MASTER THESIS

15th July 2019

Evidence Requirements of Deep Learning Based Medical Support Systems for Mental Disorders

Dario Müller

S1443623

Behavioural, Management and Socials sciences (BMS)

Intelligent Transport Systems and Human Factors

EXAMINATION COMMITTEE

First Supervisor:	Dr. Simone Borsci
Second Supervisor:	Dr. Martin Schmettow
External Supervisor:	Dr. Marjan Hummel

Table of Contents

1 Summary

Overall goal: The following works copes with the potential of Deep Learning based Medical Decision Support Systems (DLMSS) for the use in the mental healthcare setting. It explores potential requirements for the use of Deep Learning based technology as decision support systems in the mental healthcare sector. The work provides a first approach towards the alignment of the technology into the mental health sector and may inform experts and designers to better understand the needs of the sector.

Methods: The thesis explores aspects that are on the edge of innovation with very limited information available. Due to that, qualitative methods that resembled a grounded theory approach were chosen and were divided into a theoretical literature research, qualitative interviews and an online questionnaire.

Procedure: Existing literature on the topic was reviewed and tailored towards potential requirements for Deep Learning technology. Informed by this literature research, interviews with experts from the field of psychotherapy, neuropsychology and radiology (n=8) were conducted, the information was matched with the literature research and a list of potential evidence requirements for the technology was created. The process was guided with the feedback of four experts from the field of Human Factors and Health Care Innovation from the University of Twente and the Philips Research Institute in Eindhoven.

Findings: The results indicate that the synthesis of information that cannot be produced through the current means is the most important driver for such technology and that this information must directly connect to the therapeutic process. The central aspect here is the creation of a benefit for the patient. It could also be shown that a certain number of regulatory barriers must be overcome, and that regulatory approval and adoption of current medical guidelines is needed to implement the technology into practice. For the individual appraisal of the technology, trust and the concept of convergent validation emerged as important factors for the use of the technology as well as ethical considerations that must be made in the context of its use.

1.1 Acknowledgment

The study is done as part of a master thesis at the University of Twente and in collaboration with NIHR London IVD and Philips Healthcare. Thanks is given to the Experts and Supervisors who guided the process and delivered constant feedback and evaluation far beyond the scope of what can be expected. Special Credit is given to the work of Huddy, Misra, Mavroveli, Barlow & Hanna, (2018) which was the foundation for the questionnaire and an important source of information for the thesis.

Huddy, J. R., Ni, M., Misra, S., Mavroveli, S., Barlow, J., & Hanna, G. B. (2018). Development of the Point-of-Care Key Evidence Tool (POCKET): a checklist for multi-dimensional evidence generation in point-of-care tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*.

2 Introduction

Our modern society is characterized by a constant increase of technological diffusion into nearly all parts of life. One of the sectors in which technology is playing a crucial role is the field of medicine. Currently, the overall expenditure for the health care system seems to rise rather than to fall, not only in terms of total money spent, but also in terms of % of GDP (WHO, 2018; Cuckler et al., 2018). While the diffusion of new technology into the medical sector has the potential to increase the effectivity of treatments and the well-being of patients, it is also associated with an increase in costs (Nichols, 2002). The topic of increased costs however is somewhat controversial and there are multiple factors that influence the costs calculations of a technology like the context and frequency of use, long-term vs short term outcomes and many more (Goyen & Debatin, 2009). Even though financial aspects often produce aversive feelings in the context of the healthcare setting, the resources that can be invested into healthcare are sparse and the reduction of overall expenditure is necessary to sustain the quality of the system (WHO, 2010). A potential way to avoid the increased costs that are associated especially with new technologies is the utilization and expansion of the use of already existing technologies like MRI (Goyen & Debatin, 2009).

Another important reason for the increased costs of our healthcare system is a general shift in the age of the European society in that it grows older, leading to a higher old-age dependency ratio (European Commission and the Economic Policy Committee, 2015; Goyen & Debatin, 2009; Randall, 2017). The old-age dependency ratio is the ratio of people aged 65 and above as compared to the ratio of people aged 15-64 (European Commission and the Economic Policy Committee, 2015). Effectively, it describes the ratio of people who are dependent on the economic power of the rest of the population due to high age (Muszyńska & Rau, 2012). It is expected that the ratio of people aged 65 and above will grow from 18.9% in 2015 up to 23.9% in 2030 and 29% in 2060 (Eurostat, 2019). Surviving improvements due to higher quality of life and better provision of healthcare, are an important driver for this shift (Muszyńska & Rau, 2012). This creates health related and social problems because people above the age of 65 are more likely to exit employment and thereby do not produce tax money anymore (Muszyńska & Rau, 2012). This lack of resources needs to be compensated in all areas of which the health care sector is a very important one. One approach to compensate for this lack of resources is an increase in efficiency and in the quality of the health-care system to increase general population health. Poor Health is shown to be the most important predictor of leaving paid employment (Muszyńska & Rau, 2012) and therefore increasing the health of the population might also prolong the duration of payment and thereby influence the old-age dependency ratio by changing its mean age. This is where the implementation of new technology into the setting can create a

potential benefit. It has the chance to make processes more efficient and to achieve high-quality healthcare under difficult economic circumstances.

2.1 Deep Learning-Based Medical Support Systems (DLMSS)

One medium of new technology that is hoped to help in sustaining and increasing the quality of our health-care system is the integration of deep learning technology into the medical sector. Deep Learning technology falls into the category of machine learning tools which indeed falls under the umbrella term of artificial intelligence (Pesapane, Codari & Sardanelli, 2018). It is aimed to aid in different forms of clinical decision-making like diagnosis or treatment selection. For the scope of this work, the technology will be labelled Deep Learning-Based Medical Support Systems (DLMSS). In its essence, the technology is an advanced algorithm that possesses 'machine learning' abilities, i.e. it can learn from patterns in data without specifically tailored programming (Zaharchuk, Gong, Wintermark, Rubin & Langlotz, 2018). Different forms of the technology exist and because of the constant progress in this field, it can be hard to create specific definitions of these technologies. A distinction that can be made is between supervised and unsupervised machine learning. With supervised learning, some truth exists that the technology can use to learn, for example having a certain disorder (Zaharchuk et al., 2018). For unsupervised learning, there is no criterion the algorithm can use, and the classification of the algorithm is based solely on the data (Zaharchuk et al., 2018). While it is important to differentiate different forms learning in the context of the technology, other categories like semi-supervised learning also exist.

