frac{f_P(x)}{f_P(x) + f_C(x)}, \end{aligned} \quad (2)$$

when $\varepsilon \rightarrow 0$. While estimating the probability densities from empirical data, one needs to smooth the estimates e.g. by using a sufficiently wide kernel estimate, or to use a large enough ε to compensate for the measurement noise and errors caused by the finite number of samples. If ε is chosen to be too large, $p(P|x \in A)$ approaches the a priori fraction of the positive cases $p(P)/(p(P) + p(C))$, and equation (2) loses its predictive power. The values $p(P)$ and $p(C)$, also called the "a priori" probabilities, are needed when applying the Bayes' rule. This can be a great asset, but could also be regarded as a drawback when used or interpreted incorrectly. In addition, distributions of the form (2) have also some inconvenient properties, which can make their interpretation difficult, as shown later in Fig. 2.

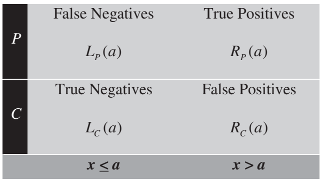
Instead of using conditional probability (2), let us introduce a *fitness* function *Fit(a)*, which increases monotonously. In a sense, *fitness* describes the location of the subject with value a relative to distributions f_C and f_P. Let us first define the left and right integrals for f_C and f_P,

$$L_P(a) := \int_{-\infty}^{a} f_P(x)\,dx \quad \text{and}$$

$$R_C(a) := \int_{a}^{\infty} f_C(x)\,dx, \quad (3)$$

which are also illustrated in Fig. 1. For completeness, $R_P(a)$ and $L_C(a)$ are defined in an analogous manner. If one consider value a as the clinical threshold for classification between the controls C and positives P, one can construct a new boolean classifier, described in Table 1.

Table 1
Classification performance when using the value a as the threshold to discriminate between controls C and positives P



P	False Negatives $L_P(a)$	True Positives $R_P(a)$
C	True Negatives $L_C(a)$	False Positives $R_C(a)$
	$x \leq a$	$x > a$

Table 1 shows that $P(x \leq a) = L_C(a) + L_P(a)$ and $P(x > a) = R_C(a) + R_P(a)$ for the columns and $P(P) = L_P(a) + R_P(a)$ and $P(C) = L_C(a) + R_C(a)$ for the rows. In particular, the fraction of rejection errors (false negatives) from all the errors (both false negative and false positive) can be written as

$$Fit(a)^{*} : = \frac{FN(a)}{FN(a) + FP(a)} = \frac{L_P(a)}{L_P(a) + R_C(a)}, \quad (4)$$

where the abbreviations *FN* and *FP* refer to false negatives and positives, i.e. the counts of incorrectly classified instances. It is obvious from equation (4) that $Fit(a)* \in [0, 1]$ and one can intuitively expect that $Fit(a)^*$ increases along with increasing values of a, which is proved by differentiating (4):

$$\begin{aligned} \frac{d}{da} Fit(a)^{*} &= \frac{\frac{d}{da} L_P(a) R_C(a) - L_P(a) \frac{d}{da} R_C(a)}{[L_P(a) + R_C(a)]^2} \\ &= \frac{f_P(a) R_C(a) + L_P(a) f_C(a)}{[L_P(a) + R_C(a)]^2} \\ &\geq 0 \quad \text{for each } a, \end{aligned} \quad (5)$$

In the special case $L_P(a) = R_C(a) = 0$ where (5) is not defined, the result can be interpolated from closest values of a where (5) is defined. Finally, to eliminate the influence caused by varying proportions of $p(P)$ and $p(C)$ between different populations, the normalized *fitness* value is defined as

$$Fit(a) := \frac{L_P(a)/p(P)}{L_P(a)/p(P) + R_C(a)/p(C)} \qquad (6)$$

Derivation of the *fitness* function can be conducted in an analogous manner if populations are interchanged, resulting in a monotonously decreasing function. In addition, alternate formulations of *fitness* functions to account for non-continuous variables, such as nominal and ordinal variables, can also be derived easily by counting these values as point masses while computing the integrals. The resulting *fitness* values obtained by evaluating (6) are in many situations close to the conditional probabilities (2) but *fitness* behaves in a more intuitive manner with real-life empirical distributions, as demonstrated in Fig. 2. For example, it is known that atrophy decreases the size of hippocampus in Alzheimer's disease; the smaller the size of hippocampus the higher the DSI value should be, i.e., the function should be monotonous. However, if the number of cases in the training set is small and conditional probabilities are used, posterior probability can decrease even while the hippocampus volume is decreasing.

It merits restating that the number of instances in either class of the training set does not bias *fitness* (6), which makes it robust against disparity between the numbers of class instances. Since the *fitness* values are not intended to be used solely as a machine learning classifier but accompanied with visual analysis tools, this choice offers more intuitive ratings for measured values. Additional information related to the class probabilities, i.e., disease incidence and prevalence, should be presented to clinicians via the graphical user interface.

Derivation of the composite Disease State Index

In clinical practice, multiple variables must be considered simultaneously. Combining results from several *fitness* functions would allow evaluation of large quantities of heterogeneous patient data at once. Due to its simplicity and interpretability, the weighted arithmetic mean is employed for combining *fitness* values. Let us define the composite Disease State Index (DSI) as

$$DSI(a_1, a_2, \ldots, a_n) := \frac{\sum_{i=1}^{n} \mathrm{w}_i Fit(a_i)}{\mathrm{w}_1 + \mathrm{w}_2 + \cdots + \mathrm{w}_n}, \qquad (7)$$

where $[a_1, a_2, \ldots, a_n]$ are the data measured from the subject and $\mathrm{w} = \mathrm{w}_1, \mathrm{w}_2, \ldots, \mathrm{w}_n$ are the non-negative weights for each of the variables according to their *relevance*. *Relevance* is a parameter quantifying a variable's ability to differentiate classes C and P. To compute the *relevance* of the ith variable, the classification accuracy is estimated by applying the *fitness* function to the training data itself:

$$Acc(i) = \frac{|C_T : Fit(a_i) < \frac{1}{2}| + |P_T : Fit(a_i) > \frac{1}{2}|}{|C_T| + |P_T|} \qquad (8)$$

where C_T and P_T are the corresponding training sets for the controls and positives (with the i:th variable present) and $\frac{1}{2}$ is the classifier threshold value for a. Now, *relevance* of a variable is formally defined as

$$Rel(i) := \max\left\{0, \left(Acc(i) - \frac{1}{2}\right) * 2\right\}. \qquad (9)$$

If the *relevance* is zero, it discriminates the classes as poorly as a random label. A *relevance* of one indicates that the variable is capable of fully discriminating between training classes C and P, thus being an excellent candidate for estimating the disease state. Substituting w_i in (7) with (9) yields

$$DSI(a_1, a_2, \ldots, a_n) := \frac{\sum_{i=1}^{n} Rel(i) Fit(a_i)}{\sum_{i=1}^{n} Rel(i)}. \qquad (10)$$

It is clear from (10) that like *fitness*, composite $DSI(a_1, a_2, \ldots, a_n) \in [0, 1]$. It must be emphasized that DSI cannot be considered as the probability of having the disease. Instead, it is a score that increases with the probability of having the disease, taking into account the assumption that having abnormally high (or low) values is worse than being inside the normal range. Thus, DSI is defined as a value derived from a series of observed facts that describes the rank of patient data relative to control and positive cases.

PUBLICATION II

Design and application of a generic clinical decision support system for multiscale data

In: IEEE Transactions on Biomedical
Engineering 59(1), pp. 234–240.

Design and Application of a Generic Clinical Decision Support System for Multiscale Data

Jussi Mattila*, Juha Koikkalainen, Arho Virkki, Mark van Gils, *Member, IEEE*, and Jyrki Lötjönen; *for the Alzheimer's Disease Neuroimaging Initiative*

***Abstract*—Medical research and clinical practice are currently being redefined by the constantly increasing amounts of multiscale patient data. New methods are needed to translate them into knowledge that is applicable in healthcare. Multiscale modeling has emerged as a way to describe systems that are the source of experimental data. Usually, a multiscale model is built by combining distinct models of several scales, integrating, e.g., genetic, molecular, structural, and neuropsychological models into a composite representation. We present a novel generic clinical decision support system, which models a patient's disease state statistically from heterogeneous multiscale data. Its goal is to aid in diagnostic work by analyzing all available patient data and highlighting the relevant information to the clinician. The system is evaluated by applying it to several medical datasets and demonstrated by implementing a novel clinical decision support tool for early prediction of Alzheimer's disease.**

***Index Terms*—Clinical diagnosis, decision support systems, software architecture, supervised learning.**

I. Introduction

ADVANCES in multimodal data acquisition instrumentation have resulted in a deluge of data that have contributed significantly to scientific research of diseases [1]. It has also altered the daily clinical practice by increasing the amount of patient information that clinicians must manage. Everything from questionnaire answers to laboratory results and information obtained with sophisticated imaging methods must be considered when making diagnostic decisions. Furthermore, new knowledge about diseases is unveiled at an unparalleled rate, making the deliberate application of evidence-based medicine a challenging and time-consuming effort.

One approach to managing this complexity is to develop detailed computer-based multiscale modeling, simulation, and analysis systems. Multiscale is defined here as patient data obtained at several scales, e.g., with genetic, molecular, structural, and neuropsychological tests. Models that describe phenomena of human physiology at a particular scale may be combined, usually with considerable effort, for understanding of larger entities [2]. Physiological multiscale models are often custom-built to target a single organ, disease, or condition, and they help develop treatments, biomarkers, and even personalized disease models for use in clinical work. They have already proven useful and shall remain the focus of much of future research [3], [4].

The increasing number and scale of measurements can improve one's understanding of a system even without detailed physiological modeling. There are established machine learning methods that can classify a patient as being healthy or diseased or provide the probability of having a disease when trained with previously diagnosed patient data [5]. Recent research has introduced mathematical and statistical models, which derive composite disease indicators from quantitative multiscale and multimodal data. Their goal is to give prognoses, e.g., in the context of prostate cancer [6] or Alzheimer's disease (AD) [7]. An alternative method is to employ data-driven techniques that divide all the experimental data into components for analysis. Study of the components can provide insight into the subsystems and ultimately to the system as a whole. Such an approach, able to handle empirical patient data and implemented within a clinical decision support system (CDSS), could transform existing patient data into knowledge applicable in the clinical setting [8].

One major hurdle for the widespread use of these systems and CDSSs in general is that data collected at different clinics vary considerably. Consequently, most CDSSs for medical diagnostics are purpose-built expert systems targeting a single condition or a family of diseases, and also require a particular set of data [9]. Generic CDSSs for clinical diagnostics have also been developed, traditionally employing Bayesian inference [10], text-mining methods [11], case-based reasoning [12], or fuzzy cognitive maps (FCM) [13]. But even with the more generic CDSS systems, most require definition of disease-specific model parameters by domain experts before they can be put into use.

This manuscript describes a data-agnostic clinical decision support system, implemented as a reusable software library. The software library uses a statistical approach to analyze multiscale data and combine them into an aggregate representation interpretable by a clinician. It supports heterogeneous patient data of virtually any type and scale and allows clinicians to study the system simultaneously as a collection of components and as a whole. The library has been designed to easily support several

Manuscript received April 8, 2011; revised June 22, 2011 and August 24, 2011; accepted September 26, 2011. Date of publication October 10, 2011; date of current version December 21, 2011. This work was supported in part by the 7th Framework Program by the European Commission (http.//cordis.europa.eu/ist; EU-Grant-224328-PredictAD; From Patient Data to Personalized Healthcare in Alzheimer's Disease). Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). *Asterisk indicates corresponding author.*

*J. Mattila is with the VTT Technical Research Centre of Finland, Tampere, Finland (e-mail: jussi.mattila@vtt.fi).

J. Koikkalainen, A. Virkki, M. van Gils, and J. Lötjönen are with the VTT Technical Research Centre of Finland, P.O. Box 1300, Finland (e-mail: juha.koikkalainen@vtt.fi; arho.virkki@vtt.fi; mark.vangils@vtt.fi; jyrki.lotjonen@vtt.fi)

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBME.2011.2170986

diseases, requiring minimal amount of configuration. The first application prototype developed using the proposed decision support library is a CDSS tool for early diagnosis of AD. The statistical methods are validated using data from several medical datasets and the clinical applicability of our proposed system is demonstrated by evaluating the implementation of the CDSS tool.

The main contributions of this work are the description of the generic decision support software library, the statistical method behind it, and evaluations of classification and computational performance of the proposed system using several medical datasets. A more thorough analysis of the statistical method and its relationship to established machine learning methods with regards to AD is available in [14].

II. Materials and Methods

A. Evaluation of Disease State

In this work, a data-agnostic statistical disease modeling method has been developed. It combines heterogeneous multiscale data to compute a value in the interval [0,1], indicating a patient's disease state, i.e., the location or rank based on data, in relation to previously known control (healthy) and positive (disease) populations. It is intended to be used mainly with quantitative features, such as standardized questionnaire answers, laboratory analysis results, automatically quantified biomedical data, and outputs of personalized disease model simulations. It can be considered a supervised classifier, where patient data are compared to previously diagnosed data. In its development, equal emphasis was given to classification accuracy and to clinical interpretability of the results.

Given the heterogeneous patient data from a single test at a single time point, e.g., an individual neuropsychological test or laboratory analysis results of a blood sample, as $x_1, x_2, \ldots, x_n$, we define the *n*-variable scalar valued disease state index (DSI) function as a weighted mean

$$\mathrm{DSI}(x_1, x_2, \ldots, x_n) := \frac{\sum_{i=1}^{n} \mathrm{Rel}(i)\mathrm{Fit}(x_i)}{\sum_{i=1}^{n} \mathrm{Rel}(i)} \tag{1}$$

where Rel(i) is a *relevance* function providing the weighting between [0,1] for variable i and Fit(x_i) is a *fitness* function providing a nonlinear transformation of value x_i into *fitness* space [0,1].

A *fitness* function computes the location, i.e., rank, of an individual variable x_i relative to values of the same variable in two different populations, denoted as controls C_i and positives P_i. Our system currently supports scalar, ordinal, and categorical (including boolean) variables, but could be extended to support others, such as value lists and complex values, by deriving appropriate *fitness* functions. Let us consider a scalar variable where the progression of a disease tends to increase its value (see Fig. 1). For these, *fitness* is defined as a monotonically increasing function

$$\mathrm{Fit}(x_i) := \frac{L_P(x_i)}{L_P(x_i) + R_C(x_i)} \tag{2}$$

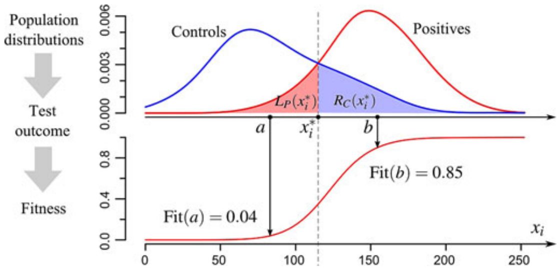


Fig. 1. Probability density functions of C_i and P_i, the resulting *fitness* (with examples at test outcome values a and b), and the optimal classification threshold x_i^*.

where $L_P(x_i)$ is the left integral of probability density function (PDF) for positive class values P_i and $R_C(x_i)$ is the right integral of PDF for control class values C_i. Derivation of the *fitness* function can be conducted in an analogous manner for ordinal variables. For a categorical variable $x_i \in \{\Omega_1, \ldots, \Omega_n\}$, we use as *fitness* the conditional probability of the subject belonging to the positive population in the case of observing $\Omega = x_i$.

The weighting factors of DSI, i.e., *relevancies* of variables, are determined by the variables' ability to correctly classify between the known classes C_i and P_i, and are independent of the patient data. *Relevance* is defined for scalar and ordinal values that increase with disease progression as

$$\mathrm{Rel}(i) := \max\{0, L_C(x_i^*) + R_P(x_i^*) - 1\} \tag{3}$$

where $L_C(x_i^*)$ is the left integral of PDF for control values C_i and $R_P(x_i^*)$ is the right integral of PDF for positive values P_i at the decision threshold x_i^* (shown in Fig. 1). For categorical variables, *relevance* is the classification accuracy of training cases given the category of the independent variable.

To combine data from multiple tests and/or multiple scales, DSI values obtained from (1) are recursively inserted back into (1) as new variables, using several levels of recursion for granularity. Recursive evaluation provides *fitness*, *relevance*, and DSI values for a tree of data, where the leaves and branches represent multiple scales but converge to a common root describing the whole system. This tree of data can be rendered for quick visual interpretation of multiscale data, using colors and shapes to quickly distinguish patient state and the relevance of all tests and variables. The nodes can also be ordered according to *relevance* to show the most important features at the top (see Fig. 2).

In summary, DSI uses available multiscale data to model the state of having a disease. It does so first with the individual measurement values, then transforms the values nonlinearly to a common classification space and combines them within that space to obtain aggregate results. The recursive computation produces classification results at multiple levels of abstraction, which can be visualized using a tree hierarchy.

B. Decision Support Library

We have developed a software library implementing the DSI computational method and supporting features using the C#

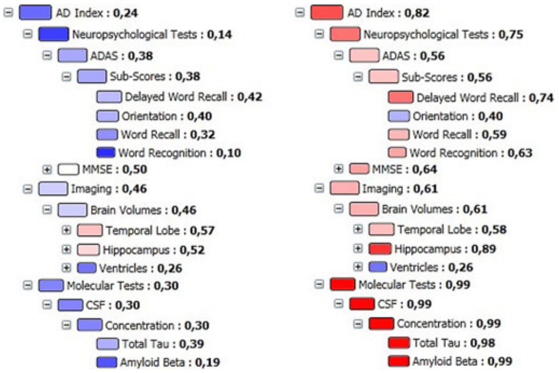


Fig. 2. DSI tree visualizations for two patients, one healthy, one with AD. Larger node sizes indicate higher *relevance* (i.e., better discrimination of training classes), with irrelevant features omitted. Shades of red indicate similarity of the patient data to the disease population, shades of blue similarity to healthy.

language (see Fig. 3). The library is context independent, and thus is applicable to several domains.

Since the DSI can use any available multiscale data, the library supports accessing multiple data repositories with a layered approach. Data access implementations, called persistence stores$^{(a)}$ in Fig. 3, are free to connect with data sources in any way that is needed, e.g., through an object relational mapping (ORM) service, web services, or simply reading a flat text file. An interface defines how the persistence stores can transfer data to and from the library.

A data definition layer$^{(b)}$ comprises descriptions of entries (e.g., types of tests done to a patient) and feature values (types of individual data points) within those entries. Definitions are application-specific metadata and must be configured in source code or by Extensible Markup Language (XML) when initializing the library for use. In addition to all features existing at the leaf nodes, the organization of the DSI tree hierarchy is also described within this layer. The actual data that are analyzed are contained within another layer$^{(c)}$, where all the subjects, entries, and feature values are represented by matching object instances, as described in Table I.

Performing DSI computations requires the library to construct control and positive classes in a generic manner, using entities from one or more persistent stores that provide training data. For this, we have developed a rule-based grouping system$^{(d)}$, where a grouping rule interface is called to check whether a training entity belongs to a particular class, e.g., to healthy controls or Alzheimer's disease patients. A CDSS tool using this decision support library is aware of the context and is responsible for defining the group forming rules, e.g., "if diagnosis equals AD, assign patient to group AD." A graphical user interface (GUI) component is available in the decision support library to allow interactive modification of the rules that have been implemented so far. If necessary, new rule implementations can be created. They are able to use all available patient information when deciding whether he or she is to be included in a training class or not.

After applying grouping rules, entities in control and positive classes are known$^{(e)}$. Now, the library must collect all types of values from the entities in a generic manner. For this, we have developed a sampling system$^{(f)}$, where sampling policies control how data from a single entity are chosen for training. One can, e.g., use the mean of all scalar values for a particular feature or pick the value that was obtained most recently. As with the grouping rules, the sampling policy implementations can be configured with a GUI component and new ones can be implemented in source code if complex sampling policies are necessary. Custom grouping and sampling may be used, e.g., for personalized healthcare, where stratification is employed to collect feature values with age and gender constraints.

Now, having the training data$^{(g)}$, data from the patient we are studying, and with the definition layer$^{(b)}$ describing the feature hierarchy, the library has all the necessary information for evaluating the DSI$^{(h)}$. Training data obtained through grouping and sampling is organized in the tree hierarchy where the leaves contain actual measurement values for the training set. *Fitness* and *relevance* are evaluated at the leaf level, DSI and *relevance* values in internal nodes are computed recursively, and, finally, a total DSI value for the whole dataset at the root of the DSI tree is obtained.

The library provides implementations of GUI components for displaying DSI trees$^{(i)}$, data distributions$^{(j)}$, entry timeline$^{(k)}$, and entry details$^{(l)}$. These are implemented on top of the logic tier using Windows Presentation Foundation (WPF) platform.

C. Data Access Implementations

Currently, there exist two implementations of persistence stores$^{(a)}$ for accessing patient information to be used with the decision support library. One of them uses an entity-attribute-value (EAV) scheme, which is a common methodology for database design in healthcare applications, thanks to its applicability to storing heterogeneous and sparse patient data [15]–[17]. EAV is well suited for querying data of individual patients, but it is well known to be inefficient for bulk queries, which are needed for collecting large quantities of training data [18]. These require the use of a normalized database where the patient and all record types are represented by their own tables [19]. Unfortunately, this is a conflicting requirement for the decision support library, which strives to be a generic one, accepting any kind of data from any clinic to be incorporated into it. To overcome the conflicting requirements, a normalized database and persistence store generators have been developed to go along with the library. They are based on C# language features, such as partial classes and reflection [20], with Entity Framework 4 (EF4) [19] and Text Template Transformation Toolkit (T4) [21] engine used for generating all the necessary constructs without hard-coding any data descriptions. Reflection is a mechanism in object-oriented programming languages that is used for examining, instantiating, and using unknown types. Partial classes allow splitting class definitions to several source files. It is often used to combine machine generated source code in one file with manually written source code in another.

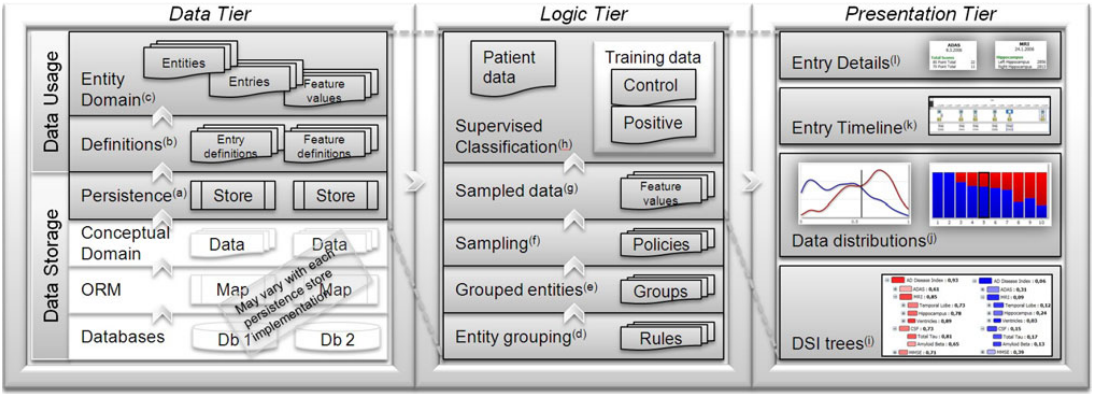


Fig. 3. Tiers, layers, and components of the generic decision support library, also showing the main direction of data flow.

TABLE I
LIBRARY RUNTIME DATA STRUCTURES

Instance	Contains	Purpose	Example usage
Entity	Entries	Object of interest	A patient
Entry	Features	Data container	Blood analysis results
Features			
Text	free text	Value container	Verbatim answer
Scalar	double	Value container	Blood pressure
Nominal	category	Value container	Multiple choice question
Ordinal	position	Value container	# of words remembered

Exact composition of entries and features are described in the definition layer. E.g., an entry definition provides a list of feature values that it can contain, while a nominal feature definition defines the list of allowed values.

The process of generating normalized databases utilizes the data definition layer(b), which is also used within the library for describing the CDSS data organization. A T4 script reads the data definitions and automatically transforms this metadata to database generation commands, which can be executed to create a new database containing data tables adhering to the given data definitions. With the database structure in place, one can create, using EF4, an object relational mapping (ORM) that allows writing and reading data in the database tables. More specifically, the EF4 tooling environment builds a conceptual model of the database by inspecting its structure and generates the necessary code for transferring data between the database and an application using the data. The EF4 generated conceptual model uses strongly typed C# classes, again working against the requirement of providing a generic decision support library. With strongly typed classes, it is normally required to explicitly declare the type of the class before using it, which in this case is impossible since the database structure is unknown to the library. To overcome this, another T4 script is used for generating partial class definitions that augment the EF4-generated conceptual model classes. The partial class definitions add functionality that allows the augmented object instances to be created and manipulated, using reflection, in a manner that can be considered weakly typed. Through these mechanisms and with information from the data definition layer(b), generic implementations of persistence stores(a) are able to access normalized databases and transform patient data contained within those into the data structures (entities/patients, entries, and feature values) used by the decision support library. Finally, there are tools to populate persistence stores(a) with data from other persistence stores as necessary.

Together, the decision support library and data access implementations form a data-agnostic end-to-end system, which can generate and populate appropriately designed databases based on the data definitions and provide evidence-based decision support using the statistical DSI method.

D. *Evaluation of the Proposed CDSS*

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu, accessed September 2, 2010). Primary goal of ADNI has been to measure the progression of mild cognitive impairment (MCI) and early AD using biomarkers, and clinical and neuropsychological assessment. MCI is a heterogeneous state of cognitive decline, with multiple possible outcomes and increased risk of AD [22]. ADNI recruited approximately 400 people with MCI to be followed for 3 years, in addition to recruiting 200 normal elderly individuals and 200 AD patients.

From the MCI patients recruited to ADNI, this study included those whose last clinical diagnosis during the study was still MCI or had converted to AD, forming the classification groups of stable MCI (SMCI, $n = 190$) and progressive MCI (PMCI, $n = 154$, average time to getting AD diagnosis: 19 months), respectively. Using baseline measurements alone, we tested our method's ability to predict conversion to AD using sparse multiscale measurement data that included neuropsychological tests, magnetic resonance imaging data, molecular test data, and genetic test data (see Table II).

The ability of DSI to predict AD was compared to three reference classifiers; support vector machine (SVM), Naïve Bayes, and Logistic Regression (LR). All methods were given exactly the same data. Data preprocessing, parameter search, and feature selection was done for the reference classifiers to attain the best performance possible. The generic DSI method has been

TABLE II
ENTRIES AND THEIR FEATURE COUNTS FOR EARLY DIAGNOSIS OF AD

Entry	Features	Available[c]	Description
Neuropsychological Tests			
MMSE	30	100 %	Mini-Mental State Examination
ADAS	13	99 %	AD Assessment Scale
Biomedical Imaging			
MRI	13	89 %	Volumes of structures from brain MRI[a]
Molecular Tests			
CSF	2	52 %	Amyloid-β and Total Tau from CSF[b]
Genetic Tests			
APOE	2	100 %	Alleles of apolipoprotein E

[a]Magnetic resonance imaging, [b]Cerebrospinal fluid, [c]Percentage of patients for whom the data existed in the ADNI database at baseline.

TABLE III
CLASSIFICATION PERFORMANCE WITH ADNI MCI DATASET

Method	AUC[a]	Accuracy	Sensitivity	Specificity
DSI	0.75 ± 0.08	0.68 ± 0.08	0.70 ± 0.12	0.66 ± 0.10
SVM	0.75 ± 0.08	0.67 ± 0.07	0.64 ± 0.11	0.69 ± 0.11
Bayes	0.76 ± 0.08	0.67 ± 0.07	0.65 ± 0.12	0.69 ± 0.11
LR	0.69 ± 0.09	0.62 ± 0.07	0.73 ± 0.10	0.53 ± 0.11

Table shows means and standard deviations (SD) over ten iterations of 10-fold cross-validation. [a]Area under curve from receiver operating characteristic (ROC).

designed not to require preprocessing of any kind, and was used as such. Ten iterations of 10-fold cross-validation were done to obtain robust performance metrics.