Deep learning technology falls under the category of supervised learning. The technology consists out of an input layer which receives information, hidden layers which analyze the data, and an output layer which presents the results (Pesapane, Codari & Sardanelli, 2018). The hidden layers are the parts of the technology that processes the information by assigning and changing the weight or importance of the input for a certain output or class (Pesapane, Codari & Sardanelli, 2018). In this way, the DLMSS have the potential to learn the correlates of a specific illness and to use this acquired knowledge as a sort of imaging biomarker for this disorder. Biomarkers are 'objective biological measures that can predict clinical outcomes' (Abi-Dargham & Horga, 2016, p.1248). The form of automatic feature extraction that the DLMSS possess is a big advantage because it reduces a potential bias on which features to choose for the biomarker since the technology identifies them on its own (Zaharchuk et al., 2018). On the other hand, it also imposes certain problems for the validity of these biomarkers. One of the fields in which there is special interest is the field of radiology (McBee et al., 2018). In comparison to 'standard' statistical models, deep learning algorithms exceed in their performance especially when an increasing amount of information needs to be analyzed (Pesapane, Codari & Sardanelly, 2018). Because radiology has evolved into a data-rich and very complex

domain, the field shows perfect conditions for deep-learning based medical support systems (DLMSS) (McBee et al., 2018). Additionally, the approach allows to utilize already existing technology which could potentially ease the associated barrier of increased costs (Goyen & Debatin, 2009; Nichols, 2002).

2.2 Human Decision Making and Radiology – Radiomics

As an umbrella term, radiology describes the use of medical imaging in order to diagnose and treat diseases in the body. Different forms of medical imaging exist and what kind of imaging is used depends on the illness and specific guidelines like the ACR Appropriateness Criteria exist for this topic (American College of Radiology, 2019). The main function of radiology is to provide additional information in order to reduce diagnostic uncertainty (Manning, Gale & Krupinski, 2005). Because of the nature of radiology, the interpretation of the data is highly qualitative and relies on the perception and the decision making of the radiologist (Manning, Gale & Krupinski, 2005). As we know today, decision making is not always rational as defined by comparing different ways of acting and choosing the one with the highest subjective profit (Kahneman & Egan, 2011). Even though this kind of decision making could often result in a 'better' outcome, we simply do not possess the cognitive abilities to do so. Because of that, we tend to rely on heuristics, especially when we need to make fast decisions under ambitious parameters (Kahneman, Slovic & Tversky, 1982). Kahneman referred to these different methods of decision making as system 1 and 2, system one being the fast and intuitive system while system two is the slower and analytical one. This also appears congruent with research from medical decision making. It was observed that clinicians possess a fast, similarity-based reasoning process and a more knowledge based one (Norman, 2005). Research about medical image analysis shows the same results, indicating a rapid, non-selective pathway that can extract global information and a slower, selective pathway for individual identification of targets (Drew, Evans, Võ, Jacobson & Wolfe 2013; Wolfe, Võ, Evans & Greene 2011). What is striking is that through the non-selective pathways, radiologists were able to detect 70% of lesions in a 200-msec time frame in chest radiographs, giving additional evidence for the existence and the use of this (Drew et al., 2013). Drew et al (2013), propose that the non-selective pathway gives an overview of the picture and provides a structure that can then be used to identify and to guide the visual identification of specific objects (e.g.: lesions) through the selective pathway (Drew et al., 2013). This process however is somewhat limited by our visual perception. Signal detection theory can help to explain this. In order to detect an 'odd' stimulus, some form of visual differentiation must be apparent between it and the 'normal' one (Prins, 2016). The smaller the visual difference, the harder it is to detect the odd stimuli against the 'background noise' (Prins, 2016). Other factors playing a role in the visual detection are feature binding and object recognition (Wolfe, Võ, Evans & Greene 2011). What is important is that our visual perception creates somewhat of a bottleneck for the detection of radiologic anomalies (Wolfe,

Võ, Evans & Greene 2011). This is where the implementation of the DLMSS might bring an advantage. It has the potential to read the data that constitutes the image and to find differences in it that are too subtle to be detected by human radiologists. Since the bottleneck in this regard appears to focus on the visual apparatus of the human being, the technology might present the information in a way that eliminates this bottleneck while also utilizing the higher cognitive aspects of human decision making.

2.3 Deep Learning for Mental Disorders - Psychoradiology

One of the possible areas of interest is the utilization of radiology imaging techniques for the field of mental disorders. Due to the prevalence of major psychiatric disorders and the associated costs, there is also an incentive to enhance the effectivity in the treatment of these illnesses (Lui, Zhou, Sweeney & Gong, 2016; Olesen et al., 2012). Due to these factors and the described capabilities of deep learning tools, the utilization of radiological techniques is hoped to aid in the treatment of mental disorders and appears to be a promising branch of research (Arbabshirani, Plis, Sui & Calhoun, 2017; Lui, Zhou, Sweeney & Gong, 2016). Recently, the term 'Psychoradiology' is used to describe this intersection of psychiatry and radiology (Lui, Zhou, Sweeney & Gong, 2016). However, while the use of radiography to aid in the diagnosis of neurogenerative illnesses like Alzheimer's disease is already implemented in practice, the use of radiography for psychiatric disorders like depression is very low at best (Lui, Zhou, Sweeney & Gong, 2016). One of the reasons for this is that the neuronal correlates of mental disorders or other cerebral deficits are often functional and highly heterogenous (Lui, Zhou, Sweeney & Gong, 2016). Because of the nature of human decision making and of our visual perception, qualitative analysis of these differences is not enough to find the neuronal correlates that constitute mental illnesses (Lui, Zhou, Sweeney & Gong, 2016). This indeed asks for quantitative forms of image analysis (Lui, Zhou, Sweeney & Gong, 2016). In the field of medicine, and especially for the detection of cancer, the quantitative approach called 'radiomics' is already tested quite successfully (Gillies, Kinahan & Hricak, 2016).

In the context of Psychoradiology, deep learning could be the technology that enables the shift towards the use of radiology (Zaharchuk, Gong, Wintermark, Rubin & Langlotz, 2018). This is due to the huge amount of multimodal imaging information that is apparent in neuroradiology (Zaharchuk et al., 2018) and the limits of our perception. With these capabilities, deep learning can help to understand mental disorders in terms of their neurological correlates' and not of their perceptual symptoms (Lui, Zhou, Sweeney & Gong, 2016).

At the moment, Philips is working on a deep-learning technology that aims at this field and is called Neuro-AI. The technology aims to correlate and weight a certain input, in this context of brain imaging data, to a certain output or class, in this context a certain type of mental disorder. For

instance, Neuro-AI may be used in the future to diagnose a patient who has or who will develop a disorder or to support the decision about what type of treatment is best suited for the patient.

2.4 Innovation in the Medical Field

Even though the early results of Deep Learning in the context of psychoradiology appear very promising, the way from being a research tool towards being implemented into clinical practice is a complex process and affected by interacting ecological factors at the individual, the organizational and even the community level (Cresswell & Sheikh, 2013; Fixsen, Naoom, Blase & Friedman, 2005). Additionally, there is a steady increase in technological advancements in all fields of medical science and the number of potential technologies to aid clinical practice is huge, which indeed increases the number of potential competitors. At the same time, the medical field has very high standards for the evaluation of new technologies or forms of therapy and a substantial amount of evidence is needed to pass this evaluation (Pesapane, Volonté, Codari & Sardanelli, 2018; Huddy, Ni, Misra, Mavroveli, Barlow & Hanna, 2018). This leads to a constant divergence of what is achieved in theoretical science and what is done in practice, which is metaphorically labelled as the 'Valley of Death' (Beach, 2017). It can be observed that this gap between science and practice is relatively consistent despite the increase in resources that are invested into medical research (Roberts, Fischhoff, Sakowski & Feldman 2012). One of the reasons for this gap is the declining role of clinical scientists in the field of medical research (Roberts, Fischhoff, Sakowski & Feldman 2012). Especially for innovative technologies like deep learning, that rely on a multidisciplinary team, an incomplete understanding for the needs of the clinical sector, which is a highly dynamic and complex field, is a huge potential barrier for the implementation of the technology (Roberts, Fischhoff, Sakowski & Feldman 2012; Woods & Hollnagel, 2006, Borsci, Uchegbu, Buckle, Ni, Walne & Hanna, 2018). The detrimental effects of this gap go even further. If the success of technological implementation into the health-care setting is uncertain, the investment into new technology might not deliver a financial benefit which can limit the investments of companies like Philips into the sector and hinder the development of better medical treatments in general. Because of this circumstance, approaches to evaluate medical technology before it gets implemented receive more attention and increased importance of translational research can be observed (Borsci et al., 2018, Roberts, Fischhoff, Sakowski & Feldman 2012; Woolf, 2008).

2.5 Translational Research

Translational research is a broad field and exact definitions can differ. In general, the term describes research that is aimed on the effective translation of new scientific knowledge into approaches usable for clinical practice (Woolf, 2008). A very important factor in this regard is the creation of evidence

for the usefulness of a new device (Huddy, Ni, Misra, Mavroveli, Barlow & Hanna, 2018). Scientific evidence about the functionality of a new device is a crucial requirement for its implementation and can help to reduce potential barriers or accelerate potential facilitators (Huddy et al., 2018).

The creation of scientific evidence is also one of the major factors which clinicians deemed necessary before considering the use of DLMSS (Philips Premium Report, 2018). At the moment, the pathway for evidence generation is fragmented and non-linear for diagnostic devices (Huddy, Ni, Misra, Mavroveli, Barlow & Hanna, 2018). This is problematic because it makes the process of accumulating the appropriate evidence less efficient and prolongs the time until implementation in the setting, potentially reducing the chances of overall success. A more systematic approach to the generation of appropriate evidence has the potential to save costs and time in the evidence generations by focusing the resources that are invested on the evidence that is important (Huddy et al., 2018). In order to accumulate the necessary evidence, it is important to understand the needs of the clinical setting and to know what kind of evidence must be created for the specific scenario (Borsci et al., 2018).

2.6 Evidence Requirements and the POCKET

In order to find out what kind of evidence must be generated to implement a technology into a certain setting, the requirements of the agents working in this setting must be elicited. These are called requirements and resemble the values that the user hold in regard of a (specific) technology or software (Davis, 1998). One example of a systematic approach that is aimed to aid in the generation of appropriate evidence to fulfill the requirements of different stakeholders is the Point-Of-Care Evidence Tool (POCKET) (Huddy, Ni, Misra, Mavroveli, Barlow & Hanna, 2018*)*. It consists out of a list of evidence that is considered important by different clinical end-users when using Point-Of-Care devices. Point of care devices allow to test patients and receive results on the bedside. This has the advantage of being a more cost-efficient way of diagnosis but also brings with it potential barriers like reduced accuracy. The POCKET is a checklist that allows the evaluation of Point-Of-Care devices before they are implemented into practice. This evaluation is focused around evidence requirements to either indicate factors that facilitate the implementation of the technology (e.g.: evidence for decreased time to treatment) or decrease potential barriers for that (e.g.: Evidence that disconfirms decreased the accuracy of point-of-care devices). The list was made using qualitative stakeholder-feedback to create different types of evidence requirements for these devices. By using literature and semi-structured interviews, different themes relating to the creation of evidence emerged and were then translated into concrete evidence requirements. This was achieved with help of a Delphi questionnaire study and expert workshops with stakeholders involved in the process. The stakeholders were patients, presenters from the industry, clinicians, commissioners and regulators. The end product was the POCKET; a checklist comprised of 64 statements divided into seven main topics:

1. Technical Description of Test
2. Clinical Pathway
3. Stakeholders
4. Economic Evidence
5. Test Performance
6. Usability and Training
7. Clinical Trials

Every statement consisted out of a concrete evidence-requirement (E.g.: Diagnostic accuracy study) and referred to one of the seven topics (E.g.: clinical trials). On this way, a checklist was created that allows the systematic evaluation of Point-Of-Care devices before they are implemented into practice.