In addition to the MCI dataset, we tested the DSI method with three other medical datasets (Pima Indian Diabetes, Cleveland Heart Disease, and Hepatitis) available online [23]. Performance with these datasets was compared to publicly available benchmark results [24]. It is not possible to make a completely objective comparison between the benchmark values and the DSI method since many of the reported values are expressed only as a single number giving the classification accuracy, without standard deviation or information about the validation process. Also, some benchmark results were computed only after excluding subjects with missing values. To robustly assess the DSI method, ten iterations of 10-fold cross-validation were performed with all available data and compared against the best benchmark method whose standard deviation was available, and against the average of benchmark methods that performed better than a simple majority classifier, i.e., one that assigns every case to whichever class is in the majority in the training set.

Applicability of the software library was demonstrated by developing a CDSS tool for early prediction of AD. The complexity of implementation work was evaluated qualitatively and the computational performance of the interactive DSI method was measured quantitatively on a laptop PC with Windows XP SP3, 2 GB of memory, and a 2.4 GHz dual core processor.

III. RESULTS

A. Classification Performance

With the MCI dataset from ADNI, the DSI method performed on a level similar to established machine learning methods, as seen in Table III.

Results obtained with other medical datasets show that the DSI method tends to perform slightly worse than the best benchmark methods, but similar to the average of them. With the diabetes dataset, the best benchmark method was SVM. For heart disease data, the maximum was obtained with a 28-nearest neighbors (k-NN) classifier, using Euclidean distance, and trained only with a subset of features. With the hepatitis dataset, accuracy was best with an 18-NN classifier, this time using Manhattan distance. Results of these evaluations are listed in Table IV.

TABLE IV
CLASSIFICATION ACCURACY WITH BENCHMARK DATASETS

Dataset	Controls / Positives	DSI[a]	Benchmark Maximum[b]	Benchmark Average[c]
Diabetes	500/268	0.75±0.04	0.78±0.04	0.74±0.03
Heart disease	164/139	0.81±0.06	0.85±0.01	0.79±0.06
Hepatitis	123/32	0.84±0.08	0.90±0.01	0.85±0.04

Table shows class counts, or means and standard deviations (SD) of classification accuracy from ten iterations of 10-fold cross-validation[a], 10-fold cross-validation[b], or all methods beating the majority class classifier[c].

B. Implementing the CDSS Tool

Relying on the generic decision support library for much of the necessary functionality, a prototype of a CDSS tool for early prediction of AD was developed. The prototype uses two persistence stores that connect to local databases, one using EAV scheme that provides MCI patients for analysis, and a normalized database for accessing training data. Definitions of entries and feature values are described in an XML file and provided to the decision support library during initialization of the application.

The tool provides a comprehensive overview of all available patient data to clinicians. GUI components from the library visualize entries, the DSI tree, and data distributions on a single screen. The patient details panel and the rendering of brain MRI images were custom built for this application. From the user interface, clinicians can select entries to see the data in more detail, select nodes from the DSI tree to see patient and training data distributions, change classification groups, and change included features to customize classification. In summary, the tool allows mining of multiscale patient data and evidence-based study of their relation to known Alzheimer's disease profiles.

The software library facilitated rapid implementation of the CDSS prototype. Taking it into use required configuration of persistence stores and data definitions, providing the necessary sampling and grouping rules, and finally wiring the GUI components into the application.

C. Computational Performance of the DSI Implementation

Training of the DSI model and computation of the initial set of DSI values was done for all patients sequentially, taking on average 860 ms/patient (standard deviation 74 ms). Re-evaluation of DSI values after user initiated exclusion or inclusion of a feature was virtually instantaneous, consistently taking less than 1 ms. Grouping and sampling of training data, including the necessary queries to the database, took on average 10 s. This is done only once after the application launches, but could be performed again if the training data are changed while running the application.

IV. Discussion

To the authors' knowledge, there are no other CDSS tools or decision support libraries for clinical diagnostics developed with a similar philosophy, i.e., using any available sparse and unprocessed patient data, and not requiring manual tuning or decision parameters defined by clinical experts. To use the decision support system presented here, one only needs data definitions, which can in several cases be derived in a straightforward manner, using the structure of the original data. Data hierarchy definitions can be modified manually if a particular organization is preferred. Computer-based methods for organizing the data hierarchy could also be developed, possibly grouping features automatically along the dimensions of a disease, e.g., effect to motor dysfunction or to delayed recall performance. Further studies are required to assess the effect of different hierarchy structures on the classification accuracy of the statistical DSI method behind the library.

The generic clinical decision support library was found to be a good basis for developing a CDSS tool for early diagnosis of AD. Features of the library aim to support clinical requirements, e.g., they accommodate workflows where patient data are collected sporadically. The statistical methods are not computationally intensive, and could be further optimized with parallelization. Computational performance of the decision support library is more limited by access to training data. Retrieving bulk patient data for training sets in a generic manner was made feasible by developing tools and defining processes that can be used for creating and populating normalized databases from existing electronic datasets.

The DSI method behind the decision support library was able to provide values for quickly interpretable visualizations of multiscale data without compromising prediction accuracy. The visualizations were designed to be transparent, i.e., to clearly disclose the origin of the derived values, since even accurate diagnostics obtained with a black box classifier are not very easily applied in clinical practice. Compared to the reference classification methods, the DSI also emphasizes clinical interpretability by 1) providing information about all subsystems of different scales (e.g., genetic, molecular, structural, and neuropsychological) individually and also as a part of the whole, 2) computing a rank of the patient data in relation to diagnosed populations instead of maximizing class separation, which leads to 3) consistency in output that should reflect the magnitude of changes in the raw data. In addition to highlighting important details to clinicians, the DSI and *relevance* values can facilitate building of expert systems.

Classification accuracy of the DSI was found comparable to benchmark methods when applied to various medical datasets, even though it is designed not to require feature selection or searching of optimal classifier parameters. In other words, the generic DSI method obtained classification accuracies close to the best benchmark results, which were manually tuned to work with the given data as well as possible. The relatively low classification accuracies with MCI data are in line with other studies [25] and underline the fact that data alone are not enough for reliable prediction of conversion from MCI to AD at an early phase of the disease. This is also true for ADNI data, partly due to a relatively short follow-up time and also due to errors in the diagnoses which have not been confirmed pathologically. Correlation between features was also considered. It appears that the tree hierarchy and the recursion resulting from it partially nullify issues due to correlation. For datasets with a large number of features, we have implemented a method that explicitly addresses correlation by applying principal component analysis (PCA) to the leaf nodes of the data hierarchy. In the evaluation datasets, this did not, however, increase classification accuracy.

Healthcare is slowly moving towards electronic health records. Eventually, patient data could be automatically loaded for analyses inside a tool such as this. A clinician diagnosing a patient would not need to observe hundreds of individual measurements at different scales, available from several sources. Instead, they could see all available data at once, hypothesize a disease, and immediately see which data are relevant in that context and which point toward the disease. This could save both time and frustration from information overload. For now, manual work is needed, either entering patient records into the tool, implementing a custom persistence store implementation, or implementing a data adapter which reads existing electronic sources of a particular clinic into a database supported by the library. This limits the presented solution to specialist clinics in the immediate future. The authors also acknowledge that routinely collected clinical data contain more artifacts and missing information than research data that affect the performance of the methods. Therefore, there are plans for future studies using less well-curated patient data from realistic sources.

The main disadvantage of the presented DSI method and the decision support library implementation is that in addition to the patient measurements for analyses, they require properly validated datasets for control and disease cases. This training data could be local to a particular clinic, but could also be collected regionally or nationally, greatly decreasing the burden of creating validated training datasets. The authors believe that data obtained in research studies should be a good starting point for compiling the initial training datasets.

Another limitation of the proposed system is that currently the library has proper support for two-class problems only. Future research will address how these methods are appropriately applied when multiple diseases are in consideration, which is a clinically important requirement for differential diagnostics.

V. Conclusion

In this manuscript, the design and implementation of a generic decision support system was presented. It is implemented as a reusable software library employing a statistical disease state modeling method, which is able to robustly analyze heterogeneous multiscale patient data with minimal preprocessing. The context-agnostic data access, analysis, and visualization methods allow the library to be rapidly applied in several contexts. When presented with a new problem or data, there is no searching of parameters, handling of missing values, or development of new user interfaces. As long as definitions of the data and the data itself are provided to the library, it can organize available values and construct interactive views that provide analyses of the recently defined information to clinical decision makers. The ultimate goal is to provide evidence-based decision support for clinicians during diagnostic work. Application of the decision support library was demonstrated by developing a prototype CDSS tool for early prediction of AD. We are currently evaluating the prototype at two memory clinics in Europe, comparing it to traditional diagnostic methods. We are also applying the DSI method and the decision support library to several other datasets to assess their robustness more comprehensively.

Acknowledgment

The authors thank participants of project PredictAD, funded partially by the 7th Framework Program by the European Commission under the ICT theme Virtual Physiological Human (Grant Agreement 224328). Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wpcontent/uploads/how_to_apply/ADNI_Authorship_List.pdf.

References

[1] H. Chen, S. Fuller, C. Friedman, and W. Hersh, *Medical Informatics: Knowledge Management and Data Mining in Biomedicine*. New York: Springer, 2010.

[2] J. Southern, J. Pitt-Francis, J. Whiteley, D. Stokeley, H. Kobashi, R. Nobes, Y. Kadooka, and D. Gavaghan, "Multi-scale computational modelling in biology and physiology," *Prog. Biophys. Mol. Biol.*, vol. 96, pp. 60–89, 2008.

[3] D. Noble, "Modeling the heartFrom genes to cells to the whole heart," *Science*, vol. 295, pp. 1678–1682, 2002.

[4] T. S. Deisboeck and G. S. Stamatakos, *Multiscale Cancer Modeling*. London: CRC Press, 2010.

[5] E. Alpaydin, *Introduction to Machine Learning*, 2^{nd} ed. Cambridge: MIT Press, 2009.

[6] A. Madabhushi, S. Agner, A. Basavanhally, S. Doyle, and G. Lee, "Computer-aided prognosis: Predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data," *Computerized Med. Imaging and Graphics*, to be published.

[7] J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, M. Bae, R. Janardan, H. Liu, G. Alexander, and E. Reiman, "Heterogeneous data fusion for alzheimer's disease study," in *Proc. 14th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, Las Vegas, 2008.

[8] C. Bennett and T. Doub, "Data mining and electronic health records: Selecting optimal clinical treatments in practice," in *Proc. 2010 Int. Conf. Data Mining*. Las Vegas, Jul. 2010.

[9] R. A. Greenes Ed., *Clinical Decision Support the Road Ahead*. New York: Elsevier, 2007.

[10] J. C. Horrocks, A. P. McCann, J. R. Staniland, D. J. Leaper, and F. T. de Dombal, "Computer-aided diagnosis: Description of an adaptable system, and operational experience with 2, 034 cases," *British Med. J.*, vol. 2, pp. 5–9, 1972.

[11] G. O. Barnett, J. J. Cimino, J. A. Hupp, and E. P. Hoffer, "DXplain. An evolving diagnostic decision-support system," *The J. Amer. Med. Assoc.*, vol. 258, pp. 67–74, 1987.

[12] I. Watson and F. Marir, "Case-based reasoning: A review," *The Knowledge Eng. Rev.*, vol. 9, pp. 327–354, 1994.

[13] C. D. Stylios, V. C. Georgopoulos, G. A. Malandraki, and S. Chouliara, "Fuzzy cognitive map architectures for medical decision support systems," *Appl. Soft Comput.*, vol. 8, pp. 1243–1251, Jun. 2008.

[14] J. Mattila, J. Koikkalainen, A. Virkki, A. Simonsen, M. van Gils, G. Waldemar, H. Soininen, and J. Lötjönen, "A disease state fingerprint for evaluation of alzheimer's diseases," *The J. Alzheimer's Dis.*, to be published. Available: http://iospress.metapress.com/content/kg54325631131n10/.

[15] P. Beck, T. Truskaller, I. Rakovac, B. Cadonna, and T. R. Pieber, "On-the-fly form generation and on-line metadata configuration—A clinical data management Web infrastructure in Java," *Stud. Health Technol. Inform.*, vol. 124, pp. 271–276, 2006.

[16] C. A. Brandt, A. M. Deshpande, C. Lu, G. Ananth, K. Sun, R. Gadagkar, R. Morse, C. Rodriguez, P. L. Miller, and P. M. Nadkarni, "TrialDB: A web-based clinical study data management system," in *AMIA Annu. Symp. Proc.*, 2003, p. 794.

[17] P. M. Nadkarni, C. Brandt, S. Frawley, F. G. Sayward, R. Einbinder, D. Zelterman, L. Schacter, and P. L. Miller, "Managing attribute—Value clinical trials data using the ACT/DB client-server database system," *J. Am. Med. Inform. Assoc.*, vol. 5, no. 2, pp. 139–151, 1998.

[18] R. S. Chen, P. M. Nadkarni, L. Marenco, F. W. Levin, J. Erdos, and P. L. Miller, "Exploring performance issues for a clinical database organized using an entity-attribute-value representation," *J. Amer. Med. Inf. Assoc.*, vol. 7, pp. 475–487, 2000.

[19] J. Lerman, *Programming Entity Framework: Building Data Centric Apps with the ADO.NET Entity Framework.*, Sebastopol, CA, O'Reilly Media, 2010.

[20] J. Albahari and B. Albahari, *C# 4.0 in a Nutshell: The Definitive Reference*, Sebastopol, CA, O'Reilly Media, 2010.

[21] P. Vogel, *Practical Code Generation in .NET: Covering Visual Studio 2005, 2008, and 2010 (Addison-Wesley Microsoft Technology Series)*. Boston, MA: Pearson Education Inc., pp. 249–284, 2010.

[22] R. C. Petersen, R. O. Roberts, D. S. Knopman, B. F. Boeve, Y. E. Geda, R. Ivnik, G. Smith, and C. R. Jack, "Mild cognitive impairment: Ten years later," *Arch. Neurol.*, vol. 66, pp. 1447–1455, 2009.

[23] A. Frank and A. Asuncion, *UCI Machine Learning Repository*, [Online] Available: http://archive.ics.uci.edu/ml, 2010.

[24] W. Duch, *Comparison of Classification Results*, [Online] Available: http://www.is.umk.pl/projects/datasets.html, 2011.

[25] D. A. Llano, G. Laforet, and V. Devanarayan, "Derivation of a new ADAS-cog composite using tree-based multivariate analysis: prediction of conversion from mild cognitive impairment to Alzheimer disease," *Alz. Dis. Assoc. Dis.*, vol. 25, pp. 73–84, 2011.

Authors' photographs and biographies not available at the time of publication.

PUBLICATION III

Quantitative evaluation of disease progression in a longitudinal mild cognitive impairment cohort

In: Journal of Alzheimer's Disease 39(1),
pp. 46–61.

Journal of Alzheimer's Disease 39 (2014) 49–61
DOI 10.3233/JAD-130359
IOS Press

Quantitative Evaluation of Disease Progression in a Longitudinal Mild Cognitive Impairment Cohort

Hilkka Runtti[a,*], Jussi Mattila[a], Mark van Gils[a], Juha Koikkalainen[a], Hilkka Soininen[b], Jyrki Lötjönen[a] and for the Alzheimer's Disease Neuroimaging Initiative
[a] *VTT Technical Research Centre of Finland, Tampere, Finland*
[b] *Department of Neurology, University of Eastern Finland and Kuopio University Hospital, Kuopio, Finland*

Handling Associate Editor: Javier Escudero

Accepted 20 August 2013

Abstract. Several neuropsychological tests and biomarkers of Alzheimer's disease (AD) have been validated and their evolution over time has been explored. In this study, multiple heterogeneous predictors of AD were combined using a supervised learning method called Disease State Index (DSI). The behavior of DSI values over time was examined to study disease progression quantitatively in a mild cognitive impairment (MCI) cohort. The DSI method was applied to longitudinal data from 140 MCI cases that progressed to AD and 149 MCI cases that did not progress to AD during the follow-up. The data included neuropsychological tests, brain volumes from magnetic resonance imaging, cerebrospinal fluid samples, and apolipoprotein E from the Alzheimer's Disease Neuroimaging Initiative database. Linear regression of the longitudinal DSI values (including the DSI value at the point of MCI to AD conversion) was performed for each subject having at least three DSI values available (147 non-converters, 126 converters). Converters had five times higher slopes and almost three times higher intercepts than non-converters. Two subgroups were found in the group of non-converters: one group with stable DSI values over time and another group with clearly increasing DSI values suggesting possible progression to AD in the future. The regression parameters differentiated between the converters and the non-converters with classification accuracy of 76.9% for the slopes and 74.6% for the intercepts. In conclusion, this study demonstrated that quantifying longitudinal patient data using the DSI method provides valid information for follow-up of disease progression and support for decision making.

Keywords: Alzheimer's disease, biomarkers, data mining, decision support techniques, early diagnosis, mild cognitive impairment

INTRODUCTION

Alzheimer's disease (AD) is a neurodegenerative disease that develops gradually over the years and finally results in loss of cognitive function and dementia [1]. Mild cognitive impairment (MCI) is an intermediate state between normal cognition and dementia. Patients with MCI have cognitive problems that are not normal for their age and do not yet interfere with their daily activities [2–4]. MCI with memory dysfunction is a risk factor for AD, however, not all MCI patients will progress to AD [2, 3].

There is no cure for AD, but it has been modeled that delaying the onset of the disease would reduce its prevalence considerably, and slowing down its progression would allow more cases to remain as mild AD instead of progressing to moderate or severe AD which

[*] Correspondence to: Hilkka Runtti, VTT Technical Research Centre of Finland, P.O. Box 1300, FIN-33101 Tampere, Finland. Tel.: +358 40 152 6627; Fax: +358 20 722 3499; E-mail: hilkka.runtti@vtt.fi.

ISSN 1387-2877/14/$27.50 © 2014 – IOS Press and the authors. All rights reserved

causes huge costs to society [5]. Different treatments to modify disease progression have been studied [6, 7] and it has been shown that they should be started as early as possible to be effective [7, 8]. To make earlier AD diagnosis and interventions feasible, different neuropsychological tests and biomarkers from laboratory tests and imaging have been studied extensively [9–12].

In 2010, Jack et al. [13] proposed a model describing temporal evolution of major AD biomarkers. The model was recently updated on the basis of gained knowledge, and according to it, different biomarkers of AD become abnormal in a certain temporal order and their longitudinal behavior is non-linear [14]. Biomarkers measuring deposition of amyloid-β plaques become abnormal first, years before the clinical symptoms appear. They are followed by indicators of neurodegeneration, and the last biomarkers to become abnormal are structural changes visible in magnetic resonance imaging (MRI) and changes in cerebral metabolism revealed by fluorodeoxyglucose positron emission tomography (FDG-PET). The updated model also takes into account that the severity of cognitive impairment due to pathophysiological load of AD is individual depending on, e.g., genetics, lifestyle, and other brain diseases.

New guidelines, incorporating both cognitive assessment and biomarkers for diagnosing different stages of AD, were recently published as a result of these research findings [15–18]. They state that the detection of preclinical stages of AD in research subjects should be based on biomarkers and that MCI and AD are diagnosed using clinical and cognitive evaluation and biomarkers can provide complementary information.

All the different tests and investigations done in modern diagnostics produce large amounts of data that clinicians need to explore carefully. Assessing the heterogeneous data and measuring longitudinal changes in them may be difficult. Several studies have successfully combined multimodal data to classify subjects into classes of healthy, MCI, or AD using established classification methods, e.g., logistic regression or support vector machines [19–24]. There also exists a statistical Disease State Index (DSI) method which estimates the state of a patient in the continuum from healthy to disease on the basis of measured data. The DSI method has been developed and extensively studied by most of the authors of this manuscript. Mattila et al. [22] demonstrated that it discriminated well between healthy cases, MCI cases that do not convert to AD, MCI cases that convert to AD, and AD cases. A recent study, also by Mattila et al. [25], showed that approximately half of the MCI patients who developed into AD could have been classified with a high accuracy already a year before receiving the clinical diagnoses using the DSI. However, it has not been studied yet how DSI values develop over time in subjects with MCI.

DSI values can be visualized with a Disease State Fingerprint (DSF) technique which shows how results from different tests contribute to the disease state of a patient. The DSF allows rapid interpretation of large amounts of patient data and helps clinicians to discern relevant information from irrelevant [22]. Until now, only data from a single time point have been visualized using the DSF.

The objective of this work was to study disease progression quantitatively using heterogeneous longitudinal data in an MCI cohort. First, it was studied whether it is possible to discern significant trends in the severity of AD as reflected by the DSI and whether subjects that convert from MCI to AD have a different longitudinal DSI behavior than subjects that do not convert. Second, classification of MCI subjects to converters and non-converters on the basis of the trend parameters from longitudinal DSI values was tested. Third, to facilitate interpretation of data, the DSF visualization was developed further for the presentation of longitudinal data.

MATERIALS AND METHODS

Study population

Data used in the analyses were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [26]. ADNI is a 5-year study aiming at developing and testing methods for acquiring and analyzing biological markers that measure the progression of MCI and AD [27]. ADNI was launched in 2004, and approximately 800 subjects of age 50 to 90 years have been recruited at around 50 sites in the United States and Canada. The enrolled subjects included 200 healthy elderly controls, 400 subjects with MCI, and 200 subjects with early AD. The subjects underwent cognitive assessment, neuropsychological testing, and MRI at intervals of six or twelve months for two to four years. Other tests, such as FDG-PET and blood and cerebrospinal fluid samples (CSF), were performed less frequently [28].

In the present study, MCI cases with at least 24 months of follow-up data were included. The selected MCI cases were divided into two groups: a stable

Table 1
Demographics of the study population at the baseline

	Stable MCI	Progressive MCI	*p*
Subjects	149 (51.6%)	140 (48.4%)	
Gender			0.373
Female	51 (34.2%)	55 (39.3%)	
Male	98 (65.8%)	85 (60.7%)	
Age (years)	75.1 ± 7.4	75.4 ± 6.7	0.916
Education (years)	15.9 ± 3.0	15.6 ± 3.0	0.239

Data presented as number of subjects (percentage of subjects %) or mean ± standard deviation. p: Group differences were examined using appropriate tests based on whether their distribution was normal or not as determined by the Kolmogorov-Smirnov test: Pearson χ^2 test (gender) and Mann-Whitney U test (age and education).

MCI group (SMCI, $n = 149$), who did not obtain the diagnosis of AD during the follow-up period, and a progressive MCI group (PMCI, $n = 140$), whose diagnosis changed from MCI to AD during the follow-up. Subjects whose diagnosis changed from MCI to healthy or from MCI to AD and then back to MCI were excluded from the study. Demographics of these two groups are presented in Table 1.

The data were downloaded from the ADNI website (http://adni.loni.ucla.edu) in September 2011. The data used in the analyses comprised Mini-Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale-cognitive subscale (ADAS), Neuropsychological Battery (NeuroBat), brain volume measures based on MRI, amyloid-β and total tau in CSF, and apolipoprotein E (APOE). Details of the included variables are presented in the Supplementary Material. MRI brain volume measures provided to ADNI by Anders Dale Lab (University of California, San Diego) were used. They performed volumetric segmentation of MRI with the FreeSurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). Technical details of the segmentation are described in [29].

Diagnosis of MCI and AD in the ADNI is based on evaluation of memory, cognition, and functional performance (memory complaints by a subject or a study partner, Logical Memory II, MMSE, and Clinical Dementia Rating) [28]. In addition, diagnosis of probable AD requires fulfillment of the AD criteria defined by the NINCDS-ADRDA (the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association) [30, 31]. Although the diagnosis is partly based on MMSE and Logical Memory II, they were included in the data analyses in this study because 1) MMSE is widely used making it interesting in clinical sense, 2) the diagnosis is not based only on the MMSE and Logical Memory II, and 3) the ADNI criteria to decide between MCI versus AD does allow overlap in MMSE score and Logical Memory II score.

Variables summarizing the tests, e.g., total MMSE score and ADAS 13 point total, were excluded as independent variables from the analysis because the subscores and the individual items contain the same information as the total scores. Justification for the use of individual items instead of total scores is that some items may differentiate between SMCI and PMCI cases better than others and part of the available information is lost if only the total scores are used. For example, Llano et al. [32] weighted individual items of ADAS with coefficients derived using data-driven approach and constructed a new composite ADAS score. Their composite score differentiated normal controls, MCI, and AD cases better than the ADAS total score and the composite score also predicted conversion to AD slightly better than the ADAS total score.

Disease State Index

The DSI is a statistical method for deriving a scalar value that estimates the state of a disease in a patient [22]. The DSI method is based on the computation of two different values: DSI values and relevance values. The DSI value of an individual variable is computed by comparing a measurement value from a patient to the distributions of known healthy and diseased cases using a so-called fitness function. DSI values are between zero and one, with higher values indicating that the patient fits better to the disease than to the control population on the basis of the measured data. The relevance value describes how well the variable differentiates between the known healthy and diseased cases. In other words, relevance is a measure of the differences in the data measured from healthy and diseased cases. Relevance values, like the DSI values, are also between zero and one, with higher values representing better discrimination. A composite DSI combining different variables is computed as a weighted arithmetic mean of the individual DSI values weighted by the relevance values. This averaging is done several times recursively to yield a hierarchy of DSI values that reveals the overall position or rank in relation to the disease, i.e., quantifies the progression of a disease based on available patient data. In this work, the study population consisted of SMCIs as control cases and PMCIs as disease cases.

The DSI method is robust against overfitting by its design. Estimation of the DSI and relevance values

Table 2
Number of available patient visits at different time points

	Baseline	Month 6	Month 12	Month 18	Month 24	Month 30	Month 36	Month 42	Month 48
Total	289	287	287	279	281	0	233	0	51
SMCI	149	148	147	143	142	0	121	0	19
PMCI	140	139	140	136	139	0	112	0	32

SMCI, stable mild cognitive impairment; PMCI, progressive mild cognitive impairment.

for individual variables is done independently from other variables, thus, there is no over-dimensionality at the variable level because only two parameters are estimated for each variable (the DSI value and the relevance value). In addition, weighting of features and the use of the hierarchy lead in practice to feature selection. As a result, any few values alone will not determine the resulting composite DSI value, but it is an amalgam of all relevant data sources. Mathematical details of the computation of the DSI and relevance values are explained in [22].

The DSI values can be calculated on the basis of a single variable or multiple variables together. In this study, it was investigated whether combining different data modalities would yield better results than utilizing data from a single modality alone. Thus, DSI values were calculated using two different approaches: 1) using all available variables together (MMSE, ADAS, NeuroBat, MRI, CSF, and APOE) and 2) using data from individual data modalities independently (MMSE, ADAS, NeuroBat, and MRI). CSF was measured less frequently so it was not analyzed individually and neither was APOE genetics, which do not change with disease progression. For the calculation of the DSI values, subjects were divided into ten training and test sets for stratified 10-fold cross-validation in which each fold contains the same proportions of class labels. The training data used for building the model of AD progression included actual measurement values from SMCI baseline visits and actual measurement values from the time of receiving AD diagnosis for PMCI cases. This kind of selection of training data sets the dynamic range of the DSI method between SMCIs at the baseline and early AD, i.e., the dynamic range of the DSI method was optimized for the purposes of the study and clinical problem at the hand. The test sets included data from the complete series of visits of the remaining SMCI and PMCI cases. The number of patient visits available at the different time points is shown in Table 2. Missing values in the raw data (e.g., a missing result in MMSE) were replaced with the values from the patient's previous available visit. This allowed having complete data sets for the analysis at each patient visit. Although using previous data can result in slightly outdated data and conservative disease progression estimates for some patient visits, that data were known to have been available at those time points.