2.7 Research

The aim of the Master thesis is to explore potential factors that have the potential to influence the adoption of the technology and to investigate what evidence requirements should be delivered in this regard. It is designed as preliminary and translational research and aims to facilitate the implementation of the technology into the mental healthcare setting. Based on literature, a consolidated list of requirements for diagnostic devices was created and, by using expert feedback, adapted towards the use of Deep Learning technology as decision support systems in the mental healthcare sector. The product of the master thesis can be used as a guidance for early evaluation of the technology in the context of user-requirements and to decrease a potential mismatch between the technology and the clinical setting. The end-product is aimed to be a checklist for the evidence requirements of deep learning based medical decision support systems. The POCKET will be used as a starting point.

Research Objective:

Exploration of factors and associated evidence requirements that have the potential to influence the adoption of deep-learning based medical decision support systems (DLMSS) in the field of mental disorders.

3 Methods

3.1 Overview

A qualitative approach was chosen for the work. The reason for that is the novelty of the topic. The available literature of requirements for the use of deep learning classifiers as decision support systems for mental disorders is very sparse and theoretical interferences must be made to adopt literature from other fields towards the use of DLMSS for mental disorders. Additionally, the use of machine learning techniques and deep learning technology for mental disorders is limited to very few and highly experimental research settings like the NeuroMiner tool used by the European PRONIA project (https://www.pronia.eu/neurominer/). This makes it hard to obtain empiric data that is valid. Due to this, the focus of this research lies on the discovery and synthesis of new information for which qualitative research methods are most suitable. A grounded theory approach was adopted and took the form of different iterative research methods which will be explained below. The results were then matched and gradually synthesized into a coherent description of possible requirements for the technology. Eight experts from the field of psychology and radiology were interviewed and the research process was additionally guided with feedback from five experts from the field of Human Factors and Health Technology and Economics. The research can be divided into two different steps.

First, a preliminary systematic review was conducted to explore the information available on the topic. This was done to synthesize a pool of information which was used to inform the materials of the data collection process.

The second part of the study was the data collection which was also divided into two parts; qualitative interviews and an online questionnaire with professionals in the field of radiology, psychotherapy and neuropsychology. The qualitative interviews were designed to evaluate the themes found during the literature research and explore additional factors that might influence the implementation of the technology into the mental health sector. The questionnaire was designed to evaluate the specific evidence requirements that were synthesized from the POCKET.

Finally, the reviewed and integrated list was analyzed, and a preliminary version of the checklist was created: InSilicioEvidenceTool (ISET). As we will discussed is section 5, it was also attempted to analyses this new checklist with expert feedback. However, due to a lack of participants for the checklist, it was not possible to gather the specific evidence requirements. The list can be found in appendix A.

3.2 Preliminary Systematic Review

3.2.1 Review POCKET list.

As a starting point, the POCKET was reviewed and used to get a better overview about the field of evidence requirements. Together with four experts in the field of Human Factors, Health Economy and Health Technology, the list was reviewed and requirements that did not translate to the new technology were excluded. The adapted list was used as a template for the new list. Additionally, the help of experts at the University of Twente and Philips Research was provided in form of literature and feedback. This decision was made because the POCKET is a checklist that is substantially validated in the clinical context. Different parties were involved in its evaluation, including clinicians, methodologists, industry representatives and regulators. This leads to the creation of a multidimensional list of evidence requirements for medical Point of Care Devices. Despite this focus on invitro diagnostics, the higher-level themes appeared to be translatable towards medical devices in general. Additionally, a lack of literature on specific requirements for DLMSS made it hard to create specific inclusion criteria that are justified reasonable in this context. Because the POCKET was matched and adopted with literature and then ought to receive expert validation through means of interviews and an online questionnaire, the decision was made to only exclude items that were based specifically on the mobile nature of Point of Care devices which cannot be translated towards DLMSS. Eleven lower-level evidence requirements were excluded. These factors The Exclusion Criteria are presented in table 1.

[Table of excluded items]

Table 1.
Exclusion Criteria for POCKET evidence Requirements

- The evidence resolves around physical features of the device influencing mobility (E.g.: Weight, Seize, etc.)
- The evidence resolves around a direct comparison of laboratory and mobile test results.
- The evidence requirement is not deemed important by the clinician.

3.2.2 Literature Research

The literature research was conducted with the aim to achieve a better understanding of potential requirements, facilitators and barriers for A.I. based classifiers in the field of mental disorders. After the review, a narrative literature research was conducted. This was done in an iterative way. Based on the POCKET and in agreement with four healthcare and human-factors experts, different search topics were developed and systematically searched for on PubMed and Scopus. Additionally, Google Scholar was used to explore specific topics that came up during the literature search and to finalize the literature research. The results were discussed with the experts and used to create new search topics. This process was iterated until thematic saturation was achieved. In total, 233 articles were deemed acceptable for narrower inspection. The abstracts were read and 34 articles that were relevant for the task were identified. An overview of the literature research can be seen in figure 1.