Disease State Fingerprint

The DSF is a method for visualizing the patient data and the hierarchy of the DSI values [22]. Example visualizations are shown in the left panel of Fig. 1. DSF consists of a tree with nodes of different sizes and colors. The size of the node indicates the relevance value, i.e., how well a variable or a test differentiates between SMCI and PMCI, and color indicates the DSI value. Higher DSI values refer to PMCI and result in shades of red. Lower values represent SMCI and result in shades of blue. In this study, the progression of AD was visualized using the DSF technique extended with support for longitudinal data.

Synchronization of the time stamps

The initial visits of MCI patients to a memory clinic occurred in different phases of the disease. For example, some PMCI cases converted from MCI to AD at follow-up month 6 and others at month 36. To take this into account, the time stamps of the patient visits were synchronized. The moment of receiving AD diagnosis was set as the zero time point (Z) of PMCIs. For SMCIs, the last available time point up to month 36 was set as their Z. The time points preceding the zero point were labeled as Z-6, Z-12, etc. DSI values from Z-42 and Z-48 months were excluded from the analysis because they contained only a few cases. Thus, DSI values computed from visit data at Z, Z-6, Z-12, Z-18, Z-24, Z-30, and Z-36 months were used in the analysis. Only those subjects who had at least three DSI values available in all approaches (DSI calculated using all variables, MMSE, ADAS, NeuroBat, or MRI), were included for further analysis. The purpose was to perform linear regression (see below) and using only two points would have yielded in perfect regression, making the comparison of goodness of fit values between the different datasets unfair. The number of available

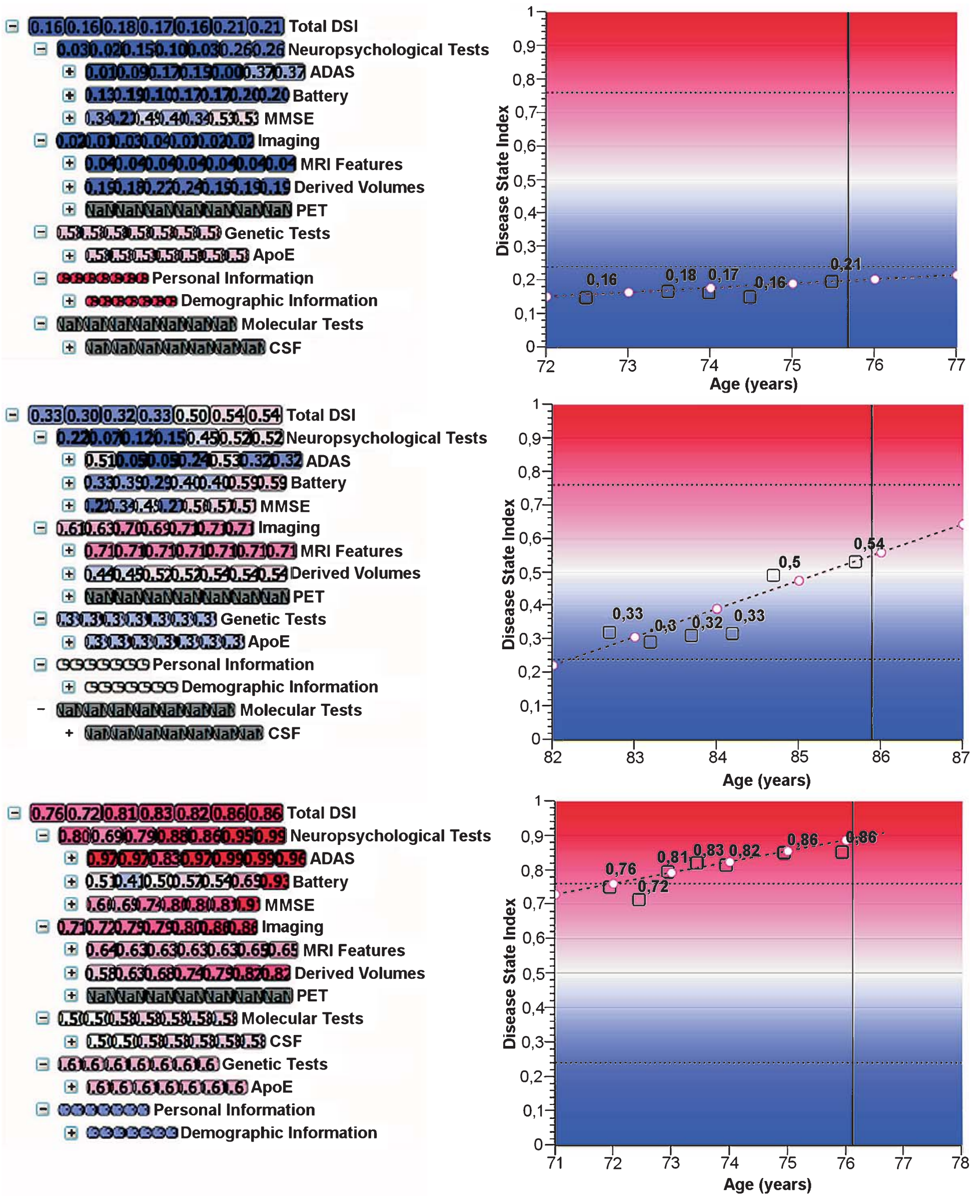


Fig. 1. Visualizations of three sets of longitudinal patient data. Left panel: Disease State Fingerprints (DSF) in which Disease State Index (DSI) values of the individual tests at different time points are shown on the rows. Total DSI values (the topmost rows of the DSFs) combines the results from the individual tests. Sizes of the boxes indicate how well the variable discriminates between the stable (SMCI) and progressive (PMCI) mild cognitive impairment cases. Color indicates to which group the data fits the best. Blue corresponds to SMCI and red to PMCI. Right panel: linear regression of the total DSI values (red dashed line with white circles). Black squares present the total DSI values of a patient. The horizontal lines indicate a threshold where the classification accuracy of 85% is achieved. The vertical line shows the current age of a patient. Data from two SMCI cases are presented in the topmost panels and data from a PMCI case is presented in the lowest panel.

Table 3
Number of Disease State Index values of the SMCI and PMCI cases at synchronized time points. The last available time point up to month 36 was selected as the zero time point (Z) of SMCI. The moment of receiving Alzheimer's disease diagnosis was set as the Z of PMCI. The time points preceding the Z were labeled as Z-6 etc

	Z-36	Z-30	Z-24	Z-18	Z-12	Z-6	Z
SMCI	147	147	147	147	147	147	147
PMCI	29	29	64	90	126	126	126

SMCI, stable mild cognitive impairment; PMCI, progressive mild cognitive impairment. The number of SMCI cases stays the same because the visit Z-36 is their baseline visit and any missing values have been replaced with the values from the previous available visit. The number of PMCI cases changes over time because some have converted in an early phase of the study. Only the cases having at least three available DSI values were included.

DSI values of the included SMCI and PMCI cases at the synchronized time points is presented in Table 3.

Modeling progression of AD

In this work, it was assumed that the change of the DSI values over time, and thus the progression of AD, can be modeled linearly:

$$DSI = a * t + b \tag{1}$$

where a is the slope of regression (rate of change for DSI values), b is the intercept of regression (DSI value at the time point zero), and t is time measured in months. A linear model was selected because it is the simplest method to model the progression of AD and it is also the simplest to interpret. Another reason was that due to the synchronization of the time stamps some subjects had only few DSI values available for the regression. Thus, there were not enough data points for more complicated models. The third reason supporting the linear model was that the follow-up times were relatively short compared with the time span of disease progression in AD in overall. Linear regression was performed for each subject separately to model each individual's disease progression.

Differentiation using the trend parameters

Classification of subjects as SMCI or PMCI cases on the basis of their regression parameters (slope, intercept) was studied as follows. First, optimal classification thresholds for the regression parameters were defined on the basis of the receiver operating characteristic (ROC) curves. Then, the regression parameters were compared to the threshold value and if it was exceeded the subject was classified as PMCI. Otherwise he or she was classified as SMCI. The thresholds and classification performance measures (classification accuracy, sensitivity, and specificity) were calculated using the stratified 10-fold cross-validation.

Statistical methods

Normality of the continuous demographic variables was studied using Kolmogorov-Smirnov test. Group differences in demographics between SMCI and PMCI groups were examined using non-parametric Mann-Whitney U test for continuous variables and Pearson χ^2 test for categorical variables.

Linear regression was performed using the longitudinal DSI values which were derived using 1) all available variables together (total) and 2) data from individual tests separately. Goodness of fit of the linear regression using 1) and 2) was compared using R^2, adjusted R^2, and mean square errors. Residuals of the regression were also examined using histograms and by plotting residuals versus predicted values. The regression parameters of the SMCI and PMCI groups were compared to zero using one-sample Wilcoxon Signed Rank test and the differences between the groups were studied using Mann-Whitney U test.

Normality of the regression parameters was studied using histograms. On the basis of the initial histogram analysis, it appeared that the slopes of the SMCI group may have a bimodal distribution. Fits of unimodal and bimodal distributions were compared and details of these analyses are explained in the Supplementary Material.

Subjects were classified as SMCIs or PMCIs on the basis of their regression parameters. Classification performance was measured using the area under the ROC curve (AUC), classification accuracies, sensitivities, and specificities. To study whether using all data modalities together would yield in significantly greater classification performance than using only a single data modality, classification accuracies of the individual tests were compared to the classification accuracies derived using all data. Thus, four comparisons with both the slopes and the intercepts (total-MMSE, total-ADAS, total-NeuroBat, total-MRI) were performed. The classification accuracies of the slopes and the intercepts derived using all data were also compared. Paired samples t-test was used if the classification accuracies were normally distributed according to Kolmogorov-Smirnov test, otherwise, related-samples Wilcoxon Signed Rank test was performed. In all analyses, $p < 0.05$ was considered significant. In pairwise comparisons of classification accuracies, Bonferroni

Table 4
Goodness of fit for the linear regression of longitudinal Disease State Index values derived using different data modalities

Dataset	R^2	Adjusted R^2	Mean square error
Total	0.553 ± 0.289	0.422 ± 0.369	0.006 ± 0.008
MMSE	0.364 ± 0.295	0.172 ± 0.390	0.014 ± 0.016
ADAS	0.388 ± 0.298	0.196 ± 0.413	0.024 ± 0.026
NeuroBat	0.475 ± 0.318	0.315 ± 0.426	0.005 ± 0.004
MRI	0.721 ± 0.259	0.642 ± 0.321	0.001 ± 0.001

Total, All available variables included when calculating DSI values; MMSE, Mini-Mental State Examination; ADAS, Alzheimer's Disease Assessment Scale-cognitive subscale; NeuroBat, Neuropsychological Battery; MRI, brain volumes derived from magnetic resonance imaging. The values are mean ± standard deviation because the linear regression was performed for each subject independently.

Table 5
Regression parameters of longitudinal Disease State Index values for SMCI and PMCI groups

	SMCI	PMCI
Slope*	0.002 (0.000, 0.006)$^+$	0.010 (0.005, 0.015)$^+$
Intercept*	0.295 (0.139, 0.621)$^+$	0.754 (0.626, 0.860)$^+$
n	7 (7; 7)	5 (3; 5)

Values are median (25th percentile, 75th percentile). SMCI, stable mild cognitive impairment; PMCI, progressive mild cognitive impairment, n, number of points in the regression, *statistically significant difference between the groups (Mann-Whitney U test, $p < 0.0005$), $^+$significantly different from zero (one-sample Wilcoxon Signed Rank test, $p < 0.0005$). Disease State Index values were derived using all variables together.

correction was applied and $p < 0.0056$ was considered significant (number of comparisons was nine).

All analyses were performed in Matlab R2012a (The Mathworks, Natick, MA) and IBM SPSS Statistics 19 (IBM, Armonk, NY). Visualizations were processed in GNU Image Manipulation Program 2.0 (GIMP 2.0, freely available at http://www.gimp.org/).

RESULTS

Modeling progression of AD

Goodness of fit for linear regression of the longitudinal DSI values is shown in Table 4. On the basis of R^2, adjusted R^2, and mean square error, the linear association was the strongest when DSI values were calculated using only MRI-derived volumes. The linear model fitted the second best when all available variables were used together. The longitudinal DSI values derived on the basis of cognitive and neuropsychological tests had the smallest association values. Plots of residuals versus predicted values supported the interpretation that the DSI values calculated on the basis of ADAS and MMSE were the least linear over time: points in the plots were not as randomly distributed as they were when the DSI values were based on all available data, MRI, or NeuroBat (results not shown here).

The linear regression of the DSI values over time was performed for each subject independently. Medians of the regression parameters for SMCI and PMCI groups are shown in Table 5. The slopes and the intercepts of both groups were higher than zero ($p < 0.0005$). There were also clear differences between the two groups: PMCIs had five times higher slopes and almost three times higher intercepts than SMCIs ($p < 0.0005$).

The distributions of the slopes of both groups are presented in Fig. 2. On the basis of the visual inspection, the SMCI curve deviated from a Gaussian distribution containing also cases with higher slopes. Therefore, a hypothesis was put forth that the SMCI group actually contained two subgroups: one with truly stable DSI values and one with non-stable DSI values having signs of disease progression. A mixture distribution of two normal curves was fitted to the slopes of the SMCIs. The fits of unimodal and bimodal distributions were compared, and the results and estimated parameters are shown in the Supplementary Material. The results showed that the bimodal distribution fitted better to the slopes of the SMCIs than the unimodal distribution supporting the idea that two subgroups do exist within the SMCI group.

Visualizing progression of AD

In Fig. 1, the progression of AD is visualized using the DSF and the regression line of the DSI values. Most of the nodes in the DSF of a clear SMCI case are blue indicating that the patient data remained constantly unlike the data of those with AD. Also, the slope and the intercept of the regression line have low values (Fig. 1, topmost panel). On the contrary, almost all nodes of a clear PMCI case are red, indicating strong resemblance to previously diagnosed AD cases, and the slope and the intercept are higher as well (Fig. 1, lowest panel). A SMCI case with clearly increasing DSI values and the DSF changing from blue to red is also shown (Fig. 3, mid-panel). This case belongs to the subgroup of SMCI cases with non-stable DSI values in Fig. 2.

Differentiation using the trend parameters

MCI cases were classified as SMCI or PMCI using the regression parameters of the longitudinal DSI values, and the classification performance results are

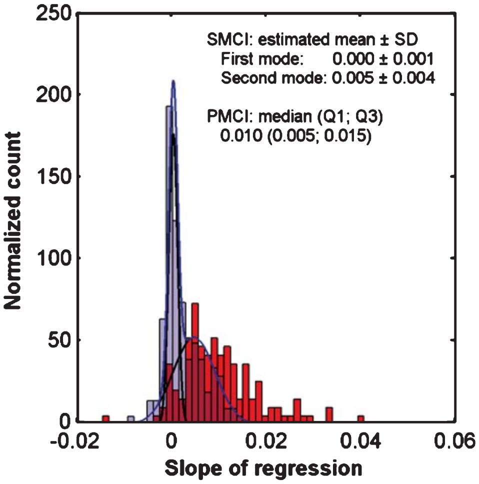


Fig. 2. Histograms of the slopes for stable (SMCI, blue) and progressive (PMCI, red) mild cognitive impairment cases. There appears to be two separate subgroups in the SMCI group. A mixture distribution of two normal curves fitted to the slopes of SMCIs is also shown. The areas of the histograms are scaled to one. (SD = standard deviation, Q1 = 25th quartile, Q3 = 75th quartile).

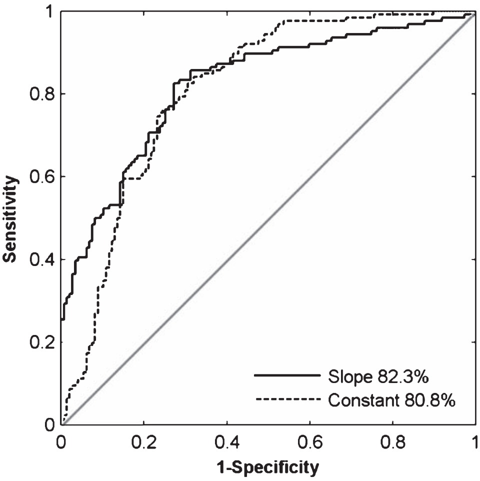


Fig. 3. Receiver operating characteristic curves of the slope (solid line) and the intercept (dashed line). Regression parameters were defined using total Disease State Index values over time.

presented in Table 6. AUCs were the highest when all available variables were used in the analysis (total). Classification accuracies were normally distributed, except for the slopes derived using NeuroBat. The classification accuracy of the slopes (total) was significantly higher than the classification accuracies of the slopes derived using ADAS or MRI ($p = 0.001$ for total-ADAS and $p = 0.005$ for total-MRI comparisons). The classification accuracy of the intercepts (total) was significantly higher than classification accuracy of the MRI-derived intercepts ($p = 0.004$). Other pairwise comparisons of the slopes and the intercepts were not statistically significant (all $p > 0.01$, Bonferroni-corrected significance level was 0.0056). The classification accuracies of the slopes (total) and the intercepts (total) were very similar (76.9% and 74.6%, respectively, $p = 0.309$). ROC curves of the slopes (total) and the intercepts (total) are presented in Fig. 3.

Table 6
Classification performance of the regression parameters of the longitudinal Disease State Index values derived using different datasets

	AUC (%)	Accuracy (%)	Sensitivity (%)	Specificity (%)
Slope				
Total	82.3	76.9 ± 8.8	82.2 ± 13.7	73.0 ± 15.0
MMSE	77.1	71.8 ± 7.6	55.5 ± 15.5	86.5 ± 5.5
ADAS	76.8	68.7 ± 10.2	51.1 ± 19.2	83.6 ± 10.2
NeuroBat	76.6	69.2 ± 5.8	60.2 ± 13.2	76.9 ± 15.3
MRI	71.0	66.8 ± 8.1	49.5 ± 14.4	80.6 ± 14.7
Intercept				
Total	80.8	74.6 ± 8.7	75.1 ± 17.4	74.4 ± 12.2
MMSE	79.0	72.0 ± 5.0	84.2 ± 11.6	61.5 ± 11.6
ADAS	80.3	74.9 ± 8.8	74.4 ± 15.6	75.7 ± 10.5
NeuroBat	79.3	66.9 ± 6.1	74.4 ± 21.7	61.0 ± 14.0
MRI	69.6	60.4 ± 8.9	55.6 ± 16.2	63.9 ± 16.2

Results are mean ± standard deviation from the stratified 10-fold cross-validation, except for the AUC. Total, all available variables included when calculating Disease State Index values; MMSE, Mini-Mental State Examination; ADAS, Alzheimer's Disease Assessment Scale-cognitive subscale; NeuroBat, Neuropsychological Battery; MRI, brain volumes derived from magnetic resonance imaging; AUC, area under the receiver operating characteristic curve.

DISCUSSION

Quantification of disease progression from MCI to AD was studied by applying the DSI method to heterogeneous longitudinal patient data and analyzing the behavior of the DSI values over time in subjects with MCI. Trend parameters of the longitudinal DSI values were obtained from regression and ability of them to differentiate between the groups of stable and progressive MCI was also studied.

In this study, it was assumed that the behavior of the longitudinal DSI values can be modeled linearly. The linear association was the strongest when the DSI values were based only on MRI features. Behavior of the total DSI values was not as linear because neuropsychological tests were included and their temporal behavior was the least linear. The linear model may not necessarily be the best model for progression of AD but it was selected because of simplicity and due to paucity of data. Some subjects with PMCI had only a few DSI values available for the regression due to synchronization of the time stamps.

Jack and his colleagues [13] proposed that changes in biomarkers over time would be sigmoidal and biomarkers would become abnormal in a certain temporal order. These assumptions gained support in several studies and they still are core components of the recently revised model [14]. Caroli et al. [33] provided the first evidence supporting the first version of the model. They compared the fit of linear and sigmoidal model and concluded that the sigmoidal model fitted better for hippocampal volume, and amyloid-β and total-tau in CSF. The linear model fitted better for FDG-PET data. Instead of real longitudinal data, Caroli et al. [33] used data from healthy controls, PMCIs, and early and late ADs at the baseline to reflect the progression of AD. Mouiha and Duchesne [34] used the same kind of cross-sectional setting to study the relationship between biomarkers and disease severity. They fitted six different models (linear, quadratic, robust quadratic, local quadratic regression, penalized B-spline, and sigmoid) to baseline data from healthy controls, PMCI, and AD cases [34]. According to them, amyloid-β had a piece-wise quadratic relationship, hippocampal volume and CSF measures of phosphorylated tau and total tau were best modeled with penalized B-splines, and linear model was the best fit for FDG-PET [34].

The results in this study show that the change of DSI values over time as reflected by the slope of the linear regression equation is clearly different in the SMCI and PMCI groups. The slope of PMCI cases was five times higher than the slope of SMCI cases. When the slopes of SMCI cases were studied more thoroughly, it was noticed that there were two different subgroups in the SMCI group: a group with lower slopes and another group with higher slopes that overlap with the slopes of the PMCI cases. It is expected that the peak with higher slopes represents MCIs that would convert to AD or other dementia later if the follow-up was continued. Davatzikos et al. [20] and Cui et al. [19] also found in their studies that subjects in the SMCI group did not have uniform results. Some SMCI cases had markers similar to AD, suggesting that they may convert to AD in the future [19, 20].

Samtani et al. [35] modeled a subject's rate of disease progression using a logistic model with several covariates. Severity of the disease was measured using ADAS and the analysis was restricted to an AD population [35]. Another approach for modeling disease progression was presented by Escudero et al. [36]. They found profiles of disease and normality using an unsupervised learning method (k-means clustering). Escudero et al. [36] calculated a so-called Bioindex that describes a subject's degree of membership to the profile of disease on the basis of measured data. To study evolution of Bioindeces over time, a sigmoid function was fitted to the Bioindex values at different time points. They used the same approach as here and fitted an individual function to the Bioindeces of each subject and studied evolution of Bioindeces in the groups of SMCI and PMCI. As in this study, they found that converters had steeper progression towards AD than non-converters. However, Escudero et al. [36] did not take into account that MCI patients arrived in the study at different phases of the disease, and they did not synchronize the time stamps as we did.

Patient visits in this study were synchronized according to the time of receiving AD diagnosis. Using this method, the accuracy of the synchronization depends on the accuracy of the actual AD diagnoses. Also, data points of the SMCI cases are not synchronized because they do not have an AD diagnosis. Jedynak et al. [37] and Yang et al. [38] proposed more sophisticated methods for synchronization. Jedynak et al. [37] used multiple biomarkers to create a disease progression score, which set the subjects on the same timeline [37]. Biomarkers were assumed to follow a sigmoidal function when constructing the disease progression score [37]. Yang et al. [38] modeled evolution of ADAS 13 score over time with an exponential model and then defined the start of the cognitive decline using the model. Other biomarkers were then synchronized using the estimated period of cognitive decline. After the synchronization, evolution of biomarkers over time and relations between them were clearer and they supported the model presented by Jack et al. [13, 14, 38]. In the approach presented in [38], one needs to define an accurate model for the progression of ADAS 13 score over time, and the accuracy of the synchronization depends on the suitability of the model.

The dynamic range for the DSI depends on training sets used. In this study, the DSI values were calculated on the basis of data from SMCI cases at baseline and PMCI cases at the point of conversion to AD. Thus, the dynamic range lies between MCI and early AD. Using the same model of disease progression to study healthy controls and late AD groups would saturate DSI values close to zero and one, respectively. On the other hand, if the training set consisted of PMCI and AD groups, the DSI would characterize changes at the later phase of the disease. Thus, if different training sets are used, the longitudinal behavior of the DSI values can be somewhat different. As another example, if training set included healthy and AD cases, slopes of the SMCI and PMCI groups should be closer to each other than they are in this study.

Training data for this study was selected from SMCI cases at the baseline and PMCI cases at the point of conversion because the initial purpose for the proposed method is in early diagnosis of AD. The main use case for the method is a situation where a subject with memory complaints arrives at a clinic. After some tests have been administered, computer-based decision support tools could help in objective assessment of patient data and possibly provide help for earlier diagnosis of AD. If the diagnosis cannot be made at the baseline, longitudinal quantification of progressing disease state provides additional information to base the diagnosis on. By selecting SMCI cases at the baseline and PMCI cases at the moment of receiving diagnosis as the training set, the system is optimized to detect early AD cases from an MCI population referred to a memory clinic. The DSI method is currently incorporated in a decision support tool that will be used in pilot studies and the training set used in the tool comprises SMCI and PMCI cases, similar to this study. When studies with other purposes (e.g., focus on conversion from normal cognition to MCI) are done in the future, then the practical issues of selecting the most appropriate training population will be addressed.

Recently, several studies have predicted the conversion from MCI to AD by combining multiple data modalities and identifying converters and non-converters on the basis of the data [19–23]. In these studies, multimodal data were combined using logistic regression [21, 22], the DSI method [22], support vector machine classifiers [19, 22, 23, 39], and a Naive Bayes classifier [22]. In [19, 20, 22, 40], it was found that combination of multimodal data resulted in better classification performance than the use of a single modality of data, e.g., using only neuropsychological tests. However, those studies did not report whether the differences were statistically significant. Ewers et al. [21] found that increasing number of variables in the model from one to four increased the classification accuracy, but the increase was not significant according to the 95% confidence intervals. Cui et al. [39] also combined different data modalities for predicting conversion from normal cognition to MCI. They reported that combination of neuropsychological test scores and MRI features resulted in significantly higher classification accuracy for the predictions than using either of the data modalities alone. Results from our study are in line with the previous research findings. Combination of all available data resulted in higher classification accuracies and AUCs than using only a single modality of data and increases in classification accuracies were not always statistically significant. To account for multiple comparisons, we used Bonferroni correction which is known to be a rather conservative method. However, in many comparisons, *p*-values were higher than 0.05.

It is worth noting that the calculation of the linear regression included DSI values from the point of conversion for PMCI cases. Thus, the classification performance measures presented here do not describe the ability of the trend parameters to predict conversion from MCI to AD. However, they demonstrate that the trend parameters of the DSI values are clearly different between the groups of SMCI and PMCI. Prediction of MCI to AD conversion with the DSI method using data from the ADNI database has already been studied in [22] and [25].

One interesting finding was that the MRI-derived longitudinal DSI values had the strongest linear association but the regression parameters of the MRI-based DSI values performed the worst in the classification. One explanation could be that changes related to normal aging in the brain may interfere with the results. For example, Koikkalainen et al. [41] removed effects of age and other confounding factors by dividing patients into subgroups and using linear regression. These procedures improved classification accuracies in their study. Another explanation could be that MRI may be a better indicator of the rate of disease progression than of the disease stage. Stronger linearity of the MRI-derived DSI over time might also be caused by the fact that MRI measures are not as prone to daily variations as neuropsychological tests may be.

Missing values were imputed with the values from the previous available visit. This approach resulted in slightly outdated data for some patient visits and biased the results towards non-progression. This approach was chosen so that all data used in the analyses really were available from a patient at the specific moments. This would not be the case, e.g., if missing values were replaced with the next available values or using other more complex imputation methods. Replacing missing values with next available values would have biased results toward progression to some extent and there would still have been missing values because some patients did not have any values available beyond the last time point. If the missing values had not been imputed at all, the DSI values at different time points would have been calculated using different variables for each visit and this would have hindered the interpretation of the longitudinal results.

The study had some limitations. The final diagnoses for the subjects were determined on the basis of clinical evaluation and they were not verified with postmortem histological samples taken from the brain. Also, the study period of 48 months is relatively short. Thus, some subjects diagnosed currently as stable MCI may convert to AD later. This study utilized longitudinal data from a period of 2–4 years. In clinics, where the patients are diagnosed, there may not be data from such a long period available. Less longitudinal data will probably produce more variation in the slopes and the intercepts of the regression equation. On the other hand, this study suggests that quantifying longitudinal patient data using the DSI method provides valid information for decision support and is a valid methodology to follow-up a patient's condition in a quantitative manner.

In conclusion, this study demonstrates that combining sparse and heterogeneous data with the DSI method can be used for deriving a quantitative measure related to early AD progression. Significant trends were found in longitudinal DSI values: rate of change of DSI values was five times higher in the PMCI group than in the SMCI group. Classification of the subjects as converters and non-converters on the basis of the regression parameters (the slope and the intercept) also showed that SMCI and PMCI cases can be differentiated on the basis of the trend parameters.

ACKNOWLEDGMENTS

This work was partially funded under the 7th Framework Programme by the European Commission (http.// cordis.europa.eu/ist; EU-Grant-224328-PredictAD; Name: From Patient Data to Personalized Healthcare in Alzheimer's Disease) and the SalWe Research Program for Mind and Body (Tekes – The Finnish Funding Agency for Technology and Innovation, grant 1104/10).

Authors' disclosures available online (http://www.j-alz.com/disclosures/view.php?id=1923).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129 and K01 AG030514.

*Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

SUPPLEMENTARY MATERIAL

The supplementary material and tables are available in the electronic version of this article: http://dx.doi.org/10.3233/JAD-130359.

REFERENCES

[1] Nestor PJ, Scheltens P, Hodges JR (2004) Advances in the early detection of Alzheimer's disease. *Nat Med* **10** Suppl, S34-S41.

[2] Gauthier S, Reisberg B, Zaudig M, Petersen RC, Ritchie K, Broich K, Belleville S, Brodaty H, Bennett D, Chertkow H, Cummings JL, de Leon M, Feldman H, Ganguli M, Hampel H, Scheltens P, Tierney MC, Whitehouse P, Winblad B, International Psychogeriatric Association Expert Conference

on mild cognitive impairment (2006) Mild cognitive impairment. *Lancet* **367**, 1262-1270.

[3] Petersen RC (2009) Early diagnosis of Alzheimer's disease: is MCI too late? *Curr Alzheimer Res* **6**, 324-330.

[4] Brooks LG, Loewenstein DA (2010) Assessing the progression of mild cognitive impairment to Alzheimer's disease: current trends and future directions. *Alzheimers Res Ther* **2**, 28.

[5] Sloane PD, Zimmerman S, Suchindran C, Reed P, Wang L, Boustani M, Sudha S (2002) The public health impact of Alzheimer's disease, 2000-2050: potential implication of treatment advances. *Annu Rev Public Health* **23**, 213-231.

[6] Salloway S, Mintzer J, Weiner MF, Cummings JL (2008) Disease-modifying therapies in Alzheimer's disease. *Alzheimers Dement* **4**, 65-79.

[7] Galimberti D, Scarpini E (2011) Disease-modifying treatments for Alzheimer's disease. *Ther Adv Neurol Disord* **4**, 203-216.

[8] Duara R, Barker W, Loewenstein D, Bain L (2009) The basis for disease-modifying treatments for Alzheimer's disease: the Sixth Annual Mild Cognitive Impairment Symposium. *Alzheimers Dement* **5**, 66-74.

[9] Ballard C, Gauthier S, Corbett A, Brayne C, Aarsland D, Jones E (2011) Alzheimer's disease. *Lancet* **377**, 1019-1031.

[10] Hampel H, Bürger K, Teipel SJ, Bokde AL, Zetterberg H, Blennow K (2008) Core candidate neurochemical and imaging biomarkers of Alzheimer's disease. *Alzheimers Dement* **4**, 38-48.

[11] Borroni B, Premi E, Di Luca M, Padovani A (2007) Combined biomarkers for early Alzheimer disease diagnosis. *Curr Med Chem* **14**, 1171-1178.

[12] Craig-Schapiro R, Fagan AM, Holtzman DM (2009) Biomarkers of Alzheimer's disease. *Neurobiol Dis* **35**, 128-140.

[13] Jack CR Jr, Knopman DS, Jagust WJ, Shaw LM, Aisen PS, Weiner MW, Petersen RC, Trojanowski JQ (2010) Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol* **9**, 119-128.

[14] Jack CR Jr, Knopman DS, Jagust WJ, Petersen RC, Weiner MW, Aisen PS, Shaw LM, Vemuri P, Wiste HJ, Weigand SD, Lesnick TG, Pankratz VS, Donohue MC, Trojanowski JQ (2013) Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol* **12**, 207-216.

[15] Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, Fagan AM, Iwatsubo T, Jack CR Jr, Kaye J, Montine TJ, Park DC, Reiman EM, Rowe CC, Siemers E, Stern Y, Yaffe K, Carrillo MC, Thies B, Morrison-Bogorad M, Wagster MV, Phelps CH (2011) Toward defining the preclinical stages of Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 280-292.

[16] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 270-279.

[17] McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, Klunk WE, Koroshetz WJ, Manly JJ, Mayeux R, Mohs RC, Morris JC, Rossor MN, Scheltens P, Carrillo MC, Thies B, Weintraub S, Phelps CH (2011) The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 263-269.

[18] Jack CR Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH (2011) Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 257-262.

[19] Cui Y, Liu B, Luo S, Zhen X, Fan M, Liu T, Zhu W, Park M, Jiang T, Jin JS, the Alzheimer's Disease Neuroimaging Initiative (2011) Identification of conversion from mild cognitive impairment to Alzheimer's disease using multivariate predictors. *PLoS One* **6**, e21896.

[20] Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ (2011) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* **32**, 2322.e19-2322.e27.

[21] Ewers M, Walsh C, Trojanowski JQ, Shaw LM, Petersen RC, Jack CR Jr, Feldman HH, Bokde AL, Alexander GE, Scheltens P, Vellas B, Dubois B, Weiner M, Hampel H, North American Alzheimer's Disease Neuroimaging Initiative (ADNI) (2012) Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. *Neurobiol Aging* **33**, 1203-1214.

[22] Mattila J, Koikkalainen J, Virkki A, Simonsen A, van Gils M, Waldemar G, Soininen H, Lötjönen J, the Alzheimer's Disease Neuroimaging Initiative (2011) A disease state fingerprint for evaluation of Alzheimer's disease. *J Alzheimers Dis* **27**, 163-176.

[23] Zhang D, Shen D, the Alzheimer's Disease Neuroimaging Initiative (2012) Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers. *PLoS One* **7**, e33182.

[24] Gray KR, Aljabar P, Heckemann RA, Hammers A, Rueckert D, the Alzheimer's Disease Neuroimaging Initiative (2012) Random forest-based similarity measures for multi-modal classification of Alzheimer's disease. *Neuroimage* **65**, 167-175.

[25] Mattila J, Soininen H, Koikkalainen J, Rueckert D, Wolz R, Waldemar G, Lötjönen J, the Alzheimer's Disease Neuroimaging Initiative (2012) Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects. *J Alzheimers Dis* **32**, 969-979.

[26] Alzheimer's Disease Neuroimaging Initiative, http://adni.loni.ucla.edu, Accessed on May 4, 2013.

[27] Weiner MW, Aisen PS, Jack CR Jr, Jagust WJ, Trojanowski JQ, Shaw L, Saykin AJ, Morris JC, Cairns N, Beckett LA, Toga A, Green R, Walter S, Soares H, Snyder P, Siemers E, Potter W, Cole PE, Schmidt M, the Alzheimer's Disease Neuroimaging Initiative (2010) The Alzheimer's Disease Neuroimaging Initiative: progress report and future plans. *Alzheimers Dement* **6**, 202-211.

[28] ADNI1 Procedures Manual (2011), http://www.adni-info.org/Scientists/Pdfs/ADNI1_Procedures_Manual_Revised_12052011.pdf, Posted on May 12, 2011, Accessed on January 23, 2012.

[29] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM (2002) whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341-355.

[30] Dubois B, Feldman HH, Jacova C, DeKosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P (2007) Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *Lancet Neurol* **6**, 734-746.

[31] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-944.

[32] Llano DA, Laforet G, Devanarayan V, the Alzheimer's Disease Neuroimaging Initiative (2011) 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* **25**, 73-84.

[33] Caroli A, Frisoni GB, the Alzheimer's Disease Neuroimaging Initiative (2010) The dynamics of Alzheimer's disease biomarkers in the Alzheimer's Disease Neuroimaging Initiative cohort. *Neurobiol Aging* **31**, 1263-1274.

[34] Mouiha A, Duchesne S, the Alzheimer's Disease Neuroimaging Initiative (2012) Toward a dynamic biomarker model in Alzheimer's disease. *J Alzheimers Dis* **30**, 91-100.

[35] Samtani MN, Farnum M, Lobanov V, Yang E, Raghavan N, DiBernardo A, Narayan V, the Alzheimer's Disease Neuroimaging Initiative (2012) An improved model for disease progression in patients from the Alzheimer's Disease Neuroimaging Initiative. *J Clin Pharmacol* **52**, 629-644.

[36] Escudero J, Ifeachor E, Zajicek JP, the Alzheimer's Disease Neuroimaging Initiative (2012) Bioprofile analysis: a new approach for the analysis of biomedical data in Alzheimer's disease. *J Alzheimers Dis* **32**, 997-1010.

[37] Jedynak BM, Lang A, Liu B, Katz E, Zhang Y, Wyman BT, Raunig D, Jedynak CP, Caffo B, Prince JL, the Alzheimer's Disease Neuroimaging Initiative (2012) A computational neurodegenerative disease progression score: method and results with the Alzheimer's Disease Neuroimaging Initiative cohort. *Neuroimage* **63**, 1478-1486.

[38] Yang E, Farnum M, Lobanov V, Schultz T, Verbeeck R, Raghavan N, Samtani MN, Novak G, Narayan V, DiBernardo A, the Alzheimer's Disease Neuroimaging Initiative (2011) Quantifying the pathophysiological timeline of Alzheimer's disease. *J Alzheimers Dis* **26**, 745-753.

[39] Cui Y, Sachdev PS, Lipnicki DM, Jin JS, Luo S, Zhu W, Kochan NA, Reppermund S, Liu T, Trollor JN, Brodaty H, Wen W (2012) Predicting the development of mild cognitive impairment: a new use of pattern recognition. *Neuroimage* **60**, 894-901.

[40] Zhang D, Wang Y, Zhou L, Yuan H, Shen D, the Alzheimer's Disease Neuroimaging Initiative (2011) Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* **55**, 856-867.

[41] Koikkalainen J, Pölönen H, Mattila J, van Gils M, Soininen H, Lötjönen J, the Alzheimer's Disease Neuroimaging Initiative (2012) Improved classification of Alzheimer's disease data via removal of nuisance variability. *PLoS One* **7**, e31112.

PUBLICATION IV

Optimizing the diagnosis of early Alzheimer's disease in mild cognitive impairment subjects

In: Journal of Alzheimer's Disease 32(4),
pp. 969–979.

Journal of Alzheimer's Disease 32 (2012) 969–979
DOI 10.3233/JAD-2012-120934
IOS Press

Optimizing the Diagnosis of Early Alzheimer's Disease in Mild Cognitive Impairment Subjects

Jussi Mattila[a,*], Hilkka Soininen[b], Juha Koikkalainen[a], Daniel Rueckert[c], Robin Wolz[c], Gunhild Waldemar[d], Jyrki Lötjönen[a] and for the Alzheimer's Disease Neuroimaging Initiative[1]

[a] *VTT Technical Research Centre of Finland, Tampere, Finland*

[b] *Department of Neurology, University of Eastern Finland, Kuopio University Hospital, Kuopio, Finland*

[c] *Department of Computing, Imperial College London, London, UK*

[d] *Department of Neurology, Memory Disorders Research Group, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark*

Handling Associate Editor: J. Wesson Ashford

Accepted 2 July 2012

Abstract. In the diagnostic process of Alzheimer's disease (AD), there may be considerable delays between first contact to outpatient services and a final, definitive diagnosis. In Europe the average delay is 20 months. Nevertheless, patient data preceding clinical AD diagnoses often contains early signs of the disease. Several studies have analyzed data of mild cognitive impairment (MCI) subjects, showing that conversion from MCI to AD can be predicted with a classification accuracy of 60–80%. This accuracy may not be high enough for influencing diagnostic decisions. In this work, the prediction problem is approached differently; a target prediction accuracy is defined first and is then used for identifying MCI patients for whom the required accuracy can be reached. The process uses a novel disease state index method in which patient data are statistically compared to a high number of previously diagnosed cases. It is shown that the disease index values derived from heterogeneous patient data can be used for identifying groups of patients for whom the prediction accuracy reaches the previously set target level. The results also show that 12 months before receiving clinical AD diagnoses, approximately half (51.5%, 95% confidence interval: 48.6–54.2%) of MCI subjects who progressed to AD can be classified with a high accuracy of 87.7%, possibly enough to support earlier diagnostic decisions.

Keywords: Clinical decision support, early Alzheimer's disease, mild cognitive impairment, patient selection

INTRODUCTION

Evidence shows that Alzheimer's disease (AD) pathology begins several years or even decades prior to onset of dementia, but the symptoms that eventually draw medical attention appear only after the disease has reached a certain stage [1]. After the initial visit to a memory clinic, many patients are considered to be in a transitional state, referred to as mild cognitive impairment (MCI), where they have memory problems that are abnormal for their age, but their functional capacity

[1] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[*] Correspondence to: Jussi Mattila, VTT Technical Research Centre of Finland, Sinitaival 6, 33101 Tampere, Finland. Tel.: +358 40 592 7979; Fax: +358 20 722 3499; E-mail: jussi.mattila@vtt.fi.

in activities of daily living is intact [2]. It is known that MCI is a heterogeneous entity with an increased probability of developing to AD, but MCI may also remain stable, progress to other dementias, or even return to normal cognition [3, 4]. For subjects with MCI who later progress to AD, it takes on average one to three years to get an AD diagnosis in the clinical phase, due to uncertainties of the diagnosis or slow disease progression [5–9]. Though there is no cure for AD yet, diagnosis in the early prodromal phase would allow the prescription of therapies and medications, which are believed to work better the earlier they are started [10, 11].

Currently, there exists no specific test that would confirm AD *in vivo*. Therefore, clinicians are required to make an informed judgment based on available information. When compared to neuropathological gold standards — obtained with postmortem brain autopsy — clinical diagnosis of AD reaches accuracies ranging from 70–90% [12–14]. Recent research has shown that biomarkers could improve diagnostic accuracy even at early phases of the disease [15–17]. These findings are contributing to the new diagnostic guidelines of predementia and prodromal AD, which are emphasizing the combination of biomarkers, neuropsychological tests, and clinical assessment [18–20]. The generally accepted AD biomarkers are:

- Low amyloid-β levels and/or elevated (phospho) tau levels in cerebrospinal fluid (CSF);
- atrophy of the temporal lobe in magnetic resonance imaging (MRI);
- temporo-parietal hypometabolism as assessed with 18-labeled fluoro-2-deoxy-D-glucose positron emission tomography (PET) or identification of amyloid accumulation in the brain with Pittsburgh Compound B (PIB) PET, and
- known causative genetic mutations in immediate family.

There has been a vast amount of studies aiming at predicting which MCI subjects will convert to AD and which will not, based on biomarker data measured at early phases of the disease [21–29]. Most of these studies rely on supervised classifiers, which assign the most probable class label for a given patient using a decision model derived from training data. They have shown that predicting conversion from MCI to AD from early measurement data is possible at an accuracy of 60–80%, with larger cohorts generally around 70%. With the overall accuracy of classification so low, predictions of outcomes can influence diagnostic decisions only marginally. Another issue with many classifiers is that they only provide a label (AD/no AD) or a disease probability without any estimate of the reliability of the result for each case individually. In this work, the challenge of prediction was approached from another angle, which may better address the clinical need. A target prediction accuracy is defined first and is then used for identifying groups of patients for which this accuracy can be reached. The approach is made possible by using a novel disease state index (DSI) method, which has been designed for clinical decision support [30]. The DSI takes as input quantifiable heterogeneous patient data and provides as output values from zero to one, indicating the proportion of patient's data that match a disease profile, e.g., that of early AD. The DSI is also transparent, i.e., the rationale for all predictions is provided to those who wish to use the results.

In this study, subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI) [31] were analyzed with the DSI method to find the proportion of MCI subjects whose measurement data contain strong evidence of early AD. First, an appropriate target level for the prediction accuracy was defined. Second, the DSI method was applied to MCI data at 6, 12, 18, and 24 months before receiving clinical AD diagnoses. Third, threshold DSI values that indicate strong evidence of early AD were determined. Finally, the numbers of MCI subjects identified as having strong evidence of early AD, based on their DSI values, were analyzed. The main contributions of this study are 1) a data-driven target classification accuracy for predicting conversion from MCI to AD; 2) DSI thresholds indicating strong evidence of early AD at 6, 12, 18, and 24 months before receiving AD diagnoses; and 3) proportion of MCI subjects having strong evidence of early AD in their data, i.e., the number of MCI patients who could possibly be diagnosed earlier than is current clinical practice. The eventual use case for the proposed method is in clinical decision support. If data from an individual with MCI exceeds a previously set DSI threshold, one could predict conversion to AD within a given period of time (e.g., 24 months) with an accuracy that is clinically relevant (e.g., close to 90%) and also see the justification for the prediction due to transparency of the DSI method.

MATERIALS AND METHODS

Disease State Index (DSI)

To make better use of available patient data, including a variety of biomarkers and neuropsychological tests, a novel clinical decision support method has been developed [30]. It computes DSI values between zero and one from patient measurements by comparing them comprehensively to a large number of previously diagnosed cases. The DSI supports any quantifiable, heterogeneous, and sparse data. It analyses the available measurements and their combinations without requiring manual cleaning of data, feature selection, or other pre-processing steps. Visualizations of a patient's disease state, called disease state fingerprints (DSF), inform the clinician about important diagnostic measures and about their relationship to patient measurements (see Fig. 1). DSI and DSF are not diagnostic tools *per se*. Their main purpose is to help clinicians quickly interpret and analyze large quantities of heterogeneous patient data. In addition, DSI has been shown to predict AD and other diseases with accuracy comparable to established machine-learning methods, and, together with DSF, is expected to have potential in improving both the confidence and accuracy of clinical diagnosis [30, 32]. In this study, the DSI method was used for measuring evidence of early AD in patient data, with increasing DSI values within the test cohorts modeling the progression of the disease.

Alzheimer's Disease Neuroimaging Initiative (ADNI)

Data used in the preparation of this article were obtained from the ADNI database (http://adni.loni.ucla.edu, accessed 2 September 2011). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies, and non-profit organizations, as a $60 million, 5-year public-private partnership. The primary goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD. Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials.

The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California-San Francisco. ADNI is the result of efforts of many co-investigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the US and Canada. The initial goal of ADNI was to recruit 800 adults, ages 55 to 90, to participate in the research, approximately 200 cognitively normal older individuals to be followed for 3 years, 400 people with MCI to be followed for at least 3 years, and 200 people with early AD to be followed for 2 years. For up-to-date information, see http://www.adni-info.org/.

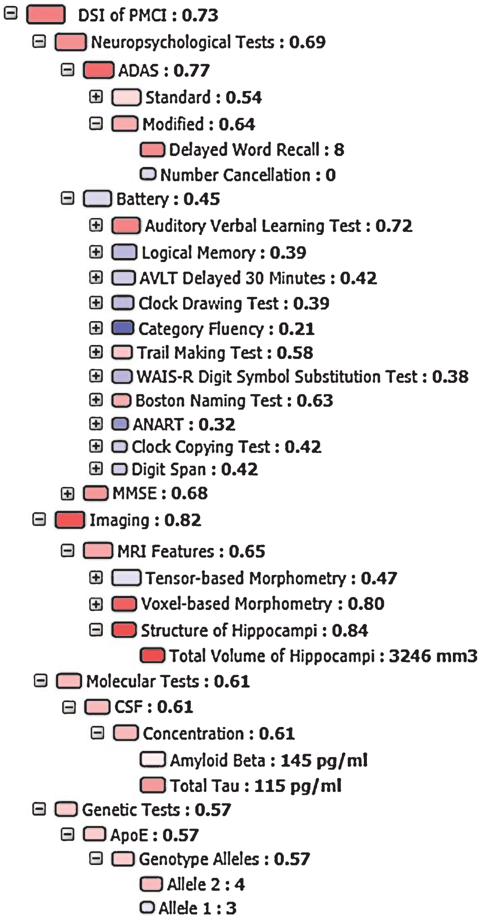


Fig. 1. Disease state fingerprint (DSF) visualization of patient data with a large share of measurement values indicating early AD. Relevance of a particular test is indicated by the size of the box next to the test's name. Red color and DSI values approaching one (1.00) indicate similarity to early AD cases. Blue color and DSI values close to zero (0.00) indicate similarity with stable MCIs. Leaves of the tree show the raw test and measurement values. Not all nodes are fully expanded in the tree; collapsed nodes show the overall DSI value from that test section.

Table 1
Demographic information of the study cohort at baseline visit

	Stable MCI (SMCI)	Progressive MCI (PMCI)
Subjects	148	140
Males	97 (65.5%)	85 (60.7%)
Age	75.1 (7.4)	75.4 (6.7)
Average time for AD diagnosis after baseline visit	n/a	20.8 months
Years of education	15.9 (3.0)	15.6 (3.0)
MMSE score[a]	27.5 (1.7)	26.6 (1.6)
ADAS-cog 13 score[a]	16.3 (6.0)	21.1 (5.3)
APOE status (ε4 carriers)[b]	62 (41.9%)	95 (67.9%)

Data presented as number of subjects (percentage of subjects %) or as mean (standard deviation).

[a] Statistically significant difference between SMCIs and PMCIs (student's t-test, $p < 0.05$).

[b] Statistically significant difference between SMCIs and PMCIs (Fisher's exact test, $p < 0.05$), APOE carriers are subjects with one or two copies of ε4 allele at the APOE locus.

From ADNI, MCI patients with at least 24 months of data were included in this study. Patients were stratified into two groups according to conversion to AD. The first group was a stable MCI group (SMCI, $n = 148$), where diagnosis remained the same (MCI) during all follow-ups. The other group was a progressive MCI group (PMCI, $n = 140$), where diagnosis changed to AD during the study and remained AD thereafter. Patients who converted otherwise (e.g., MCI $\longrightarrow$ healthy or MCI $\longrightarrow$ AD $\longrightarrow$ MCI) were excluded from the analysis. Since there are no autopsy data available, diagnoses made by ADNI (using NINCDS-ARDRA criteria [33]) were used as the reference for building the disease models and evaluating their performance. Demographic and clinical baseline data for these two groups are listed in Table 1.

Test and measurement data used in this study consisted of Mini-Mental State Examination (MMSE, 30 variables), Alzheimer's Disease Assessment Scale-cognitive subscale scores (ADAS-cog, 13 variables), neuropsychological battery (NeuroBat, 53 variables), amyloid-β and total tau levels (CSF, 2 variables), and genetic risk factors (Apolipoprotein E, APOE, 2 variables). All these data are readily available for researchers in the ADNI database. In addition, features derived from MRI images with fully automated methods of voxel-based morphometry [34], tensor-based morphometry [35], and hippocampal volume segmentation [36], were included into the study. These MRI image processing methods use an 83 region atlas to quantify the results over several structures of the brain and thus provide 83 variables each, except for the total volume of hippocampi, which is a single variable. Table 2 shows the number of visits and data modalities available for SMCI and PMCI subjects at each time point. Missing test values at months 6–48 were replaced with values from a patient's previous visit if available. This is analogous to having a comprehensive test battery available for analyses during all visits, albeit with slightly dated patient data in some cases.

Target prediction accuracy

Target prediction accuracy, or target accuracy, is defined here as a classification accuracy that must be reached when predicting conversion from MCI to AD for the prediction to be clinically relevant. As described in the Introduction, agreement between clinical and gold standard neuropathological diagnoses is between 70% and 90%. Because the ADNI database contains only clinical diagnoses, setting the target accuracy at over 90% would be questionable. An alternative data-driven approach to setting the target accuracy is to consider the discriminative power in the data at the time when clinical diagnoses of AD are made. One can assume that when clinicians make diagnoses, they have, implicitly or explicitly, concluded that measurement data contain enough evidence for making the decision. The rest of the decision is based on information that is not in the data, e.g., from meeting the patient and caregiver, and/or accepted uncertainty. Thus, appropriate target accuracy for AD diagnostics can be derived by comparing SMCI data to PMCI data from the moment when clinicians have made the AD diagnoses. In this way, target accuracy models the acceptable uncertainty of clinical AD diagnoses when only quantitative measurement data are considered.

The target accuracy was determined by comparing all SMCI data to measures from PMCI subjects at the time point they received AD diagnoses (step 1 in Fig. 2). Using the DSI method, SMCI visits were exhaustively tested to find the visit that best differentiates SMCIs from PMCIs at AD conversion time (step 2). The maximum classification accuracy obtained from these comparisons represents the minimum discriminatory power needed in data for AD diagnostics (step 3). This accuracy was set as the target accuracy (step 4) which, if exceeded using earlier PMCI data, could allow patients to be considered eligible for AD diagnosis earlier. The target accuracy was hoped to exceed 85% to be comparable with the level of clinical accuracy achieved when using NINCDS-ADRDA criteria, the diagnostic reference in this study.

Table 2
Data available in ADNI for the study cohort (MCI patients with at least 24 months of visit data)

Visit month	Baseline (Month 0)	Month 6	Month 12	Month 18	Month 24	Month 36	Month 48
SMCI patient visits	148	147	146	142	141	N/A[c]	N/A[c]
Verified MCI at +24 mo.[a]	148	120	120	19	19	–	–
MMSE	148	147	145	142	141	–	–
ADAS	148	147	145	142	141	–	–
NeuroBat	148	147	145	141	141	–	–
MRI	143	133	133	119	105	–	–
CSF	82	–	65	–	–	–	–
APOE	148	–	–	–	–	–	–
PMCI patient visits	140	139	140	136	139	112	32
Converted from MCI to AD[b]	0	13	49	76	111	137	140
MMSE	140	139	140	135	139	112	32
ADAS	140	139	140	135	139	111	32
NeuroBat	140	139	140	135	139	111	32
MRI	139	135	127	118	108	70	11
CSF	74	–	68	–	–	–	–
APOE	140	–	–	–	–	–	–

[a] Number of SMCI patients that are known to have remained stable 24 months after this visit.
[b] Number of PMCI patients whose diagnosis converted to AD during or before this visit.
[c] Visit data were excluded since the subjects could not be confirmed to have remained stable MCI for the subsequent 24 months.

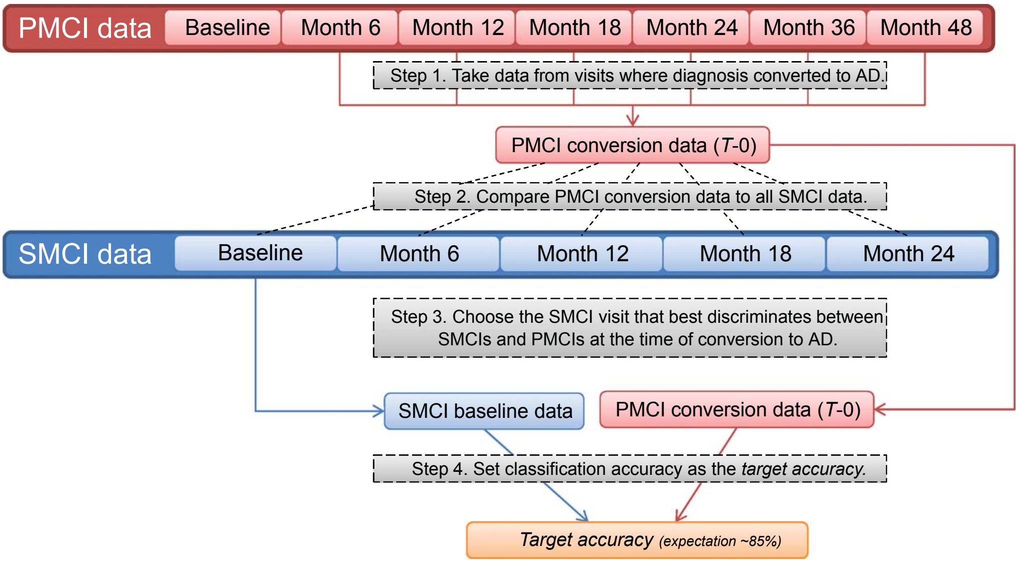

PMCI data
Baseline
Month 6
Month 12
Month 18
Month 24
Month 36
Month 48
Step 1. Take data from visits where diagnosis converted to AD.
PMCI conversion data (T-0)
Step 2. Compare PMCI conversion data to all SMCI data.
SMCI data
Baseline
Month 6
Month 12
Month 18
Month 24
Step 3. Choose the SMCI visit that best discriminates between SMCIs and PMCIs at the time of conversion to AD.
SMCI baseline data
PMCI conversion data (T-0)
Step 4. Set classification accuracy as the target accuracy.
Target accuracy (expectation ~85%)

Fig. 2. Steps 1–4 for determining the target accuracy in a data-driven manner.