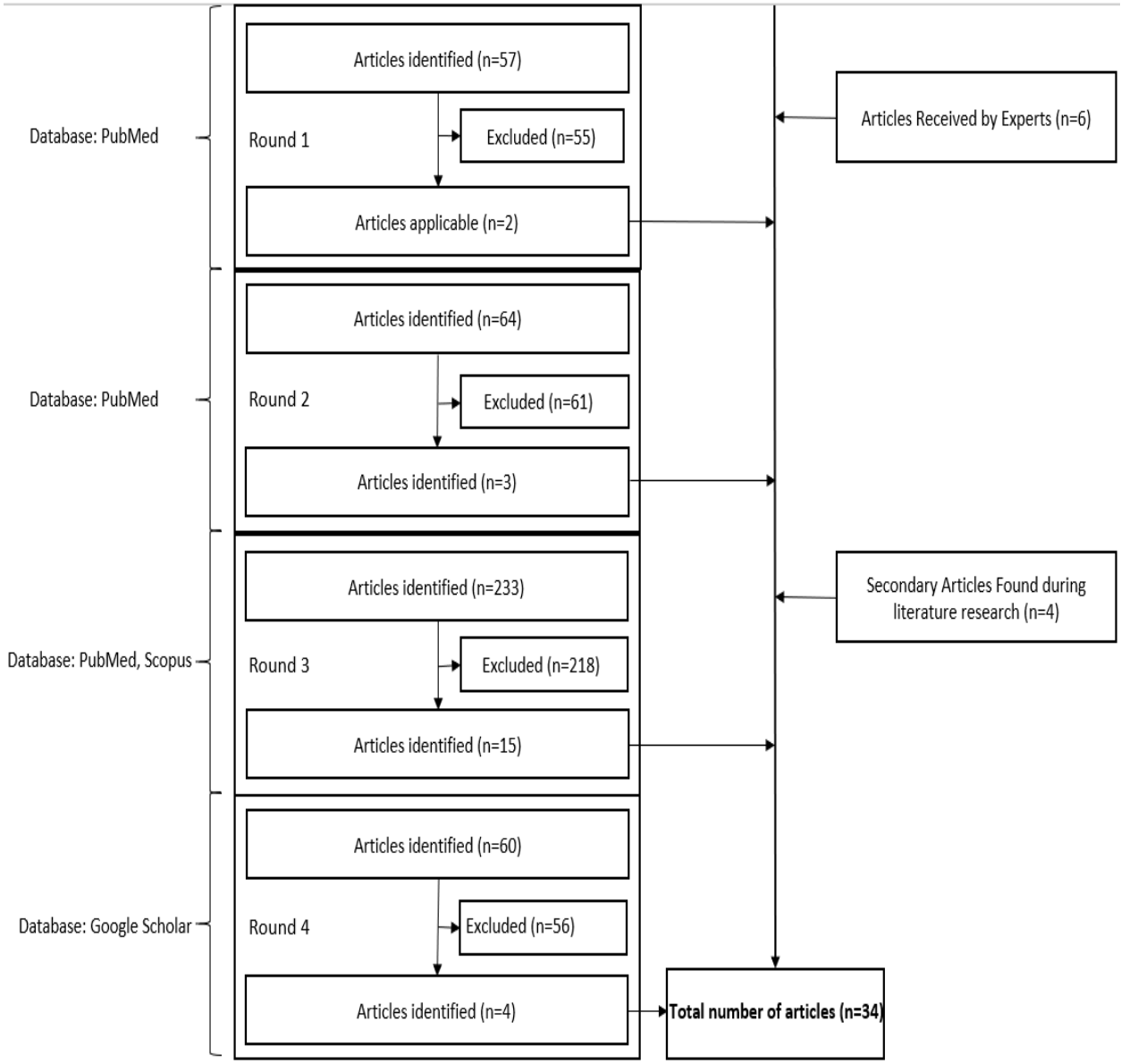

Figure 1. Overview of the literature Research

First Round

The first search was conducted on PubMed with the following search term: ((((Deep Learning) OR Artificial Intelligence) OR Machine Learning) AND Classifier) AND Requirements. In order to fulfil the inclusion criteria, the articles must cover requirements that are based on human- interaction with the technology.

Second Round

The search terms (Requirements AND Medical AND Classifier) were used. The scope of the search was limited from 2012 to the present and the term ‘Machine Learning’ was used to filter the results.

The inclusion criteria were the same as in the first search; articles must cover requirements that are based on human- interaction with the technology.

Third Round

Due to a lack of information relating to human-based requirements for deep-learning technology in the medical sector. Together with two experts from the university of Twente and two experts from Philips, it was discussed how to enhance the literature research further. The focus shifted to the use of the new, image-based biomarkers which is another distinguishing factor of the new technology. The search term (Imaging AND Biomarker AND Requirement) was used on PubMed and on Scopus. In order to be applicable for search, the sources had to cope with imaging-based Biomarkers and contain information about requirements for them**.**

Fourth Round

To conclude the literature research, a narrative search on google scholar was performed. Due to the huge numbers of sources available on google scholar, the search was limited to the time from 2014 and filtered according to 'highest relevance'. The search was aimed at factors that influence the adoption of A.I. systems in general and was not limited to the medical context. A combination of different search terms with the keywords 'Deep Learning, Artificial Intelligence, Implementation, Barriers, Facilitators' were used. Additionally, this search was utilized to gather information about specific topics that emerged during the literature research.

Analysis of Literature Research

A thematic analysis of the found literature was conducted in a narrative way. The themes that emerged out of the literature were summarized and a thematic overview was created. After this first step, the different themes were analyzed according to similarities and common topics. An overview about different topics and sub-topics was created in form of a list and topics that shared commonalities were combined. This list was discussed with three different expert and refined according to their feedback. This iterative process led to the synthesis of six overarching topics divided into 19 sub-themes.

3.2.3 Synthesis of Information into questionnaire

In order to create the new list, the adapted POCKET checklist was merged with the results of the literature research to adopt the POCKET towards the use of Deep Learning Technology in the mental healthcare setting. Factors from the literature research that matched with the POCKET were deemed independent from the InVitro nature of the technology and therefore retained in the new list while factors that related to the portability of the technology and the comparison with laboratory results

were excluded. Additionally, factors found during the literature research were added to the list. This matching procedure resulted in the creation of a new list tailored towards the use of Deep Learning Technology. The list was presented to two experts from the field of human factors and health care innovation. The feedback was used in order to revise and validate the checklist until a structure for the new list emerged. After the new structure was finished, the specific evidence requirements from the POCKET were incorporated into the structure as well as the specific requirements that emerged during the literature research. This process was iterated until the checklist was approved. After that, evidence requirements that were similar were merged in order to reduce abundant items. The final product was the 'InSilicioEvidenceTool', a list of evidence requirements divided into 4 overarching themes, 11 sub-themes and 71 potential evidence requirements, it can be found in Appendix A. As a last step, the higher-level themes were discussed in interviews with experts from the field of Radiology, Neuropsychology and Psychotherapy. Due to the high number of evidence requirements, the intended evaluation of these through the online questionnaire and the exploratory nature of the interviews, these were not evaluated during the interviews but only used as potential prompts.

3.3 Data Collection

3.3.1 Design

The data gathering was divided into two parts. Ethical approval was obtained by the ethic commission of the University of Twente.

The new questionnaire was diffused online on the Qualtrics platform. The objective was to receive expert evaluation of the evidence requirements according to their importance for the implementation of the technology.