MCI subjects with strong evidence of early AD

The primary goal of this study was to determine the proportion of MCI patients that could be classified with the target accuracy at an early phase of AD. To identify these patients, all data from PMCIs were aligned according to the visit where they received their AD diagnosis (step 5 in Fig. 3). This 'AD conversion visit' was set as *T*-0 (conversion time T minus zero months), with visits preceding the conversion aligned at *T*-6, *T*-12, *T*-18, and *T*-24 months. As some PMCIs converted during the first months of the ADNI study, there was no data for them at visits aligned to *T*-12 (number of PMCI visits available = 127), *T*-18 (*n* = 91), or *T*-24 (*n* = 64).

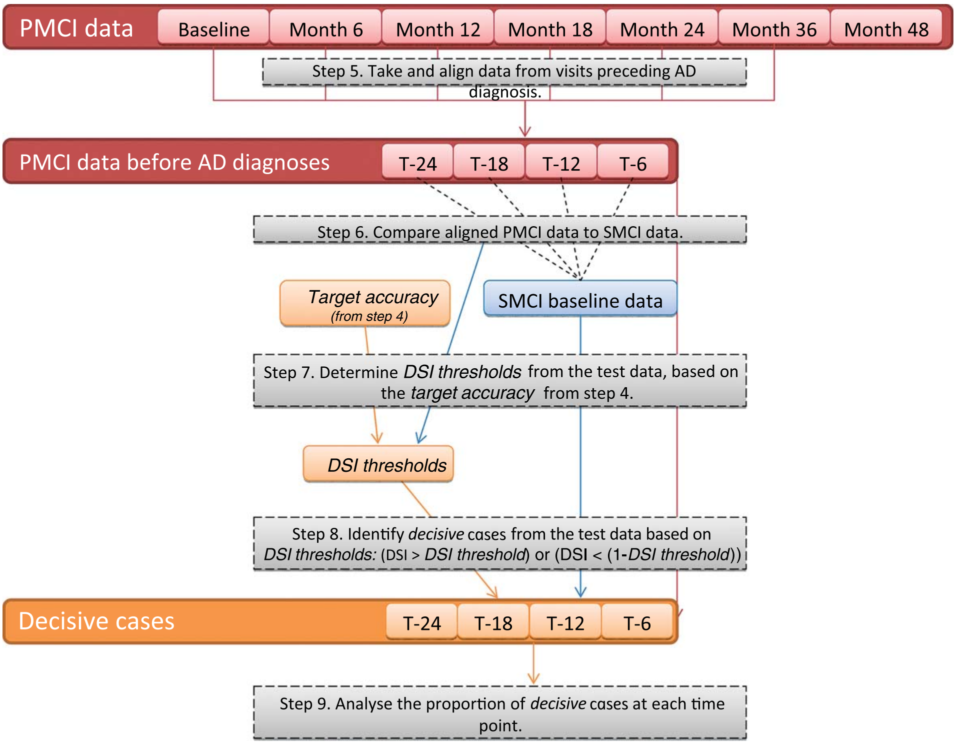

PMCI data
Baseline
Month 6
Month 12
Month 18
Month 24
Month 36
Month 48
Step 5. Take and align data from visits preceding AD diagnosis.
PMCI data before AD diagnoses
T-24
T-18
T-12
T-6
Step 6. Compare aligned PMCI data to SMCI data.
Target accuracy (from step 4)
SMCI baseline data
Step 7. Determine DSI thresholds from the test data, based on the target accuracy from step 4.
DSI thresholds
Step 8. Identify decisive cases from the test data based on DSI thresholds: (DSI > DSI threshold) or (DSI < (1-DSI threshold))
Decisive cases
T-24
T-18
T-12
T-6
Step 9. Analyse the proportion of decisive cases at each time point.

Fig. 3. Steps 5–9 for evaluating the proportion of MCI subjects whose data has strong evidence regarding early AD.

Aligned PMCI visits were analyzed against the most representative SMCI data, chosen in step 3, using the DSI method (step 6). For each aligned visit (*T*-24 to *T*-6), a DSI threshold for patient inclusion was determined (step 7). DSI thresholds define the range of DSI values, i.e., the level of evidence needed in patient data, which allows classification accuracy of included patients to reach the target accuracy. Included patients formed groups of decisive cases, where DSI values are larger than the DSI threshold, suggesting early AD, or smaller than 1 – DSI threshold, dismissing early AD (step 8). Thus, at each time point, only those patients with considerable evidence in their data, indicated by large or small DSI values, were included into the decisive groups. The rest of the patients were dismissed and not analyzed further. To reiterate, patient selection into the decisive groups using the DSI thresholds allowed classification accuracy of selected patients to reach the previously set target accuracy. Numbers of patients included in the decisive groups were evaluated, providing the proportion of MCI patients with strong evidence regarding early AD at 6, 12, 18, and 24 months before receiving their diagnoses (step 9).

To compute the DSI values used for patient selection, the DSI method must determine how much predictive power individual tests and their combinations have, i.e., how relevant they are (see [30] for details). As a final step in this study, proportions of relevance assigned to each test were evaluated to assess their contributions to the DSI method and thus to the patient selection process.

Test methodology

All data processing was performed in Matlab[2] using ten iterations of stratified (with consistent ratio of SMCIs and PMCIs) 10-fold cross-validation. The

[2] Matlab R2011b, The MathWorks Inc., Natick, MA, 2011.

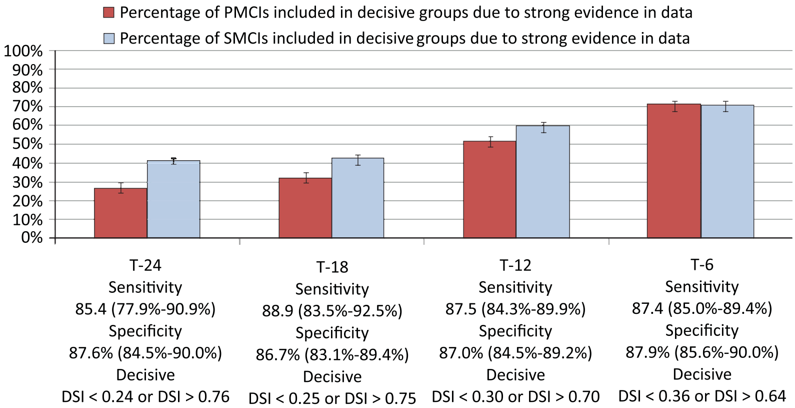


Fig. 4. Proportion of MCI subjects with strong evidence of early AD, or evidence dismissing AD, in data at visits preceding AD diagnoses. For these subjects, i.e., those assigned to the decisive groups, classification accuracy between SMCIs and PMCIs reaches the target accuracy (87.7%) at each time point. Also listed are the sensitivities and specificities for the subjects in the decisive groups, and the DSI thresholds that allowed selecting patients into the decisive groups while reaching the target accuracy.

resulting one hundred iterations of training a disease model and testing with unseen patient data give robust performance metrics and provide information about confidence intervals (CI). Statistical significances between groups were evaluated using student's t-test at significance level $p < 0.05$. All results are reported as the mean and 95% CI of the one hundred test iterations, unless otherwise noted.

RESULTS

Target prediction accuracy

PMCI conversion time data (*T*-0) was compared to all SMCI visit data using the DSI method. The best result was obtained with baseline SMCI data; discrimination between baseline SMCIs and conversion time PMCIs reached accuracy of 87.7% (CI: 86.6–88.8%), sensitivity of 86.9% (CI: 85.1–8.4%), and specificity of 88.5% (CI: 86.9–89.9%). Thus, prediction accuracy of 87.7% was set as the target accuracy, corresponding well with the maximum accuracy of clinical AD diagnoses using NINCDS-ADRDA criteria judged against neuropathological confirmations.

Previous research on ADNI data has shown that predicting MCI to AD conversion from baseline measurements attains accuracies between 60–72% [21–24]. Here, the baseline prediction accuracy between SMCI and PMCI data was in line with these studies, at 70.1% (CI: 68.4–71.5%). It should be noted that the target accuracy was set significantly higher than what can be achieved when considering all patients at baseline (87.7% versus 70.1%, respectively).

MCI subjects with strong evidence of early AD

PMCI data at visits preceding AD diagnoses (visits *T*-24 to *T*-6) were analyzed against baseline SMCI data to estimate whether data from earlier visits can reach the target accuracy. The DSI thresholds at each time point were selected such that the target accuracy of 87.7% was surpassed. Patients with low or high DSI values, limited by DSI thresholds, were selected into the decisive groups. Figure 4 shows the proportion of patients included into the decisive groups and the DSI thresholds allowing selection into them at different time points. The figure also lists the prediction sensitivities and specificities computed using only patients in the decisive groups. The proportion of PMCI patients in the decisive groups grows consistently with each visit closer to AD conversion time *T*-0. 24 months prior to AD diagnoses, measurement data allowed 26.2% (CI: 23.1–29.5%) of PMCI patients to be included into the decisive groups. At visit *T*-12, the decisive groups consisted of 51.5% (CI: 48.6–54.2%) of PMCI patients. The share of patients is maximized closest to conversion time at *T*-6, where it was it was possible to include 70.7% (CI: 68.5–73.1%) of PMCIs at the target accuracy. The ranges of DSI values allowing inclusion,

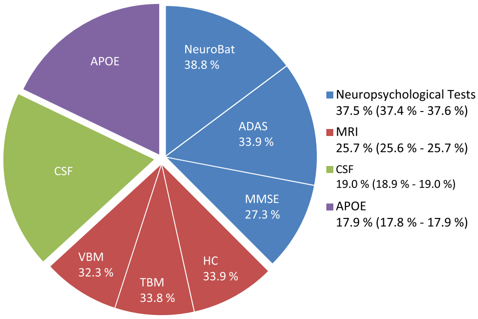


Fig. 5. Influence of the different measurement modalities and individual tests for the selection of decisive cases, as determined by the DSI method. Ranges of confidence intervals for individual neuropsychological tests and image processing methods were smaller than 0.3%.

i.e., the DSI thresholds, increase consistently with each comparison closer to *T*-0, as do the numbers of control subjects, i.e., baseline SMCIs, in the decisive groups, as seen in Fig. 4.

Driving factors for the selection of decisive cases, as determined by the DSI method, are depicted in Fig. 5. The data with the most classifying power were neuropsychological tests (their share of influence to selection being 37.5%), followed by MRI (25.7%), CSF (19.0%), and finally APOE (17.9%). Neuropsychological tests and MRI are composed of several individual tests and image processing methods, whose influence within these modalities are also noted in Fig. 5.

DISCUSSION

By statistically analyzing a large body of biomarkers and neuropsychological test data, one can get a comprehensive and objective evidence-based estimate of a patient's disease state. In this study, the DSI method was applied to early MCI data to determine the proportion of subjects whose data contain strong evidence of AD, or clear indication that the subject should remain stable during the coming months. By selecting subjects with the most and least evidence of early AD, subgroups of decisive patients were formed such that the accuracy of the selection reached a predetermined target accuracy. In this study, target accuracy was set at 87.7% by modeling the amount of evidence available in data when clinical AD diagnoses are made. This data-driven target accuracy was similar to the level of agreement between clinical and postmortem diagnoses in previous clinical samples. Ideally, the accuracies that can be reached with the DSI method should be assessed in future prospective studies with neuropathological follow-ups.

Two years before receiving clinical AD diagnoses, approximately one in four subjects had strong evidence regarding the disease in their measurement data, allowing classification at the target accuracy. At *T*-12 months, i.e., one year before AD diagnoses, approximately half of the subjects were included into the group of decisive cases. These results suggest that half of the patients who waited for their AD diagnoses for one or more years could have been considered eligible for diagnosis at least a year earlier, if identified correctly. Their early signs of AD were evident in the measurement data and being included in the decisive group implies correct prediction close to 90% of the time, similar to clinical diagnoses. In addition to potential AD converters, the method revealed, with similar accuracy, that there are subjects who will likely remain stable for 24 months, based on their data. For some subjects this information could be considered as important as having found strong indications of early AD.

Sensitivities and specificities for the decisive cases selected with the DSI method are evenly distributed, even though classification accuracy alone was driving the analyses. Confidence intervals of the results develop logically with disease progression, as do the DSI thresholds allowing inclusion into the decisive groups. Furthest from the AD diagnoses, confidence intervals are largest, and they become smaller close to *T*-0. Similarly, DSI thresholds restrict inclusion into the decisive groups the most at *T*-24, but with each visit closer to AD conversion, they allow more and more subjects to be included. The consistence of these properties gives indication that the identification process using the DSI method works in a robust manner.

Since patients exist whose measurement data could support earlier diagnosis, the next step is to provide a clinically feasible method for identifying these patients. Obviously, if similar patient data are available, it is possible to utilize the DSI thresholds determined here. In that case, DSI values above 0.76 would predict that the patient shall convert to AD within 24 months with at least 87.7% accuracy. Similarly, a value of DSI <0.24 indicates that the patient is likely to remain stable MCI for at least 24 months. The proposed method does not expect this particular dataset to be available, though. Custom DSI thresholds can be extracted from any data. The target accuracy for predicting outcomes may be manually set by a clinician if a certain level for decision support is required. After the target accuracy has been determined, either manually or in a data-driven manner, new DSI thresholds can be derived. These act as limits for the level of evidence needed in data and allow identifying patients whose disease state has progressed far enough to warrant consideration for an early diagnosis. The DSI thresholds could thus be computed to answer a particular clinical question, such as predicting conversion to AD within the coming 12 months at an accuracy of 85% or identifying MCI subjects who should not progress to AD during the next 4 years. To properly study these clinically interesting predictions with unseen patient data, there is need for a separate study where the test data are not aligned according to AD conversions, as they were here. This work is already underway to ensure that the results generalize to more realistic situations where the delays in potential conversion to AD cannot be known.

In addition to identifying the patients, analysis results must be presented to clinicians in a way that leaves making the decision in their hands. The DSF visualization has been developed for this purpose, providing an objective view to patient data as analyzed by the DSI method. In this cohort, neuropsychological testing (comprising in order of influence NeuroBat, ADAS, and MMSE) provided the most important set of variables for the patient identification process. MRI was ranked second, with all image processing methods having equal value. CSF and APOE were the least influential measures. Despite the differences in relevancies of tests, DSI as a method considers the congruence of all tests and measurements together. Thus, leaving out any one of them would impact the number of patients that can be identified. Since the DSF visualization clearly indicates which tests and measures are contributing to high DSI values and allows considering all the evidence comprehensively, clinicians using the proposed method should be better informed when making decisions regarding early diagnoses. In actual diagnostic work, APOE genotype, age, and gender, should be used as a background profile for analyzing patient data rather than being omitted or used as classification features. Issues that would then rise from stratifying training datasets into smaller groups can be countered, for example, by methods reported in [37]. In this work, a conscious choice was made against division to subgroups based on genotype, gender, age, or other features. The non-personalized group level results reported here are intended to provide a baseline for the amount of information available in patient data. Future work for developing a fully featured software tool providing automatic identification of patients with strong evidence of early AD, based on their measurement data, is currently underway. This work builds on an earlier implementation of a clinical decision support tool discussed in [32].

The main limitation of this study is the use of clinical diagnoses as a reference for building the disease models and evaluating results. Without neuropathological confirmations, it is impossible to determine whether the diagnoses suggested by an automated decision support tool indicate better or worse diagnostic accuracy than clinical diagnoses. Also, results from this study are not easily comparable to previous research, since, to the authors' best knowledge, similar analyses have not been done previously. Nevertheless, prediction accuracies obtained in the study are in line with current research and give indication that the conclusions regarding proportions of patients that could be identified early are relevant. Subjects in this study were selected into the conclusive groups of SMCIs and PMCIs based on their total DSI values. To allow more personalized diagnostics of early AD, data analysis should consider the effects of age, gender, and genotype to the results. In addition, due to ADNI inclusion criteria, results from this work apply to motivated MCI patients selected by expert memory centers, with exclusions for confounding disease, medications, etc. As such, the work demonstrates a data analysis concept, not a clinically verified solution. In regards to clinical application of the proposed method, suitable training datasets and integration to hospital information systems are essential. Studies with prospective patients are currently in preparation. In these studies, the method's usefulness and acceptance in clinical settings will be evaluated. They will also provide knowledge about applying the proposed method to datasets that are less controlled than ADNI.

CONCLUSION

A considerable proportion of MCI subjects have strong evidence of early AD in their measurement data several months before receiving their diagnoses. By selecting patients whose disease state has the strongest evidence of the disease, diagnostic outcomes can be predicted from patient data more accurately than otherwise possible. Identifying these patients at an early phase of the disease could make clinicians consider them eligible for earlier AD diagnosis, allowing administration of disease modifying therapies and medications at the earliest possible time. An alternative use for the proposed approach is to optimize patient selection for drug-trials or psycho-social treatments.

ACKNOWLEDGMENTS

This work was supported by the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist); EU-Grant-224328-PredictAD; Name: From Patient Data to Personalised Healthcare in Alzheimer's Disease). In addition, the study was supported by University of Eastern Finland for UEF-BRAIN consortium (HS).

Authors' disclosures available online (http://www.j-alz.com/disclosures/view.php?id=1430).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the Rev August 16 2011 University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

Funding parties had no active role in the research.

REFERENCES

[1] Morris JC (2005) Early-stage and preclinical Alzheimer disease. *Alzheimer Dis Assoc Disord* **19**, 163-165.

[2] Petersen RC, Roberts RO, Knopman DS, Boeve BF, Geda YE, Ivnik RJ, Smith GE, Jack CR Jr (2009) Mild cognitive impairment: Ten years later. *Arch Neurol* **66**, 1447-1455.

[3] Larrieu S, Letenneur L, Orgogozo JM, Fabrigoule C, Amieva H, Le Carret N, Barberger-Gateau P, Dartigues JF (2002) Incidence and outcome of mild cognitive impairment in a population-based prospective cohort. *Neurology* **59**, 1594-1599.

[4] Ganguli M, Snitz BE, Saxton JA, Chang CC, Lee CW, Vander Bilt J, Hughes TF, Loewenstein DA, Unverzagt FW, Petersen RC (2011) Outcomes of mild cognitive impairment by definition: A population study. *Arch Neurol* **68**, 761-767.

[5] Cattel C, Gambassi G, Sgadari A, Zuccalà G, Carbonin P, Bernabei R (2000) Correlates of delayed referral for the diagnosis of dementia in an outpatient population. *J Gerontol A Biol Sci Med Sci* **55**, M98-M102.

[6] Fiske A, Gatz M, Aadnøy B, Pedersen NL (2005) Assessing age of dementia onset: Validity of informant reports. *Alzheimer Dis Assoc Disord* **19**, 128-134.

[7] Speechly CM, Bridges-Webb C, Passmore E (2008) The pathway to dementia diagnosis. *MJA* **189**, 487-489.

[8] Bond J, Stave C, Sganga A, Vincenzino O, O'Connell B, Stanley R (2005) Inequalities in dementia care across Europe: Key findings of the Facing Dementia Survey. *Int J Clin Pract* **59**, 8-14.

[9] Ramakers IH, Visser PJ, Aalten P, Boesten JH, Metsemakers JF, Jolles J, Verhey FR (2007) Symptoms of preclinical dementia in general practice up to five years before dementia diagnosis. *Dement Geriatr Cogn Disord* **24**, 300-306.

[10] Duara R, Barker W, Loewenstein D, Bain L (2009) The basis for disease-modifying treatments for Alzheimer's disease: The sixth annual mild cognitive impairment symposium. *Alzheimers Dement* **5**, 66-74.

[11] Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, Barberger-Gateau P, Delacourte A, Frisoni G, Fox NC, Galasko D, Gauthier S, Hampel H, Jicha GA, Meguro K, O'Brien J, Pasquier F, Robert P, Rossor M, Salloway S, Sarazin M, de Souza LC, Stern Y, Visser PJ, Scheltens P (2010) Revising the definition of Alzheimer's disease: A new lexicon. *Lancet Neurol* **9**, 1118-1127.

[12] Lim A, Tsuang D, Kukull W, Nochlin D, Leverenz J, McCormick W, Bowen J, Teri L, Thompson J, Peskind ER, Raskind M, Larson EB (1999) Clinico-neuropathological correlation of Alzheimer's disease in a community-based case series. *J Am Geriatr Soc* **47**, 564-569.

[13] Petrovitch H, White LR, Ross GW, Steinhorn SC, Li CY, Masaki KH, Davis DG, Nelson J, Hardman J, Curb JD, Blanchette PL, Launer LJ, Yano K, Markesbery WR (2001) Accuracy of clinical criteria for AD in the Honolulu-Asia Aging Study, a population-based study. *Neurology* **57**, 226-234.

[14] Kazee AM, Eskin TA, Lapham LW, Gabriel KR, McDaniel KD, Hamill RW (1993) Clinicopathologic correlates in Alzheimer disease: Assessment of clinical and pathologic diagnostic criteria. *Alzheimer Dis Assoc Disord* **7**, 152-164.

[15] Jacobs HI, Van Boxtel MP, van der Elst W, Burgmans S, Smeets F, Gronenschild EH, Verhey FR, Uylings HB, Jolles J (2011) Increasing the diagnostic accuracy of medial temporal lobe atrophy in Alzheimer's disease. *J Alzheimers Dis* **25**, 477-490.

[16] Mattsson N, Zetterberg H, Hansson O, Andreasen N, Parnetti L, Jonsson M, Herukka SK, van der Flier WM, Blankenstein MA, Ewers M, Rich K, Kaiser E, Verbeek M, Tsolaki M, Mulugeta E, Rosén E, Aarsland D, Visser PJ, Schröder J, Marcusson J, de Leon M, Hampel H, Scheltens P, Pirttilä T, Wallin A, Jönhagen ME, Minthon L, Winblad B, Blennow K (2009) CSF biomarkers and incipient Alzheimer disease in patients with mild cognitive impairment. *JAMA* **302**, 385-393.

[17] Vos S, van Rossum I, Burns L, Knol D, Scheltens P, Soininen H, Wahlund LO, Hampel H, Tsolaki M, Minthon L, Handels R, L'italien G, van der Flier W, Aalten P, Teunissen C, Barkhof F, Blennow K, Wolz R, Rueckert D, Verhey F, Visser PJ (2012) Test sequence of CSF and MRI biomarkers for prediction of AD in subjects with MCI. *Neurobiol Aging* **33**, 2272-2281.

[18] Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, Cummings J, Delacourte A, Galasko D, Gauthier S, Jicha G, Meguro K, O'brien J, Pasquier F, Robert P, Rossor M, Salloway S, Stern Y, Visser PJ, Scheltens P (2007) Research criteria for the diagnosis of Alzheimer's disease: Revising the NINCDS-ADRDA criteria. *Lancet Neurol* **6**, 734-746.

[19] Clifford JR Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, Phelps CH (2011) Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 257-262.

[20] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, Gamst A, Holtzman DM, Jagust WJ, Petersen RC, Snyder PJ, Carrillo MC, Thies B, Phelps CH (2011) The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 270-279.

[21] Misra C, Fan Y, Davatzikos C (2008) Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. *Neuroimage* **44**, 1415-1422.

[22] Hinrichs C, Singh V, Xu G, Johnson SC, Alzheimers Disease Neuroimaging Initiative (2011) Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. *NeuroImage* **55**, 574-589.

[23] Davatzikos C, Bhatt P, Shaw LM, Batmanghelich KN, Trojanowski JQ (2011) Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol Aging* **32**, 2322.e19-27.

[24] Llano DA, Laforet G, Devanarayan V (2011) Derivation of a new ADAS-cog composite using tree-based multivariate analysis: Prediction of conversion from mild cognitive impairment to Alzheimer disease. *Alz Dis Assoc Dis* **25**, 73-84.

[25] Clifford JR Jr, Shiung MM, Weigand SD, O'Brien PC, Gunter JL, Boeve BF, Knopman DS, Smith GE, Ivnik RJ, Tangalos EG, Petersen RC (2005) Brain atrophy rates predict subsequent clinical conversion in normal elderly and amnestic MCI. *Neurology* **65**, 1227-1231.

[26] Borroni B, Anchisi D, Paghera B, Vicini B, Kerrouche N, Garibotto V, Terzi A, Vignolo LA, Di Luca M, Giubbini R, Padovani A, Perani D (2006) Combined 99mTc-ECD SPECT and neuropsychological studies in MCI for the assessment of conversion to AD. *Neurobiol Aging* **27**, 24-31.

[27] Toledo-Morrell L, Stoub TR, Bulgakova M, Wilson RS, Bennett DA, Leurgans S, Wuu J, Turner DA (2004) MRI-derived entorhinal volume is a good predictor of conversion from MCI to AD. *Neurobiol Aging* **25**, 1197-1203.

[28] Westman E, Simmons A, Muehlboeck JS, Mecocci P, Vellas B, Tsolaki M, Kłoszewska I, Soininen H, Weiner MW, Lovestone S, Spenger C, Wahlund LO; AddNeuroMed consortium; Alzheimer's Disease Neuroimaging Initiative (2011) AddNeuroMed and ADNI: Similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. *NeuroImage* **58**, 818-828.

[29] Tondelli M, Wilcock GK, Nichelli P, De Jager CA, Jenkinson M, Zamboni G (2012) Structural MRI changes detectable up to ten years before clinical Alzheimer's disease. *Neurobiol Aging* **33**, e825-e836.

[30] Mattila J, Koikkalainen J, Virkki A, Simonsen A, van Gils M, Waldemar G, Soininen H, Lötjönen J; Alzheimer's Disease Neuroimaging Initiative (2011) A disease state fingerprint for evaluation of Alzheimer's disease. *J Alzheimers* **27**, 163-176.

[31] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack C, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005) The Alzheimer's disease neuroimaging initiative. *Neuroimaging Clin N Am* **15**, 869-877.

[32] Mattila J, Koikkalainen J, Virkki A, van Gils M, Lötjönen J, Alzheimer's Disease Neuroimaging Initiative (2012) Design and application of a generic clinical decision support system for multiscale data. *IEEE Trans Biomed Eng* **59**, 234-240.

[33] McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM (1984) Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology* **34**, 939-944.

[34] Ashburner J, Friston KJ (2000) Voxel-based morphometry – the methods. *NeuroImage* **11**, 805-821.

[35] Koikkalainen J, Lötjönen J, Thurfjell L, Rueckert D, Waldemar G, Soininen H (2011) The Alzheimer's Disease Neuroimaging Initiative. Multi-template tensor-based morphometry: Application to analysis of Alzheimer's disease. *NeuroImage* **56**, 1134-1144.

[36] Lötjönen J, Wolz R, Koikkalainen J, Julkunen V, Thurfjell L, Lundqvist R, Waldemar G, Soininen H, Rueckert D; The Alzheimer's Disease Neuroimaging Initiative (2011) Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimer's disease. *NeuroImage* **56**, 185-196.

[37] Koikkalainen J, Pölönen H, Mattila J, van Gils M, Soininen H, Lötjönen J; Alzheimer's Disease Neuroimaging Initiative (2012) Improved classification of Alzheimer's disease data via removal of nuisance variability. *PLoS One* **7**, e31112.

PUBLICATION V

Predicting AD conversion: Comparison between prodromal AD guidelines and computer assisted PredictAD tool

In: PLOS ONE 8(2), e55246.

Predicting AD Conversion: Comparison between Prodromal AD Guidelines and Computer Assisted PredictAD Tool

Yawu Liu[1,2], Jussi Mattila[3], Miguel Ángel Muñoz Ruiz[1], Teemu Paajanen[1], Juha Koikkalainen[3], Mark van Gils[3], Sanna-Kaisa Herukka[1], Gunhild Waldemar[4], Jyrki Lötjönen[3], Hilkka Soininen[1]*, for The Alzheimer's Disease Neuroimaging Initiative¶

1 Department of Neurology, University of Eastern Finland, Kuopio University Hospital, Kuopio, Finland, **2** Department of Clinical Radiology, University of Eastern Finland, Kuopio University Hospital, Kuopio, Finland, **3** VTT Technical Research Centre of Finland, Tampere, Finland, **4** Department of Neurology, Memory Disorders Research Group, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark

Abstract

Purpose: To compare the accuracies of predicting AD conversion by using a decision support system (PredictAD tool) and current research criteria of prodromal AD as identified by combinations of episodic memory impairment of hippocampal type and visual assessment of medial temporal lobe atrophy (MTA) on MRI and CSF biomarkers.

Methods: Altogether 391 MCI cases (158 AD converters) were selected from the ADNI cohort. All the cases had baseline cognitive tests, MRI and/or CSF levels of Aβ1–42 and Tau. Using baseline data, the status of MCI patients (AD or MCI) three years later was predicted using current diagnostic research guidelines and the PredictAD software tool designed for supporting clinical diagnostics. The data used were 1) clinical criteria for episodic memory loss of the hippocampal type, 2) visual MTA, 3) positive CSF markers, 4) their combinations, and 5) when the PredictAD tool was applied, automatically computed MRI measures were used instead of the visual MTA results. The accuracies of diagnosis were evaluated with the diagnosis made 3 years later.

Results: The PredictAD tool achieved the overall accuracy of 72% (sensitivity 73%, specificity 71%) in predicting the AD diagnosis. The corresponding number for a clinician's prediction with the assistance of the PredictAD tool was 71% (sensitivity 75%, specificity 68%). Diagnosis with the PredictAD tool was significantly better than diagnosis by biomarkers alone or the combinations of clinical diagnosis of hippocampal pattern for the memory loss and biomarkers ($p \leq 0.037$).

Conclusion: With the assistance of PredictAD tool, the clinician can predict AD conversion more accurately than the current diagnostic criteria.

Citation: Liu Y, Mattila J, Ruiz MÁM, Paajanen T, Koikkalainen J, et al. (2013) Predicting AD Conversion: Comparison between Prodromal AD Guidelines and Computer Assisted PredictAD Tool. PLoS ONE 8(2): e55246. doi:10.1371/journal.pone.0055246

Editor: Kewei Chen, Banner Alzheimer's Institute, United States of America

Received May 30, 2012; **Accepted** December 27, 2012; **Published** February 12, 2013

Funding: This work was partially funded under the 7th Framework Programme by the European Commission (http://cordis.europa.eu/ist; EU-Grant-224328-PredictAD; Name: From Patient Data to Personalized Healthcare in Alzheimer's Disease), strategic funding from University of Eastern Finland for the UEFBRAIN Consortium (HS), EVO grants from Kuopio University Hospital, and by Health Research Council of The Academy of Finland, grant 121038 (HS and YL). The funders had no role in study design, analysis, decision to publish, or preparation of the manuscript. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California Los Angeles. This research was also supported by National Institutes of Health grants P30 AG010129 and K01 AG030514. The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report.

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

* E-mail: hilkka.soininen@uef.fi

¶ A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Introduction

Alzheimer's disease (AD) is the most common form of dementia in the elderly [1]. The pathology of AD starts years, even decades before any appearance of symptoms. The current hypothesis is that interventions should be started at an early phase in order to be efficient. Therefore, early diagnostics is essential 1) for detecting persons in clinical trials where pharmaceutical or psychosocial interventions are developed, and 2) for starting treatments at the earliest phase possible when efficient treatments become available in future. If one could have an intervention to delay disease onset or progression this would dramatically reduce the global burden of AD.

Mild cognitive impairment (MCI) is thought to represent the stage between normal forgetfulness due to aging and AD. Thus, MCI is a high risk factor for developing AD. However, due to heterogeneity of the MCI population the annual conversion rate varies from 4 to 31% between different studies/populations [2,3], and thus predicting which MCI cases will actually convert to AD is still a challenge. According to a recent proposal about new research criteria for AD, the diagnosis of AD requires that the patient displays the core criterion of significant episodic memory impairment, and exhibits at least one or more of the supportive biomarker criteria [4–8]. Dozens of clinical measures and AD biomarkers have been proposed [9]. The diagnosis process involves collaborative efforts from neurologists, psychologists, radiologists, geneticists, and laboratories to interpret demographic information, neuropsychological tests, and biomarkers. Several studies have shown that by combining biomarkers one achieve an improvement in accuracy of the AD diagnosis [10,11]. However, cognitive status does not always parallel the neuropathological changes due to the complex compensatory mechanisms present in AD. Therefore an accurate diagnosis of incipient/very early AD is not easy for the clinician, he/she is confronted by large amounts of quantitative and qualitative patient data, and particularly when much of the biomarker data may be ambiguous or even contradictory.

Recently, several computer-assisted support tools have been proposed as ways to help clinicians to make as accurate diagnoses as possible [12–14]. Decision support tools can provide objective and evidence-based information about the state of the patient; they are intended to integrate heterogeneous measurement data acquired from a patient in current clinical practice [12,13]. There is evidence that computer-assisted analyses of patient data can achieve comparable diagnostic accuracy as experienced clinicians [12,13]. The PredictAD tool can provide a classification and positions the patient into a continuous space between the values 0 and 1, indicating a patient's disease state in relation to previously known control (healthy) and positive (disease) populations [12,13]. This makes it possible to assess the disease severity i.e. it is not simply a yes/no diagnosis.

Many studies have been carried out to study the accuracy of biomarkers in detecting AD or predicting cognitive outcomes, however, there are few studies evaluating the relative importance of different biomarkers when they are used together. In some MCI cases, the biomarker data are ambiguous or contradict each another. It is unknown whether one of these biomarkers or their combination of them would be more sensitive, and whether quantitative values provide more information than a dichotomous rating [15]. In the present study, we grouped MCI cases from the Alzheimers disease Neuroimaging Initiative (ADNI) cohort (http://adni.loni.ucla.edu/) into four groups: high likelihood, intermediate likelihood, uninformative likelihood, and low likelihood of converting to AD [5]. We evaluated the accuracies of predicting the AD diagnosis made by quantitative analysis using the computer assisted PredictAD tool [12] and by using current guidelines of prodromal AD [4–8] as identified by combinations of dichotomized cognitive scores and visual assessment of middle temporal lobe atrophy on MRI and dichotomized CSF biomarkers. Our working hypothesis was that computer-assisted analysis could help to improve accuracy of the diagnosis.

Subjects and Methods

A total of 391 MCI cases were selected from the ADNI cohort (http://adni.loni.ucla.edu/). The demographics of the cases are summarized in Table 1. The definition of MCI is as follow: 1) subjects had Mini-Mental State Examination (MMSE) score between 24 and 30, 2) the memory complaint, 3) objective memory loss measured by education adjusted scores on Wechsler Memory Scale-Revised (WMS-R) Logical Memory II, 4) Clinical Dementia Rating (CDR) of 0.5, 5) the absence of significant levels of impairment in other cognitive domains, essentially preserved activities of daily living, and 6) the absence of dementia. All the cases had baseline ADNI cognitive testing results, including MMSE, Alzheimer's Disease Assessment Scale-Cognitive subscale (ADAS-Cog), and several other common neuropsychological tests (http://adni.loni.ucla.edu/).

Predicting AD Conversion with Current Prodromal AD Guidelines

The prediction of AD conversion was conducted with the combinations of clinical diagnosis of hippocampal pattern of memory loss [5] and biomarkers [16,17]. The episodic memory loss of the hippocampal type, which is characterized by a free recall deficit on testing not normalized with cueing [5], was defined as present when the scores of delayed recall and delayed recognition of Auditory Verbal Learning Test (RAVLT) [18] were lower than 1 standard deviation of the corresponding mean values in healthy aged people, i.e. RAVLT delayed recall <3 and RAVLT delayed recognition <10 [19]. The Scheltens Scale was used to categorize the visual medial temporal lobe atrophy (MTA) on MRI, The scale rates atrophy on a 5-point scale (0 = absent, 1 = minimal, 2 = mild, 3 = moderate and 4 = severe) [16]. A single experienced neuroradiologist (YL) evaluated MTA in all of the cases. Scheltens score ≥3 was considered as having significant MTA. CSF levels of Tau >93 pg/ml, and Amyloid beta 1–42 (Aβ1–42) <192 pg/ml were considered as positive CSF markers [17]. The likelihood of AD conversion was defined as follows [5]:

- High likelihood: all clinical core criteria (RAVLT tests), Scheltens scale and CSF markers were positive,
- Intermediate likelihood: clinical core criteria was positive, one of MRI and CSF markers was positive, but the other one was lacking, i.e., not available,
- Uninformative likelihood: clinical core criteria was positive, and one of MRI and CSF markers was positive, but the other one was negative.
- Low likelihood: all clinical core criteria, Scheltens scale, and CSF markers were negative.

Predict Conversion to AD with PredictAD Tool

The PredictAD tool [12] was used by one clinician who was blinded to the outcome during the evaluation. The PredictAD tool provided the rater with the available patient information at baseline, including demographics, apolipoprotein E (APOE) genotype, MMSE, ADAS-Cog, neuropsychological battery, MRI

Table 1. Demographics and clinical examinations for the MCI patients.

	Non-AD converter (n = 233)	AD converter (n = 158)	p value
Gender Male/Female	158/75	95/63	0.044
Age years	75±8	74±7	0.544
Years of education	16±3	16±3	0.969
ApoE alle 4 carrier	66 of 198 (33%)	145 of 193 (75%)	<0.001
MMSE	27.3±1.8	26.7±1.7	0.001
RAVLT delayed recall	3.7±3.6	1.5±2.1	<0.001
RAVLT delayed recognition	10.3±3.5	8.7±3.6	<0.001
ADAS-Cog total score (11-item)	10.3±4.2	13.3±4.1	<0.001
ADAS-Cog total score (13-item)	16.7±6.1	21.6±5.4	<0.001
Clock drawing test	4.4±0.8	3.9±1.1	<0.001
Digit span forward	8.2±2.0	8.2±2.0	0.940
Digit span backward	6.2±2.2	6.0±1.8	0.523
Category fluency	16.3±4.9	15.3±4.8	0.048
Trail making test-A	41.8±20.1	49.7±25.9	0.001
Trail making test-B	115.7±67.5	151.1±67.5	<0.001
Digit symbol substitution test	38.5±11.2	33.8±11.0	<0.001
Scheltens scale	1.8±0.9 (n = 230)	2.2±0.9 (n = 157)	<0.001
Tau	93±61 (n = 115)	118±57 (n = 84)	0.004
Aβ1–42	178±58 (n = 115)	144±39 (n = 84)	<0.001

doi:10.1371/journal.pone.0055246.t001

features automatically derived with FreeSurfer software package, and CSF laboratory analysis results. In addition, several features automatically derived from original MRI images using manifold learning [20], tensor-based morphometry [21], and hippocampus volume segmentation [22], developed in the PredictAD project (www.predictad.eu), were included. When determining with the assistance of PredictAD tool whether a subject had prodromal AD, the clinician based his opinion on presence of abnormal performances in the delayed recall and delayed recognition of Auditory RAVLT, the other neuropsychological tests were used as supportive evidences to determine the confidence of the clinical diagnosis. Given the baseline data, the clinician was then asked to categorize, i.e. diagnose, each patient into one of six categories: 1) clear indication of Non-AD, 2) probable indication of Non-AD, 3) subtle indication of Non-AD, 4) subtle indication of early AD, 5) probable indication of early AD, and 6) clear indication of early AD. One must emphasize that the clinician was asked to predict the diagnostic outcomes (Non-AD and AD converter) at the end of ADNI study using exclusively baseline data. To compare the accuracy of classification between automatically computed PredictAD diagnosis and clinician's diagnosis with assistance of PredictAD tool, Disease State Index (DSI) values, computed by the PredictAD tool, were categorized uniformly between 0 and 1 as follows: (1) Clear indication of Non-AD: DSI <0.17, (2) Probable indication of Non-AD: 0.17≤ DSI <0.33, (3) Subtle indication of Non-AD: 0.33≤ DSI <0.50, (4) Subtle indication of early AD: 0.50≤ DSI<0.67, (5) Probable indication of early AD: 0.67≤ DSI <0.83, and (6) Clear indication of early AD: ≥0.83. In the automatically computed PredictAD diagnosis, all the neuropsychological and genetic tests, MRI, and CSF data were used to calculate the DSI.

To test the reproducibility of the diagnosis by clinicians with the assistance of PredictAD tool, interobserver variability and intraobserver reproducibility were analyzed. To test the interobserver variability, two clinicians (Y.L. and M.M.) independently made diagnosis in 40 (10%) randomly selected cases. To test the intraobserver reproducibility, one clinician made diagnosis in the 40 cases with an interval of at least 6 months between the diagnosis sessions.

Statistical Analysis

The demographics and results of clinical exams were compared with Student t-test and chi square test between converters and non-converters. The conversion rates were calculated in cases with different likelihoods of AD conversion. The sensitivity, specificity, and accuracy of classification with the PredictAD tool, and different combinations of clinical scores, Scheltens scale, and CSF markers were calculated. McNemar's test was used to compare the differences in accuracy produced with the PredictAD tool and the current AD guidelines. Kappa test was used to test interobserver variability and intraobserver reproducibility. The difference was considered statistically significant if $p<0.05$.

Results

A total 387 of 391 MCI cases had undergone MRI exams, 199 MCI cases had undergone CSF examination, and 195 MCI cases had both MRI and CSF exams. During the 3-year follow-up, 158 of 391 (40%) converted to AD, 15 of 391 (4%) returned to normal cognitive status, and 218 MCI cases (56%) remained stable.

The conversion rates in different situations are summarized in Table 2.

Among the MCI cases who possessed a single positive marker (clinical core criteria or biomarker), those MCI cases who had increased Tau and decreased Aβ1–42 had the highest conversion rate (57%). The conversion rate for those MCI cases with Scheltens score≥3 was 55%. The MCI cases fulfilling the clinical

Table 2. Conversion rates of baseline MCI in different situations.

	Criteria	Cases	Converters (percentage)
Baseline MCI		391	158 (40%)
Hippocampal pattern of memory loss (clinical)	Auditory Verbal Learning Test (RAVLT) +	136	72 (53%)
Core biomarkers			
moderate to severe MTA	MRI +	92	51(55%)
increased Tau or decreased Aβ1–42	Tau or Aβ1–42 +	150	76 (51%)
increased Tau and decreased Aβ1–42	Tau and Aβ1–42 +	84	48 (57%)
High likelihood AD	RAVLT +, MRI +, CSF +	20	13 (65%)
Low likelihood AD	RAVLT − and biomarkers −	29	2 (7%)
Intermediate likelihood AD	RAVLT +, one biomarker +, and one not available	21	12 (57%)
no Scheltens scale	RAVLT + and Tau or Aβ1–42 +	2	1 (50%)
no CSF markers	RAVLT + and MRI +	19	11 (58%)
Uninformative likelihood AD	RAVLT +, one biomarker +, and one −	58	37 (64%)
negative MRI	RAVLT + and Tau or Aβ1–42 +	41	24 (59%)
negative CSF markers	RAVLT + and MRI +	17	13 (77%)

+ = positive finding.
doi:10.1371/journal.pone.0055246.t002

core criteria for episodic memory loss evident both on free recall and recognition had the lowest conversion rate (53%).

As expected, the conversion rate was highest for those MCI subjects in high likelihood AD group (65%) and lowest for MCI subjects with low likelihood (7%). For the MCI cases with intermediate and uninformative likelihood of AD, the conversion rates were 57% and 64% respectively. Among the 20 baseline MCI cases estimated as high likelihood of AD, there were no significant differences in age, Scheltens score, concentrations of CSF Tau and Aβ1–42, AVLT scores, education years, gender, frequency of APOE e4 allele, or PredictAD DSI between converters (n = 13) and non-converters (n = 7) ($p \geq 0.354$).

Sensitivity, Specificity, and Accuracy using Different Criteria and PredictAD Tool

The sensitivity, specificity, and accuracy of classification using the PredictAD tool and different criteria are listed in Table 3.

The criteria of increased CSF Tau or decreased Aβ1–42 achieved the highest sensitivity (90%), but the lowest specificity (36%). The criteria that included episodic memory loss of the hippocampal type, Scheltens scale ≥ 3, increased CSF Tau, and decreases Aβ1–42 could correctly detect 111 of 115 non-AD converters, producing the highest specificity (98%), but the lowest sensitivity (6%).

The PredictAD tool produced the highest accuracy 72%, followed by the clinician's diagnosis with the assistance of the PredictAD tool (71%). There was no significant difference in accuracy between the diagnosis by Predict tool alone and by the clinician (p = 1.0). The accuracy of the diagnosis by PredictAD tool alone was significantly higher than if one used the criteria of the biomarkers alone or combinations of clinical diagnosis of hippocampal pattern of memory loss and biomarkers ($p \leq 0.037$).

When considering the six categories of diagnosis (from clear indication of early AD to clear indication of non-AD), the interobserver variability and intraobserver reproducibility showed moderate agreements (kappa = 0.403, $p < 0.001$; kappa = 0.462, $p < 0.001$, respectively). However, when we simplified the six categories of diagnosis into AD and non-AD groups, excellent

Table 3. Sensitivity, specificity, and accuracy (percentage) of classification between AD converters and non-converters with different combinations of examinations and use of the PredictAD tool (All MCI cases).

	Criteria	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy
Neuropsychology tests (1)	Auditory Verbal Learning Test (RAVLT) +	46 (38–54)	73 (66–78)	62
Visual MTA (2)	MRI +	32 (25–40)	82 (76–87)	62
CSF (3a)	Tau or Aβ1–42 +	90 (82–96)	36 (27–45)	59
CSF (3b)	Tau and Aβ1–42 +	57 (46–68)	70 (60–78)	64
1+2		17 (12–24)	93 (89–96)	63
1+3a		44 (33–55)	78 (69–85)	64
1+2+3a		18 (11–28)	91 (84–96)	60
1+2+3b		4 (1–11)	97 (92–99)	58
PredictAD tool	Cutoff value of disease state index 0.50	73 (66–80)	71 (64–76)	72
Clinician with PredictAD tool assistance	Scale 1–3 stable MCI, scale 4–6 AD converter	75 (68–82)	68 (62–74)	71

doi:10.1371/journal.pone.0055246.t003

Table 4. Sensitivity, specificity, and accuracy (percentage) of classification between AD converters and non-converters with different combinations of examinations and use of the PredictAD tool (195 MCI cases with both MRI and CSF results).

	Criteria	Sensitivity (95% CI)	Specificity (95% CI)	Accuracy
Neuropsychology tests (1)	Auditory Verbal Learning Test (RAVLT) +	48 (37–59)	70 (60–78)	61
Visual MTA (2)	MRI +	31 (22–43)	83 (75–89)	61
CSF (3a)	Tau or Aβ1–42 +	90 (81–95)	36 (27–45)	59
CSF (3b)	Tau and Aβ1–42 +	57 (45–67)	71 (61–79)	65
1+2		19 (12–30)	94 (87–97)	62
1+3a		43 (33–55)	79 (70–86)	64
1+2+3a		18 (11–28)	91 (84–95)	60
1+2+3b		4 (1–11)	97 (92–99)	57
PredictAD tool	Cutoff value of disease state index 0.50	76 (65–84)	71 (61–79)	73
Clinician with PredictAD tool assistance	Scale 1–3 stable MCI, scale 4–6 AD converter	78 (68–86)	68 (58–76)	72

doi:10.1371/journal.pone.0055246.t004

agreements were achieved (kappa = 0.800, p<0.001 for interobserver variability; kappa = 0.850, p<0.001 for intraobserver reproducibility).

The PredictAD DSI achieved accuracy of 81% in detecting non-AD converters, and an accuracy of 63% in detecting AD converters. In the clinician's diagnosis with the assistance of the PredictAD tool, the accuracies were 80% and 62% respectively. However, with the assistance of PredictAD tool, the clinician's diagnosis of high confidence (clear non-AD, probable non-AD, probable AD, and clear AD) was dramatically improved compared to the PredictAD tool alone. The number of non-AD diagnoses made by the clinician with high confidence increased from 118 to 146 (from 30% to 37%), and the number of AD diagnosis with high confidence increased from 87 to 112 (from 22% to 29%). With help of the PredictAD tool, the clinician made diagnoses of clear non-AD or clear AD in 144 of 391 (37%) cases with overall accuracy of 84% (Tables 4, 5, 6).

Table 5. Accuracy of classification between AD converters and non-converters with the PredictAD tool.

	Final Diagnosis			Total	Accuracy
	AD	Healthy	MCI		
Clear indication of non-AD	2	9	43	54 (14%)	96%
Probable indication of non-AD	9	4	51	64 (16%)	86%
Subtle indication of non-AD	27	2	53	82 (21%)	67%
Indication of Non AD	38	15	147	200 (51%)	81%
Subtle indication of AD	58	0	46	104 (27%)	56%
Probable indication of AD	51	0	20	71 (18%)	72%
Clear indication of AD	11	0	5	16 (4%)	80%
Indication of AD	121	0	70	191 (49%)	63%

Note: Clear non-AD: disease state index <0.17, Probable non-AD: 0.17≤ disease state index <0.33, Subtle non-AD: 0.33≤ disease state index <0.50, Subtle AD: 0.50≤ disease state index <0.67, Probable AD: 0.67≤ disease state index <0.83, Clear AD: disease state index ≥0.83. 'Healthy' denotes MCI cases which converted back to the category 'healthy' during the study and belong still to the non-AD group. Overall accuracy of diagnosis was 72%.

doi:10.1371/journal.pone.0055246.t005

The clear AD diagnoses (16 cases) in the PredictAD DSI index included 5 stable MCI cases. The Probable indication of AD (71 cases) in the PredictAD DSI index included 20 stable MCI individuals. Among this subgroup there were no significant differences in age, gender, presence of APOE 4, years of education, concentrations of CSF markers, Scheltens scores, MMSE, or RAVLT results between AD converters and those with stable MCI (p≥0.236).

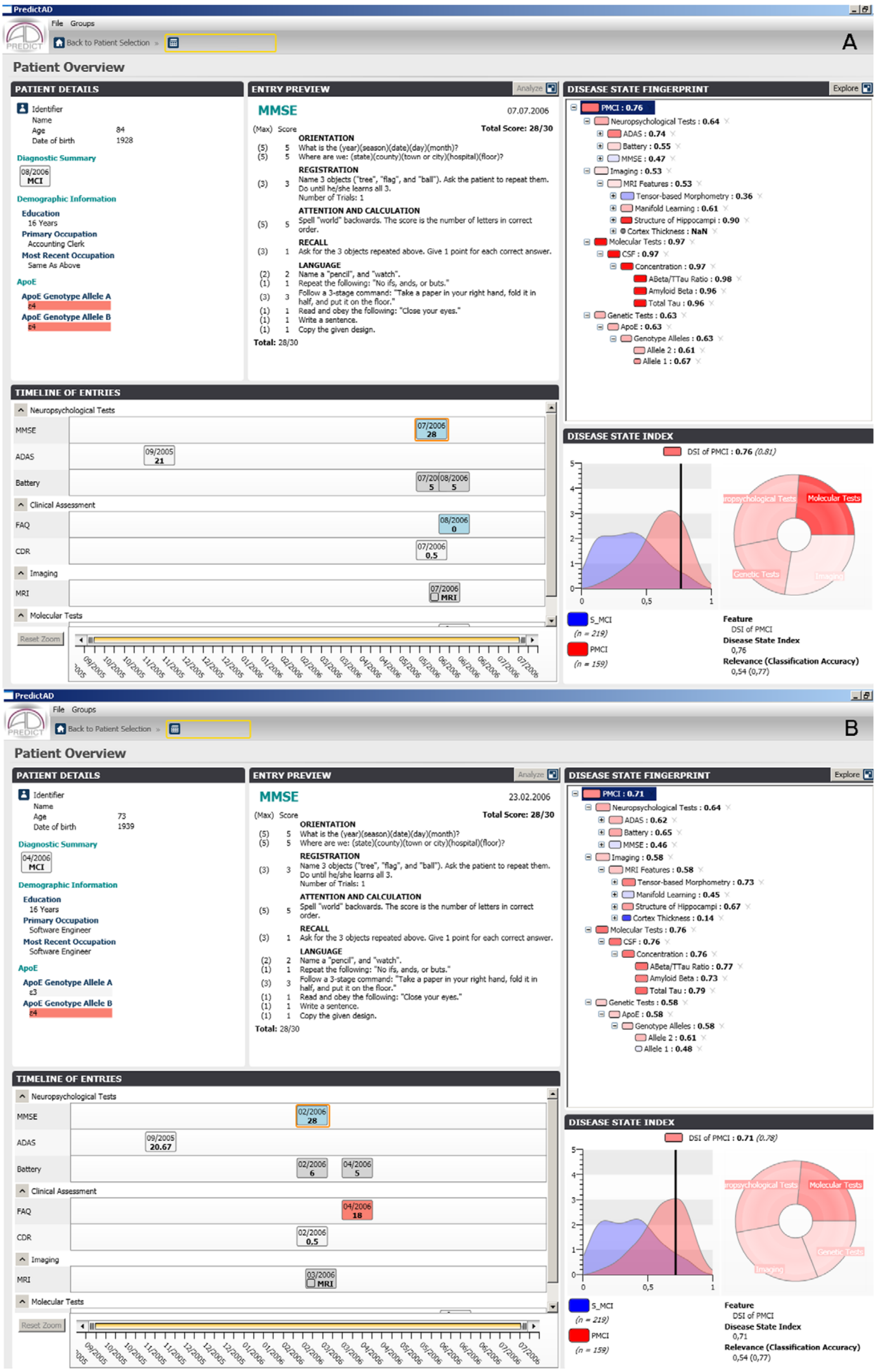
Because a variety of subject-specific factors may be influencing results in unkown ways, we also performed analyses on a subset of 195 participants who had all data available (neuropsychology, MRI and CSF) and repeated the analyses reported in Table 4 for this subset. The sensitivities, specificities, and accuracies of classifications using the PredictAD tool and different criteria in this subgroup were highly similar to those in whole group (Tables 3–4).

Table 6. Accuracy of classification between AD converters and non-converters the clinician making the diagnosis with assistance of the PredictAD tool.

	Final Diagnosis			Total	Accuracy
	AD	Healthy	MCI		
Clear indication of non-AD	6	12	64	82 (21%)	93%
Probable indication of non-AD	15	3	46	64 (16%)	77%
Subtle indication of non-AD	18	0	34	52 (13%)	65%
Indication of Non AD	39	15	144	198 (50%)	80%
Subtle indication of AD	43	0	38	81 (21%)	53%
Probable indication of AD	31	0	19	50 (13%)	62%
Clear indication of AD	45	0	17	62 (16%)	73%
Indication of AD	119	0	74	193 (50%)	62%

Overall accuracy of diagnosis was 71%.

doi:10.1371/journal.pone.0055246.t006



Patient Overview

PATIENT DETAILS

Identifier
Name
Age 84
Date of birth 1928

Diagnostic Summary
08/2006
MCI

Demographic Information
Education
16 Years
Primary Occupation
Accounting Clerk
Most Recent Occupation
Same As Above

ApoE
ApoE Genotype Allele A
ε4
ApoE Genotype Allele B
ε4

ENTRY PREVIEW

MMSE 07.07.2006

(Max)	Score	Total Score: 28/30
		ORIENTATION
(5)	5	What is the (year)(season)(date)(day)(month)?
(5)	5	Where are we: (state)(county)(town or city)(hospital)(floor)?
		REGISTRATION
(3)	3	Name 3 objects ("tree", "flag", and "ball"). Ask the patient to repeat them. Do until he/she learns all 3. Number of Trials: 1
		ATTENTION AND CALCULATION
(5)	5	Spell "world" backwards. The score is the number of letters in correct order.
		RECALL
(3)	1	Ask for the 3 objects repeated above. Give 1 point for each correct answer.
		LANGUAGE
(2)	2	Name a "pencil", and "watch".
(1)	1	Repeat the following: "No ifs, ands, or buts."
(3)	3	Follow a 3-stage command: "Take a paper in your right hand, fold it in half, and put it on the floor."
(1)	1	Read and obey the following: "Close your eyes."
(1)	1	Write a sentence.
(1)	1	Copy the given design.