Next to the checklist, eight interviews with experts from the fields of Radiology, Psychotherapy and Neuropsychology were conducted. The goal of the interviews was to explore the topic from the point of the practitioners who might use the technology in the future. The interview schedule was created based on the preliminary work and aimed to review the literature-based adoption of the original POCKET. Additionally, the interviews were designed to explore additional factors that might be important when considering requirements for this type of technology. The information obtained was synthesized into one coherent list.

3.3.2 Interviews

Participants

It was tried to include mental health experts, who are the proposed target group of the technology as well as radiologist, who have expertise in the field of medical imaging. This rather diverse group was chosen according to the qualitative background of the study. Since the focus of the study lies on the creation of new information and hypotheses, the heterogenous target group was adopted to maximize the potential amount of new information. In other words, the creation of new information was prioritized over the generalizability of this information.

In order to find participants for the interviews, the social network as well as the internet was utilized. When potential participants were found, a formal invitation was sent out and/or it was tried to reach them via the phone, if a number was available. Ultimately, the only participants were people from the direct social environment and people who were linked to this. All participants were German which is explained by the fact that they came from the social surrounding of the researcher.

Eight interviews were conducted. The sample included 2 female and 6 male participants with a mean age of 52.75 years and 23.75 years of working experience. The sample consisted of one radiologist, one psychiatrist, six therapists and two neuropsychologists (one of them also a therapist). All participants signed an informed consent form.

The interviews were conducted on a completely voluntary basis and no incentive was given to the participants. Regarding the duration of the interviews and the limited time of most health practitioners, this might explain the low participation outside of the social network of the researcher.

Procedure

At the beginning of the interviews, the general way of how the technology works was presented and potential questions about the topic were answered. A semi-structured method was chosen for the interviews. The first half of the interviews was conducted in an open manner. This was done to create a less formal atmosphere and to allow exploration of the theme and more free thinking. Again, the rationale behind that was to maximize the potential amount of new information. After this unstructured beginning, open questions were asked to explore the topics that emerged during the literature research. The Interview schedule can be found in appendix B.

Data Processing

The interviews differed in their duration from 42 minutes to 78 minutes with a mean duration of 57 minutes. The interviews were audio recorded. Four interviews were conducted in person, three were conducted via phone and one via skype. Using the ATLAS.TI software, the interviews were coded. A thematic analysis was carried out to find common themes in the codes and extract the relevant information from the interviews. Due to problems with the audio recording software, one interview was not audio recorded. The results from this interview were summarized.

3.3.3 Questionnaire

Participants

To recruit participants for the online questionnaire, different approaches were utilized. First, it was tried to contact participants through the social networks of the researcher at the University of Twente, the Philips Research Institute in Eindhoven and the Imperial College in London.

Second, it was tried to recruit participants through LinkedIn. Groups that related to the theme of radiology, machine learning and/or mental health were joined, and the survey was advertised. Additionally, the administrators of the groups were contacted directly to obtain permission to post in the groups and to advertise the list to these administrators.

As a third step, people who might have expertise in the field were searched for and a formal invitation to participate in the survey and/or the interviews was sent via email.

Despite these efforts, only five people participated in the list from which only 2 completed the list. The two participants who completed the list were native German speakers who also participated in the interviews and are part of the direct social environment of the researcher. They also completed the list in the presence of the researcher which was necessary because the items needed to be translated to them. This lack of participants made the statistical analysis of the results invalid and resulted in lack of empirical evaluation of the list. This problem will be elucidated in the discussion. The unvalidated checklist can be found in the appendix.

Procedure

To evaluate the evidence requirements that were synthesized in the preliminary work, an online questionnaire was developed. The questionnaire was diffused online using the Qualtrics Survey Software. A five-point Likert Scale was used in order to evaluate the importance of the respective evidence requirements for the implementation of the technology. Before the begin of the study the participants were asked to fill out an informed consent form. Demographic data was gathered about the gender of the participants, the professional status and the area of expertise, the age and the working experience. A short use case was presented to provide a functional example of the technology. After the introduction, the participants were asked to rate the importance of the specific requirements with the following options:

1. Extremely Important
2. Very Important
3. Moderately Important
4. Slightly Important
5. Not at all important
6. I do not know

Option 6 was integrated in order to prevent invalid feedback.

Synthesis of Results

The initial plan was to merge the evaluated evidence requirements with the results of the interviews. Because it was not possible to empirically evaluate the new evidence requirements, it was decided to desist from that approach and to present the two results separate from each other.

4 Results

4.1 Literature Research

4.1.1 Additional Value

In the context of Radiology, a potential factor for the success of DLMSS and new devices in general is that they must create an additional value for the delivery of medical care (Thrall, Li, Li, Cruz, Do, Dreyer & Brink, 2018). The general need to make the system more efficient seems to be an important driver for this. The creation of an additional value connects to the outcome expectations that physicians have for a new medical device (Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. 2018). Outcome expectations describe whether the physician thinks that the technology will achieve its

objectives and is shown to be an important predictor for the implementation of innovation and the intention to use DLMSS (Fleuren, Paulussen, Van Dommelen & Van Buuren, 2014; Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. 2018). This indicates that the provision of evidence for an actual benefit that is induced by the technology needs to match the expectations of physician and that it is therefore important to include the creation of this type of evidence in the evaluation process of the technology. An important outcome expectation that can be found in literature is the impact of the test for patient management strategies and in terms of health outcomes and costs (Kip et al., 2018). More concrete factors include increased diagnostic certainty, faster turnaround times for results, better patient outcomes and better quality of work (Thrall, Li, Li, Cruz, Do, Dreyer & Brink, 2018). It should also be tried to explore expectations that relate to the specific use case which is done in the interviews.

Creation of additional Value for clinical practice.

1. Financial Benefit
2. Better Health Outcomes
3. Increased Diagnostic Certainty
4. Faster Turnaround time for Tests
5. Better Quality of Work

4.1.2 Training Data

A potential barrier for the technology is the set of training data that is needed for the predictive model to work (Arbabshirani, Plis, Sui & Calhoun, 2017; Thrall et al., 2018). Because the algorithm needs a pre-defined outcome (e.g.: Person has the pathology or not), the lack of an objective gold standard for mental illnesses is problematic for the implementation of the technology (Fu & Costafreda, 2013). It appears important to include evidence for the process that led to the definition of the right outcome in the evaluation of the technology.