Total: 28/30

DISEASE STATE FINGERPRINT

- PMCI : **0.76**
  - Neuropsychological Tests : **0.64**
    - ADAS : **0.74**
    - Battery : **0.55**
    - MMSE : **0.47**
  - Imaging : **0.53**
    - MRI Features : **0.53**
      - Tensor-based Morphometry : **0.36**
      - Manifold Learning : **0.61**
      - Structure of Hippocampi : **0.90**
      - Cortex Thickness : **NaN**
  - Molecular Tests : **0.97**
    - CSF : **0.97**
      - Concentration : **0.97**
        - ABeta/TTau Ratio : **0.98**
        - Amyloid Beta : **0.96**
        - Total Tau : **0.96**
  - Genetic Tests : **0.63**
    - ApoE : **0.63**
      - Genotype Alleles : **0.63**
        - Allele 2 : **0.61**
        - Allele 1 : **0.67**

TIMELINE OF ENTRIES

Neuropsychological Tests
MMSE 07/2006 28
ADAS 09/2005 21
Battery 07/20(5; 08/2006 5
Clinical Assessment
FAQ 08/2006 0
CDR 07/2006 0,5
Imaging
MRI 07/2006 MRI
Molecular Tests

DISEASE STATE INDEX

🟥 DSI of PMCI : **0.76** *(0.81)*

🟦 S_MCI
(n = 219)
🟥 PMCI
(n = 159)

Feature
DSI of PMCI
Disease State Index
0,76
Relevance (Classification Accuracy)
0,54 (0,77)

Patient Overview

PATIENT DETAILS

Identifier
Name
Age 73
Date of birth 1939

Diagnostic Summary
04/2006
MCI

Demographic Information
Education
16 Years
Primary Occupation
Software Engineer
Most Recent Occupation
Software Engineer

ApoE
ApoE Genotype Allele A
ε3
ApoE Genotype Allele B
ε4

ENTRY PREVIEW

MMSE 23.02.2006

(Max)	Score	Total Score: 28/30
		ORIENTATION
(5)	5	What is the (year)(season)(date)(day)(month)?
(5)	5	Where are we: (state)(county)(town or city)(hospital)(floor)?
		REGISTRATION
(3)	3	Name 3 objects ("tree", "flag", and "ball"). Ask the patient to repeat them. Do until he/she learns all 3. Number of Trials: 1
		ATTENTION AND CALCULATION
(5)	5	Spell "world" backwards. The score is the number of letters in correct order.
		RECALL
(3)	1	Ask for the 3 objects repeated above. Give 1 point for each correct answer.
		LANGUAGE
(2)	2	Name a "pencil", and "watch".
(1)	1	Repeat the following: "No ifs, ands, or buts."
(3)	3	Follow a 3-stage command: "Take a paper in your right hand, fold it in half, and put it on the floor."
(1)	1	Read and obey the following: "Close your eyes."
(1)	1	Write a sentence.
(1)	1	Copy the given design.

Total: 28/30

DISEASE STATE FINGERPRINT

- PMCI : **0.71**
  - Neuropsychological Tests : **0.64**
    - ADAS : **0.62**
    - Battery : **0.65**
    - MMSE : **0.46**
  - Imaging : **0.58**
    - MRI Features : **0.58**
      - Tensor-based Morphometry : **0.73**
      - Manifold Learning : **0.45**
      - Structure of Hippocampi : **0.67**
      - Cortex Thickness : **0.14**
  - Molecular Tests : **0.76**
    - CSF : **0.76**
      - Concentration : **0.76**
        - ABeta/TTau Ratio : **0.77**
        - Amyloid Beta : **0.73**
        - Total Tau : **0.79**
  - Genetic Tests : **0.58**
    - ApoE : **0.58**
      - Genotype Alleles : **0.58**
        - Allele 2 : **0.61**
        - Allele 1 : **0.48**

TIMELINE OF ENTRIES

Neuropsychological Tests
MMSE 02/2006 28
ADAS 09/2005 20.67
Battery 02/2006 6; 04/2006 5
Clinical Assessment
FAQ 04/2006 18
CDR 02/2006 0,5
Imaging
MRI 03/2006 MRI
Molecular Tests

DISEASE STATE INDEX

🟥 DSI of PMCI : **0.71** *(0.78)*

🟦 S_MCI
(n = 219)
🟥 PMCI
(n = 159)

Feature
DSI of PMCI
Disease State Index
0,71
Relevance (Classification Accuracy)
0,54 (0,77)

Figure 1. Screenshots from the PredictAD tool for two cases. The cases A and B had similar baseline neuropsychological tests, biomakers, and genetic tests, but the case A did not convert to AD, case B converted to AD during 3-year follow-up period. The case A was classified by both predictAD tool and current guildline for prodromal AD. It is probable that this case will convert in longer follow-up. The MCI subjects like case A seem to be a potential interesting study group. It might be possible to identify sensitive biomarkers to detect AD at early phase or explore novel preventative factors to delay the onset of symptoms of AD by investigating this subgroup. The main window of the PredictAD tool consists of five panels. The 'Patient details' panel shows basic information about the patient. The 'Timeline of entries' panel contains information about all measurements acquired from patient. The panel is interactive: the user can click any of the entries visible and a summary isshown in the 'Entry preview' panel. The disease state fingerprint is shown in the 'Disease state fingerprint' panel. When the user selects any of the item from the fingerprint, details behind the item are shown in the 'Disease state index' panel. The distributions show the probability density functions of the corresponding item for the study and control groups, in this case PMCI and SMCI groups, and the value measured from the patient is shown by a vertical black line.
doi:10.1371/journal.pone.0055246.g001

Discussion

The results show that the PredictAD tool alone (72%) and the clinician with the assistance of the PredictAD tool produced comparable or higher accuracy in predicting 3-year MCI outcome than current research criteria for diagnosis of prodromal AD. The literature is somewhat confusing, due to differences in size of study populations, statistical methods, and length of follow-up etc., but it seems that the overall accuracy of combinations of clinical data and/or biomarkers in predicting AD conversion from MCI has varied from 67% to 93% [23–28]. Liu et al., using the 100 MCI cases from AddNeuroMed data and a combination of neuropsychological tests and structural MRI biomarkers reported overall accuracy 69% during one year follow-up [27]. Studies with the ADNI cohort reported accuracies 67–77% when using combinations of clinical measures and CSF and MRI biomarkers [23,24,26].

We acknowledge that the prediction accuracy of about 70% is not high concerning the clinical utility but the result is still comparable with the current state-of-the-art. It reflects a reality that the current prodromal AD guidelines and combinations of biomarkers are not perfect. However, our point was not to develop a novel method but to show how the current guidelines compare with computer-assisted methods. The PredictAD tool can provide objective and evidence-based information about the state of the patient by integrating heterogeneous measurement data acquired from a patient in current clinical practice. PredictAD makes it possible to assess the disease severity, i.e. it is not simply a yes/no diagnosis. Its graphical user interface can make it easy for clinician to explore every single test or biomarker, giving more confident to clinicians than a probability or yes/no diagnosis calculated with certain software with underlying complex statistical calculation. Using the PredictAD tool, the clinician was able to detect a subpopulation for which the accuracy was 84% which starts to be high enough for affecting the clinical reasoning. It is good to remember that 100% is not the correct target value in reality due to different reasons: 1) Stable MCI and progressive MCI cases in ADNI are not pathologically confirmed cases. It has been shown in different studies that the agreement of the clinical and neuropathology diagnoses is 70–90% [29–31]. In other words, even 72% is within this range and studies reporting values >90% should be interpreted with a caution. 2) Even neuropathological diagnoses are not perfect.

It is interesting that about 30% MCI cases with clear (DSI $\geq$0.83, 5 of 16 cases) and probable (0.67$\leq$ DSI <0.83, 20 of 71 cases) indications of AD did not convert to AD during the 3-year follow-up, even though they did not significantly differ from AD converters in age, gender, presence of APOE4, years of education, concentrations of CSF markers, Scheltens scores, MMSE, and RAVLT results (Figure 1). The reason why those 25 stable MCI cases did not convert to AD is still unknown. In fact, this subgroup population seems to be interesting, and a detailed investigation of this subgroup, we might uncover novel preventative factors which delay the onset of symptoms of AD.

Current research criteria for prodromal AD [4] emphasizes that the core criteria of episodic memory impairment should not only include deficit on delayed free recall but also on cued recall or recognition. In this paper we used RAVLT free recall and recognition scores to form the criteria of episodic memory impairment. Adjustments for gender or education were not used, and in addition it can be argued that results may have been different if another memory test or cut-off values would have been used. However, it is essential to remind that all MCI subjects in ADNI cohort already fulfilled a significant memory impairment measured with WMS-R logical memory II test (with education correction). Thus subjects who fulfilled the criteria of episodic memory impairment in the present paper performed lower than expected for age altogether in three memory tests.

It has been shown that the Scheltens scale can classify AD patients and healthy controls or other types of dementia with high sensitivity, specificity, and accuracy [16,32,33]. Westman et al. [34] applied Scheltens scale 2 and 3 as cutoff values in 101 MCI cases from the multicenter study AddNeuroMed study. They reported that the visually evaluated atrophy of MTL produced similar accuracy in predicting conversion from MCI to AD (68%) compared to multivariate regional MRI classification and manual hippocampal volumes at one year follow-up. We applied Scheltens scale 3 as the cutoff value in the ADNI data and found prediction accuracy (62%) during the 3-year follow-up.

In the present study, according to the most recent criteria for likelihood of AD, only 25 cases fulfilled the high likelihood of AD, i.e. all clinical core criteria, MRI and CSF markers were positive, fifteen of those 25 (60%) cases did convert to AD. Moreover, very low sensitivities (6%–57%) were achieved by using the combination of clinical core tests, and MRI and CSF markers. In contrast, by using the PredictAD tool, the number of clinician's diagnosis of a clear indication of AD was 62 cases, and 45 of those 62 (73%) cases did convert to AD. This finding indicates that the PredictAD tool uses the clinical, MRI, and CSF data in a much more efficient way than the recent criteria applied with specific cut-off values for making the diagnosis of AD.

We acknowledge that the present study has certain limitation. In the predicting AD conversion with current prodromal AD guidelines, only RAVLT tests were used to define if the subjects had prodromal AD symptoms, but in the predicting AD conversion with PredictAD tool alone, all the neuropsychological tests were used. When the clinician determined if the subjects had prodromal AD symptoms with the assistance of PredictAD tool, only RAVLT tests were used as in the predicting AD conversion with prodromal AD guidelines. However, the clinician was not blinded to the other neuropsychological tests, the performance at the other tests exploring cognitive domains other than memory were used to increase the confidence of clinical diagnosis. The overall predicting accuracy was 72%, 71%, and 64% for the

PredictAD tool alone, clinician's prediction with the assistance of the PredictAD tool, and the best combination of the core clinical and biomarkers respectively. Diagnosis with the PredictAD tool was significantly more accurate than diagnosis by biomarkers alone or the combinations of clinical core criteria and biomarkers. The methods judging if a subject presented prodromal symptoms were not equal. It may explain the differences in overall predicting accuracy. The findings imply that a single neuropsychological test is not powerful enough to replace the other neuropsychological tests in early AD diagnosis, enhancing the justification of using PredictAD tool in clinical practice.

In conclusion, with the assistance of the PredictAD tool, the clinician can predict AD conversion more accurately than than the current research criteria for prodromal AD.

Acknowledgments

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data, but did not participate in analysis or writing of this report.

Author Contributions

Conceived and designed the experiments: YL TP JL HS. Critically revised the paper: YL JM MAMR TP JK MVG SKH GW JL HS. Analyzed the data: YL JM JL. Contributed reagents/materials/analysis tools: JM JK MVG SKH GW JL HS. Wrote the paper: YL JM TP JL HS.

References

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM (2007) Forecasting the global burden of alzheimer's disease. Alzheimers Dement 3: 186–191.
2. Bruscoli M, Lovestone S (2004) Is MCI really just early dementia? A systematic review of conversion studies. Int Psychogeriatr 16: 129–140.
3. Luis CA, Loewenstein DA, Acevedo A, Barker WW, Duara R (2003) Mild cognitive impairment: Directions for future research. Neurology 61: 438–444.
4. Dubois B, Feldman HH, Jacova C, Dekosky ST, Barberger-Gateau P, et al. (2007) Research criteria for the diagnosis of alzheimer's disease: Revising the NINCDS-ADRDA criteria. Lancet Neurol 6: 734–746.
5. Dubois B, Feldman HH, Jacova C, Cummings JL, Dekosky ST, et al. (2010) Revising the definition of alzheimer's disease: A new lexicon. Lancet Neurol 9: 1118–1127.
6. R JC,Jr, Albert MS, Knopman DS, McKhann GM, Sperling RA, et al. (2011) Introduction to the recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimers Dement 7: 257–262.
7. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, et al. (2011) The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimers Dement 7: 263–269.
8. Sperling RA, Aisen PS, Beckett LA, Bennett DA, Craft S, et al. (2011) Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimers Dement 7: 280–292.
9. Drago V, Babiloni C, Bartres-Faz D, Caroli A, Bosch B, et al. (2011) Disease tracking markers for alzheimer's disease at the prodromal (MCI) stage. J Alzheimers Dis 26 Suppl 3: 159–199.
10. Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, et al. (2009) MRI and CSF biomarkers in normal, MCI, and AD subjects: Predicting future clinical change. Neurology 73: 294–301.
11. Vemuri P, Wiste HJ, Weigand SD, Shaw LM, Trojanowski JQ, et al. (2009) MRI and CSF biomarkers in normal, MCI, and AD subjects: Diagnostic discrimination and cognitive correlations. Neurology 73: 287–293.
12. Mattila J, Koikkalainen J, Virkki A, Simonsen A, van Gils M, et al. (2011) A disease state fingerprint for evaluation of alzheimer's disease. J Alzheimer's Dis 27: 163–176.
13. Kloppel S, Stonnington CM, Barnes J, Chen F, Chu C, et al. (2008) Accuracy of dementia diagnosis: A direct comparison between radiologists and a computerized method. Brain 131: 2969–2974.
14. Kawamoto K, Houlihan CA, Balas EA, Lobach DF (2005) Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success. BMJ 330: 765.
15. Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, et al. (2011) The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. Alzheimers Dement 7: 270–279.
16. Scheltens P, Leys D, Barkhof F, Huglo D, Weinstein HC, et al. (1992) Atrophy of medial temporal lobes on MRI in "probable" alzheimer's disease and normal ageing: Diagnostic value and neuropsychological correlates. J Neurol Neurosurg Psychiatr 55: 967–972.
17. Shaw LM, Vanderstichele H, Knapik-Czajka M, Clark CM, Aisen PS, et al. (2009) Cerebrospinal fluid biomarker signature in alzheimer's disease neuroimaging initiative subjects. Ann Neurol 65: 403–413.
18. Rey A (1964) Lexamen clinique en psychologie. Paris: Presses Universitaires de France.
19. Ivnik RJ, Malec JF, Tangalos EG, Petersen RC, Kokmen E, et al. (1992) Mayòs older americans normative studies: Updated AVLT norms for ages 56 to 97. The Clinical Neuropsychologist 6: 83–104.
20. Wolz R, Aljabar P, Hajnal J, Rueckert D (2010) Manifold learning for biomarker discovery in MR imaging. 6357: 116–123.
21. Koikkalainen J, Lotjonen J, Thurfjell L, Rueckert D, Waldemar G, et al. (2011) Multi-template tensor-based morphometry: Application to analysis of alzheimer's disease. Neuroimage 56: 1134–1144.
22. Lotjonen J, Wolz R, Koikkalainen J, Julkunen V, Thurfjell L, et al. (2011) Fast and robust extraction of hippocampus from MR images for diagnostics of alzheimer's disease. Neuroimage 56: 185–196.
23. Ewers M, Walsh C, Trojanowski JQ, Shaw LM, Petersen RC, et al. (2012) Prediction of conversion from mild cognitive impairment to alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. Neurobiol Aging 33: 1203–1214.
24. Devanand DP, Liu X, Brown PJ, Huey ED, Stern Y, et al. (2012) A two-study comparison of clinical and MRI markers of transition from mild cognitive impairment to alzheimer's disease. Int J Alzheimers Dis 2012: 483469.
25. Devanand DP, Liu X, Tabert MH, Pradhaban G, Cuasay K, et al. (2008) Combining early markers strongly predicts conversion from mild cognitive impairment to alzheimer's disease. Biol Psychiatry 64: 871–879.
26. Cui Y, Liu B, Luo S, Zhen X, Fan M, et al. (2011) Identification of conversion from mild cognitive impairment to alzheimer's disease using multivariate predictors. PLoS One 6: e21896.
27. Liu Y, Paajanen T, Zhang Y, Westman E, Wahlund LO, et al. (2011) Analysis of regional MRI volumes and thicknesses as predictors of conversion from mild cognitive impairment to alzheimer's disease. Neurobiol Aging 31: 1375–1385.
28. B Bouwman FH, Schoonenboom SN, van der Flier WM, van Elk EJ, Kok A, et al. (2007) CSF biomarkers and medial temporal lobe atrophy predict dementia in mild cognitive impairment. Neurobiol Aging 28: 1070–1074.
29. Lim A, Tsuang D, Kukull W, Nochlin D, Leverenz J, et al. (1999) Clinico-neuropathological correlation of alzheimer's disease in a community-based case series. J Am Geriatr Soc 47: 564–569.
30. Petrovitch H, White LR, Ross GW, Steinhorn SC, Li CY, et al. (2001) Accuracy of clinical criteria for AD in the honolulu-asia aging study, a population-based study. Neurology 57: 226–234.
31. Kazee AM, Eskin TA, Lapham LW, Gabriel KR, McDaniel KD, et al. (1993) Clinicopathologic correlates in alzheimer disease: Assessment of clinical and pathologic diagnostic criteria. Alzheimer Dis Assoc Disord 7: 152–164.
32. Burton EJ, Barber R, Mukaetova-Ladinska EB, Robson J, Perry RH, et al. (2009) Medial temporal lobe atrophy on MRI differentiates alzheimer's disease from dementia with lewy bodies and vascular cognitive impairment: A prospective study with pathological verification of diagnosis. Brain 132: 195–203.
33. Wahlund LO, Julin P, Johansson SE, Scheltens P (2000) Visual rating and volumetry of the medial temporal lobe on magnetic resonance imaging in dementia: A comparative study. J Neurol Neurosurg Psychiatr 69: 630–635.
34. Westman E, Cavallin L, Muehlboeck JS, Zhang Y, Mecocci P, et al. (2011) Sensitivity and specificity of medial temporal lobe visual ratings and multivariate regional MRI classification in alzheimer's disease. PLoS One 6: e22506.

PUBLICATION VI

Application of the PredictAD Software Tool to Predict Progression in Patients with Mild Cognitive Impairment

In: Dementia and geriatric cognitive disorders
34(5–6), pp. 344–350.

Original Research Article

Dementia and Geriatric Cognitive Disorders

Dement Geriatr Cogn Disord 2012;34:344–350
DOI: 10.1159/000345554

Accepted: November 1, 2012
Published online: December 6, 2012

Application of the PredictAD Software Tool to Predict Progression in Patients with Mild Cognitive Impairment

Anja H. Simonsen[a] Jussi Mattila[b] Anne-Mette Hejl[a] Kristian S. Frederiksen[a] Sanna-Kaisa Herukka[c,d] Merja Hallikainen[c] Mark van Gils[b] Jyrki Lötjönen[b] Hilkka Soininen[c,d] Gunhild Waldemar[a] for the Alzheimer's Disease Neuroimaging Initiative

[a]Memory Disorders Research Group, Department of Neurology, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark; [b]VTT Technical Research Centre of Finland, Tampere, and [c]Institute of Clinical Medicine-Neurology, University of Eastern Finland, Yliopistoranta 1C and [d]Department of Neurology, Kuopio University Hospital, Kuopio, Finland

Key Words
Mild cognitive impairment · Alzheimer's disease · PredictAD software tool

Abstract

Background: The PredictAD tool integrates heterogeneous data such as imaging, cerebrospinal fluid biomarkers and results from neuropsychological tests for compact visualization in an interactive user interface. This study investigated whether the software tool could assist physicians in the early diagnosis of Alzheimer's disease. ***Methods:*** Baseline data from 140 patients with mild cognitive impairment were selected from the Alzheimer's Disease Neuroimaging Study. Three clinical raters classified patients into 6 categories of confidence in the prediction of early Alzheimer's disease, in 4 phases of incremental data presentation using the software tool. A 5th phase was done with all available patient data presented on paper charts. Classifications by the clinical raters were compared to the clinical diagnoses made by the Alzheimer's Disease Neuroimaging Initiative investigators. ***Results:*** A statistical significant trend ($p < 0.05$) towards better classification accuracy (from 62.6 to 70.0%) was found when using the PredictAD tool during the stepwise procedure. When the same data were presented on paper, classification accuracy of the raters dropped significantly from 70.0 to 63.2%. ***Conclusion:*** Best classification accuracy was achieved by the clinical raters when using the tool for decision support, suggesting that the tool can add value in diagnostic classification when large amounts of heterogeneous data are presented.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2012 S. Karger AG, Basel
1420–8008/12/0346–0344$38.00/0

Accessible online at:
www.karger.com/dem

Dr. Anja Hviid Simonsen
Memory Disorders Research Group
Department of Neurology; N6702
9, Blegdamsvej, DK–2100-Copenhagen (Denmark)
E-Mail anja.hviid.simonsen@rh.regionh.dk

Introduction

Mild cognitive impairment (MCI) is a term referring to persons who do not fulfill the criteria for dementia, but who exhibit some form of cognitive impairment [1, 2]. MCI is associated with an increased risk of developing Alzheimer's disease (AD) [3, 4]. Identification of MCI patients who will progress to AD would allow the application of disease-modifying treatments to slow progression at a point where clinical manifestations are limited. A combination of results from neuropsychological testing [5], MRI [6] and cerebrospinal fluid (CSF) biomarkers [7] can aid in the prediction of which patients with MCI will progress to AD. Furthermore, measurements of brain amyloid by PET using the ligand ^{11}C PIB (Pittsburg compound B) were shown to predict a 3-year conversion to AD in a group of patients with amnestic MCI [8].

Even a modest delay of 1 year in the onset and progression of disease could drastically reduce the burden of AD on society [9]. Current symptomatic treatments and non-pharmacological interventions are assumed to be most effective at the earliest stages of the disease, underlining the importance of early diagnosis [10]. PredictAD, funded by the 7th EU framework (FP7 – 224328), is a research project where a consortium of technical and clinical partners aims to provide standardized and objective solutions for enabling earlier diagnoses of AD, improved monitoring of treatment efficacy, easier patient selection for drug trials and improved cost-effectiveness of diagnostic protocols. For this purpose, the PredictAD consortium has developed a software tool to support clinical decision-making [11]. The PredictAD tool integrates heterogeneous data from clinical investigations in an individual patient, such as imaging, CSF biomarkers and results from neuropsychological tests for compact visualization in an interactive user interface. The aim of this study was to investigate whether the PredictAD tool could assist physicians in the early diagnosis of AD. The hypothesis was that when physicians were presented with clinical and paraclinical information from a patient by the software they would be able to predict conversion from MCI to AD better than if the information were presented in a traditional form on a printed chart.

Methods

Patients

Baseline data from 140 patients with MCI were selected from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study cohort [4]. The ADNI comprises a consortium of universities and medical centers in the USA and Canada. It was established to develop imaging techniques and biomarker procedures in normal subjects, subjects with MCI and subjects with mild AD. A total of 229 cognitively normal subjects, 398 MCI patients and 192 AD patients were recruited for the ADNI [4] which is supported by the National Institutes of Health, private pharmaceutical companies and nonprofit organizations. Full inclusion and exclusion criteria are described in detail at www.adni-info.org.

The criteria for selecting MCI patients in this study were: the availability of the Mini Mental State Examination (MMSE), Functional Assessment Questionnaire (FAQ), Alzheimer's Disease Assessment scale (ADAS-cog), Clinical Dementia Rating (CDR), CSF levels of amyloid beta and tau as well as MRI-derived volume values at the baseline measurement. Furthermore, patients who dropped out of the ADNI before 3 years of follow-up were excluded from our study. Some of the excluded patients had converted from MCI to AD at that time.

All selected patients had a CDR = 0.5 and an MMSE score ≥24 at baseline. In total, 140 of the 398 ADNI MCI patients met our criteria and were included in this study.

Table 1. Average age, MMSE and years of education of the MCI patients

	n	Gender M/F	Age, years (SD)	MMSE (SD)*	Years of education (SD)*
P-MCI	64	39/25	74.7 (7.0)	26.6 (1.9)	15.3 (3.0)
S-MCI	76	51/25	73.9 (7.8)	27.3 (1.7)	16.3 (2.7)

* Statistically significant difference ($p < 0.05$ from unpaired Student t test).

Diagnostic Classification of Patients According to the ADNI

In the ADNI study, patients with MCI were diagnosed according to the criteria of Petersen et al. [1]. Of the 140 patients included in our study, 64 (45.7%) progressed to AD (progressive MCI, P-MCI) during the ADNI follow-up period according to the National Institute of Neurological and Communicative Disorders and Stroke and Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) [12], whereas 76 remained in the MCI group (stable MCI, S-MCI). Demographic information is listed in table 1. Due to our selection criteria, the average time of conversion from MCI to AD (i.e. the P-MCI group) was 20 months, which is slightly longer than when considering all ADNI P-MCIs including the ones that were excluded because they dropped out before the end of the 3-year follow-up. The mean age of the patients was 74.3 years (SD 7.5), and there were 90 males and 50 females.

Stepwise Classification by Clinical Raters Using the PredictAD Tool

In PredictAD, a clinical decision support tool was developed for the early diagnosis of AD [11]. The tool was applied by 3 clinical raters who were informed that the patients to be evaluated had been diagnosed with MCI by the ADNI team at baseline. All cases presented to the raters were anonymized. The baseline data

Table 2. Protocol for stepwise data presentation using the PredictAD tool

	Available data
Phase 1	Age, gender, years of education, primary occupation + MMSE including subscores, FAQ including subscores and AD index calculated by the software
Phase 2	All from phase 1 + ADAS-cog including subscores and CDR including subscores
Phase 3	All from phases 1 and 2 + MRI volumetrics (provided by the ADNI and computed with the FreeSurfer image analysis suite [14])
Phase 4	All from phases 1, 2 and 3 + CSF levels of amyloid beta and total tau

were presented to the raters in a stepwise manner using the PredictAD tool. After each presentation of patient data, the clinical raters were asked to categorize the patient into one of 6 categories according to the likelihood that the patient would develop AD dementia: (1) clear indication of non-AD, (2) probable indication of non-AD, (3) subtle indication of non-AD, (4) subtle indication of early AD, (5) probable indication of early AD or (6) clear indication of early AD.

In other words, the clinical raters were asked to predict the 3-year conversion outcomes (S-MCI or P-MCI) using baseline data for cognitive tests, MRI and CSF biomarkers. The non-AD categories, i.e. (1)–(3), were defined to be used for subjects with memory problems thought to be due to causes other than early-phase AD.

As shown in table 2, baseline clinical data were presented to the clinical raters in a stepwise manner in 4 phases, each phase adding more information about the patient.

In phase 1, only simple demographic and clinical data (global and functional status) were presented. In subsequent phases, more clinical information was added, mimicking the routine clinical diagnostic process, starting with the less costly and invasive procedures and ending with the addition of CSF biomarkers at phase 4. The raters did not receive any form of feedback about their performance between the rating phases.

During each phase, the PredictAD tool provided only the allowed patient information to the clinical raters (see fig. 1). A timeline panel showed when the different tests were administered. Selecting a test from the timeline displayed it in a preview panel showing detailed results from the selected test. A tree of colored nodes on the right panel of the computer screen showed how patient data as a whole relate to average values from previously diagnosed S-MCI and P-MCI cases. Tests with large node sizes indicated patient measures which differentiated well between S-MCI and P-MCI. Red indicated patient data which are similar to P-MCI cases, whereas blue pointed towards S-MCI. White indicated measurements that were intermediate between the average values for previously diagnosed S-MCI and P-MCI cases. Values within the tree visualization are Disease State Index (DSI) values [13]. The DSI method is a novel statistical classification method that was developed with the goal of supporting clinical decision-making. Here, it was used for ranking patient data against data from known S-MCI and P-MCI cases on a scale of 0–1. A DSI value of zero denotes a perfect match with S-MCIs, and a DSI value of 1 denotes a perfect match to P-MCI cases. To compute these DSI values, all MCI patients from the ADNI were used for training the disease model, except the patient being currently classified equivalent to leave-one-out cross-validation. This was done to ensure that there are enough data for good-quality computerized analysis and to ensure that raters' predictions have no bias towards the correct categories when using the software tool. More rigorous analysis of the performance and characteristics of the DSI method underlying the tool's visualizations can be found in our previously published paper [11].

To minimize influence from previous phases, all 140 patients were categorized once before moving on to the next phase and restarting from the first patient with additional clinical information. Information about previous rater categorizations was not provided to the clinical raters during subsequent phases. The presentation of the subjects was always in the same order. Any retention from previous phases influenced categorizations similarly, whether using the tool or reading the paper charts.

Patient Classification by Clinical Raters Using Traditional Paper Charts

After categorizing patients into 4 phases with the tool, each rater was presented with exactly the same patient data as at phase 4 in a paper chart format at a separate rater meeting. Raters were once more requested to classify all patients into the 6 categories. This 5th phase resembled the current state of diagnostic work, i.e. no help from tools was available, only raw data and test results. Similar to the earlier phases using the tool, the raters were not provided with test norms or MCI cut-off scores for any of the data.

Clinical Raters

The raters were all physicians who had clinical experience in the diagnostic evaluation of dementia and experience with the cognitive tests and assessment scales used in the ADNI. However, they had not previously participated in any ADNI-related studies.

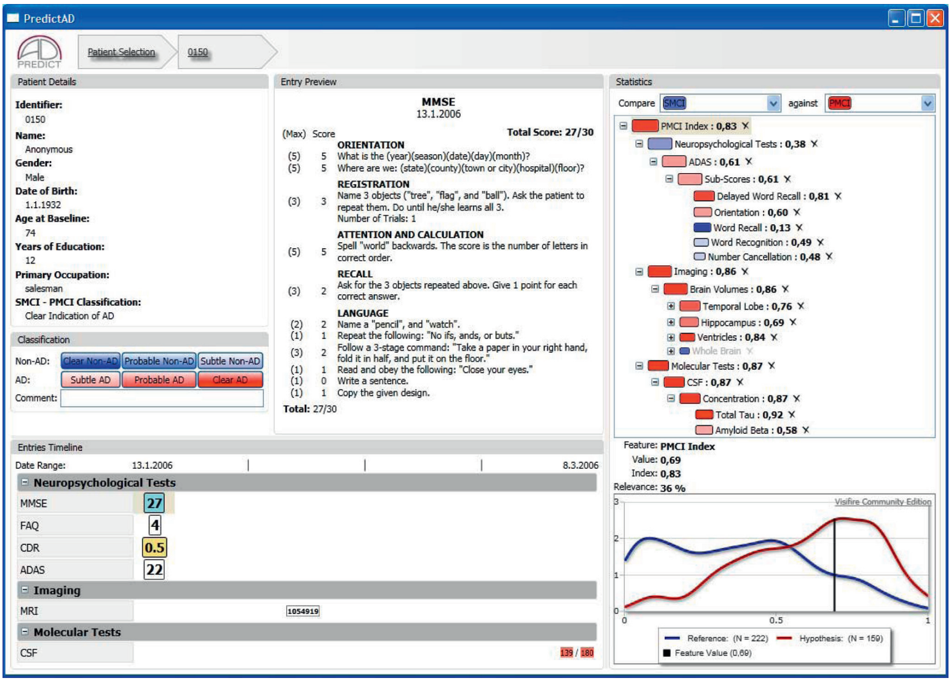
Statistical Analysis

Categorizations made by the clinical raters using baseline data alone were compared to clinical diagnoses made by the ADNI investigators after 3 years of follow-up. At each phase, assigning a patient to category (1), (2) or (3) was deemed correct if the clinical diagnosis was S-MCI and similarly, categories (4), (5) and (6) were correct if the clinical diagnosis was P-MCI.

Classification accuracy for the raters was computed from all categorizations made by the 3 clinical raters grouped together. Classification accuracy was defined as the ratio of correctly classified patients versus the total number of patients.

Differences between the diagnostic groups at baseline were evaluated with the unpaired Student t test. Changes in classification accuracy between the tool and the paper charts were evaluated using the McNemar test.

A subanalysis of cases categorized by the raters as (1) or (6) was performed separately. Differences in the relative number of *clear*



Patient Details

Identifier: 0150
Name: Anonymous
Gender: Male
Date of Birth: 1.1.1932
Age at Baseline: 74
Years of Education: 12
Primary Occupation: salesman
SMCI - PMCI Classification: Clear Indication of AD

Classification

Non-AD: Clear Non-AD | Probable Non-AD | Subtle Non-AD
AD: Subtle AD | Probable AD | Clear AD
Comment:

Entry Preview

MMSE
13.1.2006

(Max)	Score	**Total Score: 27/30**
		ORIENTATION
(5)	5	What is the (year)(season)(date)(day)(month)?
(5)	5	Where are we: (state)(county)(town or city)(hospital)(floor)?
		REGISTRATION
(3)	3	Name 3 objects ("tree", "flag", and "ball"). Ask the patient to repeat them. Do until he/she learns all 3. Number of Trials: 1
		ATTENTION AND CALCULATION
(5)	5	Spell "world" backwards. The score is the number of letters in correct order.
		RECALL
(3)	2	Ask for the 3 objects repeated above. Give 1 point for each correct answer.
		LANGUAGE
(2)	2	Name a "pencil", and "watch".
(1)	1	Repeat the following: "No ifs, ands, or buts."
(3)	2	Follow a 3-stage command: "Take a paper in your right hand, fold it in half, and put it on the floor."
(1)	1	Read and obey the following: "Close your eyes."
(1)	0	Write a sentence.
(1)	1	Copy the given design.

Total: 27/30

Statistics

Compare SMCI against PMCI

- PMCI Index : **0,83** X
  - Neuropsychological Tests : **0,38** X
    - ADAS : **0,61** X
      - Sub-Scores : **0,61** X
        - Delayed Word Recall : **0,81** X
        - Orientation : **0,60** X
        - Word Recall : **0,13** X
        - Word Recognition : **0,49** X
        - Number Cancellation : **0,48** X
  - Imaging : **0,86** X
    - Brain Volumes : **0,86** X
      - Temporal Lobe : **0,76** X
      - Hippocampus : **0,69** X
      - Ventricles : **0,84** X
      - Whole Brain X
  - Molecular Tests : **0,87** X
    - CSF : **0,87** X
      - Concentration : **0,87** X
        - Total Tau : **0,92** X
        - Amyloid Beta : **0,58** X

Feature: **PMCI Index**
Value: **0,69**
Index: **0,83**
Relevance: **36 %**

Entries Timeline

Date Range: 13.1.2006 – 8.3.2006

Neuropsychological Tests	
MMSE	27
FAQ	4
CDR	0.5
ADAS	22
Imaging	
MRI	1054919
Molecular Tests	
CSF	139 / 180

Fig. 1. Screenshot of the PredictAD tool used in this study. A patient being categorized at phase 4 has been marked as having a *clear* indication of early AD. Shown in the software are patient details, a timeline of entries where an MMSE entry is currently selected, a preview of the selected entry and a DSI visualization tree revealing how patient data relates to known S-MCI and P-MCI cases. Large nodes in the tree indicate good patient measures at differentiating between S-MCI and P-MCI. Shades of blue and red indicate where patient data are similar to S-MCIs and P-MCIs, respectively. Distributions (S-MCIs in blue, P-MCIs in red) and the patient value (in black) of the currently selected measure are displayed in a graph below the tree.

cases and in classification accuracy between phases were analyzed using the Fisher exact test.

Interrater agreement was evaluated using categorizations obtained at the 2 final phases where all data were available. This allowed us to test the agreement between clinical raters using the PredictAD software and paper charts. The test was performed by computing quadratic-weighted Cohen's kappa values.

In this study, a difference was considered statistically significant if $p < 0.05$.

Results

In this study, 140 patients with MCI at baseline were evaluated by 3 clinical raters. There were no differences between S-MCI and P-MCI regarding gender or age, but the S-MCI patients had a statistically significant higher MMSE ($p = 0.011$) and more years of education ($p = 0.044$) than the P-MCI patients (see table 1).

In total, the stepwise classification process resulted in 2,100 patient categorizations (140 patients × 3 clinical raters × 5 phases).

Increasing Accuracy and Confidence from the Availability of More Data

There was a trend towards more accurate classifications at each successive phase when the raters were presented with more clinical data, as depicted in figure 2. The increase in accuracy from phase 1 to phase 4, i.e. 62.6–70.0%, was statistically significant, $p = 0.029$.

At each successive phase, the raters gained more confidence in making the classifications. The number of patients assigned to the *Clear* categories increased by a sta-

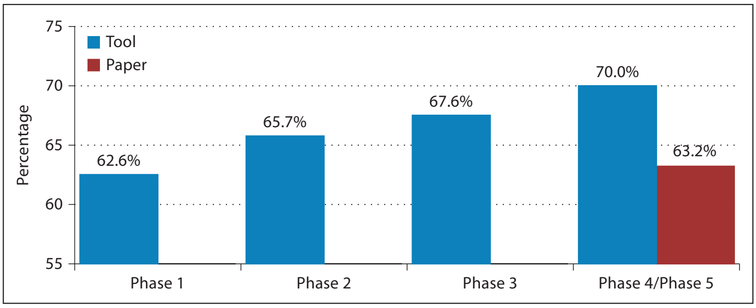


Fig. 2. Classification accuracy of clinical raters achieved for all patients during different phases of patient data presentation.

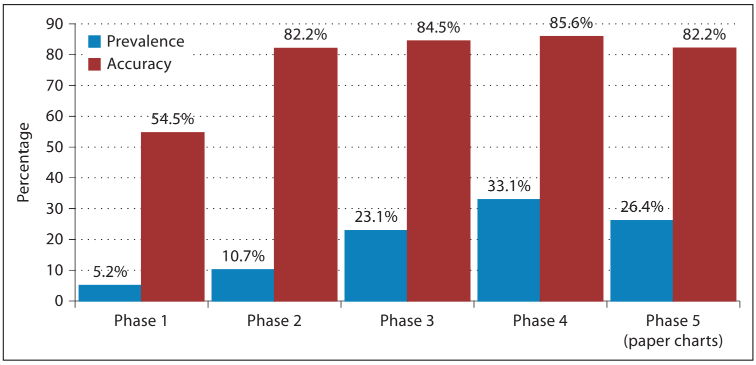


Fig. 3. Prevalence of patients assigned as *clear* (1) or (6) and classification accuracy for these cases at each phase.

tistically significant margin during each successive phase when information was added (phase 1–phase 2, p = 0.005; phase 2–phase 3, p = 0.000002; phase 3–phase 4, p = 0.002). Furthermore, the classification accuracy for the *Clear* patients increased (see fig. 3) with additional data, up to 85.6%. The increase in classification accuracy for cases categorized as *Clear* was statistically significant from phase 1 to phases 2, (p = 0.022), 3 (p = 0.004) and 4 (p = 0.002).

Using the Tool was Superior to Paper Charts

When the clinical raters were presented with patients in with exactly the same data in a paper chart format after they had been using the tool, there was a decrease in every rater's classification performance. There was a statistically significant decrease in classification accuracy from 70.0 to 63.2% (p = 0.005), as seen in figure 2.

In addition to decreased classification accuracy, the clinical raters were less confident in classifying patients as *Clear* cases when they only had the data in a paper chart format (see fig. 3). The decrease in the amount of *Clear* cases between having the tool available and not having it was close to being statistically significant (p = 0.050). Classification accuracy of *Clear* cases was also lower with the paper charts than with the tool, but the difference was not statistically significant.

Inter-Rater Agreement

When the 3 raters were using the tool, inter-rater agreement between them was very good with Cohen's kappa of 0.64, 0.76 and 0.80. When deprived of the tool, the agreement between raters was moderate (Cohen's kappa: 0.41, 0.43 and 0.71). Agreement between classifications made by a single rater using either the tool or paper charts was relatively good (Cohen's kappa: 0.58, 0.70 and 0.77).

Discussion

The results show that the best classification accuracy by clinical raters and the best agreement between raters was achieved when they used the software tool for decision support. That is to say, when clinical raters combined their clinical experience in AD with the additional information and context provided by the tool, they achieved the most accurate and consistent results.

There was a statistically significant decrease in the classification accuracy of the raters when they only had traditional paper charts with patient data. This suggests it is more challenging to apply clinical diagnostic criteria to a large amount of heterogeneous data when they are presented without any help that highlights important details. These results reinforce the case for decision-support systems that help clinicians manage large quantities of patient data obtained in modern healthcare.

Another aspect revealed by the results was that increasing amounts of data proved beneficial for the diagnostics in more than one way. It was expected that the availability of more data would improve classification accuracy, but there was also an added benefit of boosting confidence in the diagnosis. There was a statistically significant increase in the number of *clear* cases each time more data were made available, up to 33.1% of the patients being marked *clear* at the 4th phase. Classification accuracy of the *clear* cases also improved at each consecutive phase, up to 85.6%. In other words, one third of the patients were classified from baseline data at a relatively high accuracy by the clinical raters. Although the overall prediction accuracy with the tool was only 70%, having one third of the patients categorized at 85% accuracy (similar to the accuracy of clinical AD diagnoses) on average 20 months before the clinical AD diagnoses were given, could allow earlier treatments or better patient selection for drug trials. This result also suggests that there are some cases where the data contain strong evidence of early AD, perhaps allowing earlier diagnosis if interpreted correctly. In this regard, having the tool seemed to improve classification performance, as with the paper charts, the doctors categorized only 26.0% of the cases as *clear* and achieved 82.2% classification accuracy for them.

To put this study into context, it is important to stress that the 3 clinical raters had not met the patients in person and they were not able to review the patients' medical history. By interviewing and examining patients in person, valuable information regarding medical history, psychiatric symptoms and information about the onset of cognitive decline would be obtained that could further improve the diagnostic classification of the patients. The tool used here should be developed further to take other clinical parameters into account such as the ones mentioned above.

The clinical raters in our study only had the baseline data presented to them by the software or on the paper charts. Whether the software solution would add clinically meaningful value to the classification accuracy above that obtained by interview and examination of patients in person together with traditional presentation of investigational results is not clear from this study. This hypothesis should be examined in further prospective clinical studies.

A weakness of this study is that some of the stable MCI patients may have converted to AD after the 3-year follow-up period. These patients were classified as S-MCI in the data, but may in fact have been P-MCIs. This may skew the results towards a lower classification accuracy, affecting all the methods and analyses applied in the study.

Furthermore, the ADNI MCI population was a selected group of patients as opposed to a general mixed memory-clinic population. The performance of the PredictAD tool in a mixed group of patients with memory impairment has still to be clarified.

In conclusion, the results suggest that a computerized decision-support tool designed to help the reading and interpreting of heterogeneous patient data may be useful for diagnostic work. When raters were using the tool, the confidence in making a diagnostic classification, the accuracy of the diagnoses and interrater agreement were all significantly higher than their performance when only traditional paper charts were available.

Acknowledgements

This study was sponsored by the 7th Framework Programme of the European Commission EU-Grant- 224328-Predict-AD and the Simon Spies (G.W.) and Absalon foundations (G.W. and A.H.S.).

Data collection and sharing for this project was funded by the ADNI (National Institutes of Health grant U01 AG024904).

Disclosure of Conflict of Interest

J.M. has a patent application PCT/FI2010/050545 pending 2010. M.v.G. is an associate editor of the journal: Computer Methods and Programs in Biomedicine. J.L. has a patent (Disease State Index and fingerprint technology) patent US7,840,510B2 issued 2010 and a patent application PCT/FI2010/050545 pending 2010. G.W. is a board member of the Lundbeck Foundation and serves as a speaker/consultant for Pfizer, Janssen and Lundbeck. The remaining authors have no disclosures.

References

1 Petersen RC, Smith GE, Waring SC, Ivnik RJ, Tangalos EG, Kokmen E: Mild cognitive impairment: clinical characterization and outcome. Arch Neurol 1999;56:303–308.

2 Palmer K, Fratiglioni L, Winblad B: What is mild cognitive impairment? Variations in definitions and evolution of nondemented persons with cognitive impairment. Acta Neurol Scand 2003;197:14–20.

3 Petersen RC, Negash S: Mild cognitive impairment: an overview. CNS Spectr 2008;13: 45–53.

4 Landau SM, Harvey D, Madison CM, Reiman EM, Foster NL, Aisen PS, Petersen RC, Shaw LM, Trojanowski JQ, Jack CR Jr, Weiner MW, Jagust WJ: Comparing predictors of conversion and decline in mild cognitive impairment. Neurology 2010;75:230–238.

5 Madureira S, Verdelho A, Moleiro C, Ferro JM, Erkinjuntti T, Jokinen H, Pantoni L, Fazekas F, Van der Flier W, Visser M, Waldemar G, Wallin A, Hennerici M, Inzitari D: Neuropsychological predictors of dementia in a three-year follow-up period: data from the LADIS study. Dement Geriatr Cogn Disord 2010;29:325–334.

6 Julkunen V, Niskanen E, Muehlboeck S, Pihlajamaki M, Kononen M, Hallikainen M, Kivipelto M, Tervo S, Vanninen R, Evans A, Soininen H: Cortical thickness analysis to detect progressive mild cognitive impairment: a reference to Alzheimer's disease. Dement Geriatr Cogn Disord 2009;28:404–412.

7 Brandt C, Bahl JC, Heegaard NH, Waldemar G, Johannsen P: Usability of cerebrospinal fluid biomarkers in a tertiary memory clinic. Dement Geriatr Cogn Disord 2008;25: 553–558.

8 Koivunen J, Pirttila T, Kemppainen N, Aalto S, Herukka SK, Jauhianen AM, Hanninen T, Hallikainen M, Nagren K, Rinne JO, Soininen H: PET amyloid ligand [11C]PIB uptake and cerebrospinal fluid beta-amyloid in mild cognitive impairment. Dement Geriatr Cogn Disord 2008;26:378–383.

9 Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM: Forecasting the global burden of Alzheimer's disease. Alzheimers Dement 2007;3:186–191.

10 Cummings JL: Treatment of Alzheimer's disease: the role of symptomatic agents in an era of disease-modifying therapies. Rev Neurol Dis 2007;4:57–62.

11 Mattila J, Koikkalainen J, Virkki A, Simonsen A, van Gils M, Waldemar G, Soininen H, Lotjonen J: A disease state fingerprint for evaluation of Alzheimer's disease. J Alzheimers Dis 2011;27:163–176.

12 McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM: Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology 1984;34:939–944.

13 Chang C, Lin C: LIBSVM: a library for Support Vector Machines. Science 2001;2:1–39.

14 Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, van der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A, Makris N, Rosen B, Dale AM: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–355.

Series title and number
VTT Science 51

Title	**Disease State Index and Disease State Fingerprint** **Supervised learning applied to clinical decision support in Alzheimer’s disease**
Author(s)	Jussi Mattila
Abstract	Due to scientific and technological advancements, investigations in modern medicine are producing more measurement data than ever before. Since a large amount of information exists, and it is also being produced at ever-increasing rates, no single person can digest all current knowledge of diseases. Data collected from large patient cohorts may contain valuable knowledge of diseases, which could be useful to clinicians when making diagnoses or choosing treatments. Making use of the large volumes of data in clinical decision-making requires ancillary help from information technologies, but such systems have not yet become widely available. This thesis addresses the challenge by proposing a computer-based decision support method that is suited to clinical use. This thesis presents the Disease State Index (DSI), a supervised machine learning method intended for the analysis of patient data. The DSI comprehensively compares patient data with previously diagnosed cases with or without a disease. Based on this comparison, the method provides an estimate of the state of disease progression in the patient. Interpreting the DSI is made possible by its visual counterpart, the Disease State Fingerprint (DSF), which allows domain experts to gain a comprehensive view of patient data and the state of the disease at a quick glance. In the design and development of these methods, both performance and applicability in clinical use were taken into account equally. Alzheimer’s disease (AD) is a slowly progressing neurodegenerative disease and one of the largest social and economic burdens in the world today, and it will continue to be so in the future. Studies with large patient cohorts have significantly improved our knowledge of AD during the last decade. This information should be made extensively available at memory clinics to maximize the benefits for diagnostics and treatment of the disease. The DSI and DSF methods proposed in this thesis were studied in the early diagnosis of AD and as a measure of disease progression in six original publications. The methods themselves and their implementation within a clinical decision support system, the PredictAD tool, were quantitatively evaluated with regard to their performance and potential benefits in clinical use. The results show that the methods and clinical decision support tool based on these methods can be used to follow disease progression objectively and provide earlier diagnoses of AD. These, in turn, could improve treatment efficacy due to earlier interventions and make drug trials more efficient by allowing better patient selection.
ISBN, ISSN	ISBN 978-951-38-8119-1 (Soft back ed.) ISBN 978-951-38-8120-7 (URL: http://www.vtt.fi/publications/index.jsp) ISSN-L 2242-119X ISSN 2242-119X (Print) ISSN 2242-1203 (Online)
Date	December 2013
Language	English, Finnish abstract
Pages	96 p. + app. 75 p.
Name of the project	
Commissioned by	
Keywords	Supervised learning, data visualization, clinical decision support systems
Publisher	VTT Technical Research Centre of Finland P.O. Box 1000, FI-02044 VTT, Finland, Tel. +358 20 722 111

Julkaisun sarja ja numero
VTT Science 51

Nimeke	**Taudin tilan indeksi ja taudin tilan sormenjälki** **Ohjatun koneoppimisen menetelmä kliinisen päätöksenteon tueksi Alzheimerin taudissa**
Tekijä(t)	Jussi Mattila
Tiivistelmä	Nykyaikaisen lääketieteen tutkimuksissa kerätään uuden teknologian ansiosta enemmän mittaustuloksia kuin koskaan aiemmin. Koska tietoa on paljon ja sitä tuotetaan yhä nopeammin, yksittäisen ihmisen on mahdotonta sisäistää kaikki olemassa oleva ajantasainen tietämys eri taudeista. Suurista potilasjoukoista saadut tulokset saattavat sisältää arvokastakin tietoa, josta olisi apua kliinikoille diagnostiikassa ja hoitotoimenpiteitä päätettäessä. Suurten tietomassojen hyödyntäminen päätöksenteossa vaatii tietotekniikkaa apuvälineeksi, mutta tähän mennessä tehtävään sopivia järjestelmiä ei ole saatu laajamittaiseen käyttöön. Tämä väitöskirja vastaa tähän haasteeseen esittelemällä kliiniseen käyttöön soveltuvan tietokonepohjaisen päätöksenteon tukijärjestelmän. Tämä väitöskirja esittelee ohjatun koneoppimisen menetelmän nimeltään Disease State Index (DSI, suom. taudin tilan indeksi), jolla potilaiden mittaustuloksia voidaan verrata kattavasti suurissa tietokannoissa oleviin aiemmin diagnosoituihin potilaisiin. Menetelmä antaa vertailun perusteella arvion potilaan taudin tilasta ja sen etenemisestä. DSI:n tulosten tulkintaan kehitettiin visualisointimenetelmä nimeltään Disease State Fingerprint (DSF, suom. taudin tilan sormenjälki), joka mahdollistaa potilaan tietojen ja tulosten nopean mutta kattavan arvioinnin. Menetelmien suunnittelussa ja toteutuksessa otettiin yhtä lailla huomioon tarkkuusvaatimukset kuin niiden soveltuvuus käyttöönottoon klinikoissa. Alzheimerin tauti (AT) on hitaasti etenevä neurodegeneratiivinen tauti ja yksi maailman vakavista sosiaalisista ja taloudellisista ongelmista nyt ja tulevaisuudessa. Potilaista kerättyjen suurten tietomassojen avulla AT:n kuva on terävöitynyt merkittävästi kymmenen viime vuoden aikana. Tämä tieto olisi hyvä saada laajamittaisesti muistiklinikoiden käyttöön parhaan mahdollisen diagnostiikan ja hoidon varmistamiseksi. Väitöskirjassa esiteltyjen menetelmien soveltuvuutta AT:n varhaiseen diagnostiikkaan sekä taudin seurantaan tutkittiin kuudessa julkaisussa, joissa itse menetelmät sekä niiden toteutus kliinisenä päätöksenteon tukijärjestelmänä, nimeltään PredictAD tool (suom. EnnustaAT-apuväline), arvioitiin kvantitatiivisesti suorituskyvyn ja potentiaalisten hyötyjen suhteen. Tulokset näyttävät, että menetelmillä ja niiden pohjalta kehitetyllä kliinisen päätöksenteon tukityökalulla voidaan seurata potilaan taudin tilan etenemistä objektiivisesti sekä mahdollistaa AT:n varhaisempaa diagnostiikkaa. Näiden voidaan puolestaan odottaa parantavan hoitojen tehoa hoitojen aiemman aloituksen ansiosta sekä auttavan lääkekehityksessä paremmin kohdennetun potilasvalinnan myötä.
ISBN, ISSN	ISBN 978-951-38-8119-1 (nid.) ISBN 978-951-38-8120-7 (URL: http://www.vtt.fi/publications/index.jsp) ISSN-L 2242-119X ISSN 2242-119X (painettu) ISSN 2242-1203 (verkkojulkaisu)
Julkaisuaika	Joulukuu 2013
Kieli	Englanti, suomenkielinen tiivistelmä
Sivumäärä	96 s. + liitt. 75 s.
Projektin nimi	
Toimeksiantajat	
Avainsanat	Supervised learning, data visualization, clinical decision support systems
Julkaisija	VTT PL 1000, 02044 VTT, puh. 020 722 111

Disease State Index and Disease State Fingerprint

Supervised learning applied to clinical decision support in Alzheimer's disease

Due to scientific and technological advancements, investigations in modern medicine are producing more measurement data than ever before. Since a large amount of information exists, and it is also being produced at ever-increasing rates, no single person can digest all current knowledge of diseases. Data collected from large patient cohorts may contain valuable knowledge of diseases, which could be useful to clinicians when making diagnoses or choosing treatments. Making use of the large volumes of data in clinical decision-making requires ancillary help from information technologies, but such systems have not yet become widely available. This thesis addresses the challenge by proposing a computer-based decision support method that is suited to clinical use.

ISBN 978-951-38-8119-1 (Soft back ed.)
ISBN 978-951-38-8120-7 (URL: http://www.vtt.fi/publications/index.jsp)
ISSN-L 2242-119X
ISSN 2242-119X (Print)
ISSN 2242-1203 (Online)

VTT SCIENCE 51 Disease State Index and Disease State Fingerprint