Another barrier that falls into that category is the generalizability of the training sample (Thrall et al., 2018). The Accuracy of the model for populations that were not included in the training sample is a potential concern and evidence that indicates the generalizability of the algorithm should be included (Thrall et al., 2018). Evidence generation should, therefore, include a very diverse population in the training sample and a detailed description of it. Evidence that the algorithm also works in samples that show different characteristics than the training sample should also be included. For this, the importance of multisite studies using the same algorithm is highlighted by the literature (Crommelin et al., 2014; O'Connor et al., 2017). In general, the quality of the training sample is one of the biggest concerns and a potential barrier for the use of deep-learning technology because it builds the foundation for the decision function. Additionally, a huge amount of patient data is needed to not

overfit the model with features unrelated to the actual outcome (Arbabshirani, Plis, Sui & Calhoun, 2017). Overfitting describes the use of features by the model that are not related to the actual outcome (Arbabshirani, Plis, Sui & Calhoun, 2017).

Quality of Training Data

1. Lack of a gold Standard for Diagnosis of Training Sample
2. Generalizability of Predictive Model

4.1.3 Administration

The next theme that emerged from the literature addresses potential regulatory barriers. One of them is connected to the creation of training samples that stand in contrasts to the high demands of privacy and confidentiality that is apparent in the medical setting (Pesapane, Volonté, Codari & Sardanelli, 2018). The training samples must be stored in databases on which different institutions would need to have access which can be problematic considering data protection regulations in the medical field (Pesapane, Volonté, Codari & Sardanelli, 2018). The same principle is apparent in practice. In the case of mental illnesses, a therapist would, for example, need to have access the patient data from the radiologic database in which it is saved. This was also mentioned as a potential barrier for the use of DLMSS for mental illnesses (Philips Premium Report, 2018). While this appears to be a valid barrier for the implementation of the new technology, it is less of an evidence requirement for clinicians but more connected to the policies that are apparent in the field of medicine and medical device evaluation (Bergsland, Elle & Fosse, 2014). The new device must be able to receive regulatory approval before it can be implemented which is indeed depending on the relevant department like the FDA in the United States (Pesapane, Volonté, Codari & Sardanelli 2018, Van Ginneken, Schaefer-Prokop & Prokop, 2011).

Other regulatory factors include the accountability of such devices in the case of errors (Pesapane, Volonté, Codari, Sardanelli, 2018). There appears to be a need for a clear definition of accountability for the results of the algorithm before it can be implemented in the context of clinical decision making (Pesapane, Volonté, Codari, Sardanelli, 2018). Furthermore, the technology on its own also needs a clear definition before it can be evaluated (E.g.: A.I., Machine Learning, Deep Learning, Medical Support System, Medical Decision Support, etc.). In this regard, it may also be important to differentiate between decision making and data analysis, with the latter one having a higher potential to meet regulatory requirements (Pesapane, Volonté, Codari & Sardanelli, 2018). However, it is questionable whether this solves the problem or merely postpones it to the physician who then must base his decision on the data analysis and is accountable for potential mistakes which might indeed be another barrier.

<u>Regulations</u>

1. Regulatory approval
2. Confidentiality
3. Accountability

4.1.4 Explainability

Another potential barrier that is connected to the deep learning nature of the technology is explainability (Hengstler, Enkel & Duelli, 2016; Petkovic, Altman, Wong & Vigil, 2018; Park, Chang & Nam, 2018). Explainability falls under the category of usability requirement for this type of technology (Petkovic, Altman, Wong & Vigil, 2018).

The reason why explainability appears to be crucial is that deep learning technology has already reached a technical complexity in which the 'reason' with which it decides can often not be reconstructed anymore (Pesapane, Volonté, Codari & Sardanelli, 2018). For medical practice, this would mean that the clinician would have to either accept the results of the technology or not, without having any idea how it came to this conclusion. For a technology that aims to aid clinical decision making, this can be a potential barrier, especially in the context of an evidence-driven domain like medicine. It appears crucial that the technology can provide an explanation for its reasoning to the clinician.

Explainability can be further differentiated into model explainability and sample explainability (Petkovic, Altman, Wong & Vigil, 2018). Model explainability refers to why and how the algorithm works in general while sample explainability connects to how the algorithm made a specific decision for a specific sample (Petkovic, Altman, Wong & Vigil, 2018). It appears reasonable on this background to include evidence on both, the general way deep learning technology works, and the reasons why the technology made a specific decision in a specific case. It is advisable to include some representation of decision making in the interface design of the technology that the physician can work with. Potential solutions in this regard include the use of a sequential, tree-based classifier to indicate the different features and risk factors that the algorithm uses for decision making (Si, Yakushev & Li, 2017; Petkovic, D., Altman, R., Wong, M., & Vigil, A. 2018).

Explainability

1. Sample Explainability
2. Model Explainability

4.1.5 Validity

The validity of the decision that the algorithm makes is another potential barrier to the implementation of the technology (Beach, 2017; Isaac & Gispen-de Wied). The basic problem here is how to trust the

decision function of the algorithm to use valid means for the decision it makes. Especially in the case of mental disorders in which very little is known about the neuronal correlates of the pathologies, the features that are found by the technology to be indicators of a certain pathology are hard to verify causally as being connected to it. The 'black box' nature of the technology combined with the high dependability on the training sample and the lack of a gold standard for diagnosis further complicate this (Fu & Costafreda, 2013).

Despite the deep learning features of the algorithms, the use of imaging-based biomarkers is another concern for the validity. In order to be used for medical decision making, the predictive features that are identified by the technology must be valid biomarkers for the pathology that the algorithm aims to predict (Isaac & De Wied, 2015). For example; if the technology finds functional deviations in the brain to be indicators for a certain pathology, these would then be used as biomarkers for this pathology. For this, a cut-off score that enables this feature to differentiate between the people who have the pathology and people who do not have the pathology is needed which is indeed based on the training sample (Arbabshirani, Plis, Sui & Calhoun, 2017). 'Classical' biomarkers use enzymes or certain genetic variations while brain-imaging based biomarkers are potentially harder to define as being either normal or abnormal. This indeed connects to the generalizability of the training sample. In order to define a feature as abnormal or normal, a cutoff score needs to be defined for which the standard variation in the feature that the algorithm uses needs to be known. All of this is information that should be included in the generation of evidence for such a technology.

Because the implementation of imaging-based biomarkers can be observed also outside the field of mental disorders and deep learning, there exists a substantial amount of literature on what evidence is needed in order to qualify a biomarker to be used as such.

4.1.6 Imaging-Based Biomarkers

The main question is how to make sure the features that the biomarker uses are valid neuronal correlates of the pathology. In the literature, different approaches were found to indicate this relationship. They can be thematically differentiated into 'Test Performance', 'Biological Validation' and 'Clinical Validation' (Huddy et al., 2018; O'Connor et al., 2017; Medeiros, 2017; Scher, Morris, Larson & Heller, 2013). They aim at the creation of scientific evidence to prove the applicability of the technology.

Test performance refers to statistical quality standards of test and measurement theory and results like sensitivity or specificity (Huddy et al., 2018, Van Ginneken, Schaefer-Prokop & Prokop 2011, Arbabshirani, Plis, Sui & Calhoun, 2017).

The biological validation of the biomarker aims to indicate the validity of for the indicated outcome (O'Connor et al., 2017; Medeiros, 2017). If a certain neuronal correlate is known to be affected through a specific illness, a biomarker which uses this as one of its features can be credited to be biologically validated to at least some degree (Medeiros, 2017). Biological Validation also relates to aspects like Reproducibility or Repeatability of Results and potential confounding factors (O'Connor et al., 2017). Especially the use of multisite studies in which different institutions using different populations are able to produce similar results appears to be an important evidence requirement for this (Hampel,Lista & Khachaturian, 2012; Jack, 2018). It also connects to the generalizability of the decision function.

Clinical Validation refers to the validation of the biomarker in practice through clinical trials and aims at the evidence for the general increase of quality of care (Huddy et al., 2018; Hampel, Lista & Khachaturian, 2012). A very important factor here appears to be the validation using measurable clinical endpoints (Beach, 2017).

The validity of Biomarker used by the predictive model

1. Test Performance (statistical Measures)
2. Biological Validation (biological rationale behind feature)
3. Clinical Validation (effectivity of technology in practice)

4.1.7 Subjective Factors

The last theme for potential barriers that was identified relates to subjective factors which are known to impact the uptake of medical innovation (WHO, 2010). Concerns about patient safety, changes in the provider/patient relationship or anxiety of staff to lose its job were mentioned as factors that have the potential to influence the adaption of health technology systems (Ludwick & Doucette, 2009; Pesapane, Codari & Sardanelly, 2018).

Subjective Factors

1. Concerns about Patient Safety
2. Concerns about change in Provider/Patient Relationship
3. Concerns to lose Job

Out of the literature research, 6 themes emerged that have the potential to influence the implementation of Deep Learning Based Medical Support systems. These can be further differentiated into 18 sub-themes. While the potential facilitators are more general and relate to the need to make the health-care system more efficient, the barriers are more specific to the deep-learning nature of the technology and the use of imaging-based biomarkers. For the different themes that were identified, it

became apparent that they are strongly interconnected, especially concerning the quality of the training sample and the emerging concerns about the validity of the technology. A summary of the emergent themes can be seen in table 1.

Table 1
Topics of potential evidence requirements for DLMSS

Creation of additional Value for clinical practice.
• Financial Benefit • Better Health Outcomes • Increased Diagnostic Certainty • Faster Turnaround time for Tests • Better Quality of Work
Quality of Training Data
• Lack of a gold Standard for Diagnosis of Training Sample • Generalizability of Predictive Model
Regulations
• Confidentiality of Data and Security • Regulatory Approval • Accountability
Explainability
• Sample Explainability • Model Explainability
The validity of Biomarker used by the predictive model
• Test Performance (Statistical measures) • Biological Validation (Validation of Biomarker) • Clinical Validation (Effectivity of technology in practice)
Subjective Factors
• Concerns about Patient Safety • Concerns about change in Provider/Patient Relationship • Concerns to lose Job

Note. The table comprises a synthesized list of the requirements that emerged from the literature research and the POCKET. The factors were matched, integrated in the list and reviewed with expert feedback.

4.2 Interviews

4.2.1 General Findings

The interviews were designed to receive feedback on the factors that emerged during the literature research and to explore other potential factors. They delivered a very broad range of information. In general, the attitude towards the principle working mechanism of the technology was rather positive. No one of the participants had concerns about the existence of biological correlates of thinking and behavior. As one participant said:

"It is quite clear for me that every thought has some form of biological correlate in the brain".

This was further represented by the fact that all participants stated to have interest in the technology in the context of research. One of the reasons was "scientific curiosity". Despite this personal motivation, all participants agreed that to be implemented into clinical practice, the technology must deliver some form of benefit. While there is some heterogeneity in the specific benefit that the technology should deliver and in the personal motivation to use the technology, an underlying pattern was that it must provide information that:

1) Cannot be produced through the current means.
2) Has a practical benefit for the patient.

Another point that became clear during the interviews is that the technology should be used as an additional source of information, not to make decisions on its own. It was seen as crucial by every participant that the practitioner has the interpretational sovereignty and that the new test is used in a similar way to already existing pen-and-paper tests.

4.2.2 Creation of Additional Value for Clinical Practice

In accordance with the literature, an added benefit appears to be the most important driver for the adoption of the technology. However, it was also seen as the biggest potential barrier for the implementation in clinical practice. While no interviewee stated a particular concern about the general way that the technology works or the A.I. aspect of it, the usefulness of it was questioned. Especially the use of the technology to diagnose a patient was regarded very critical. The tone in the interviews was that the diagnosis of a patient was, most of the time, a very easy and short exercise with less benefit. One interview stated the following: "If I want to diagnose a mental disorder, I use the ICD 10 and ask for the symptoms". This connects to the missing usefulness of the disorder categories for actual therapy, which was mentioned by all participants. One interviewee stated that the diagnosis "is something that you write in your report to give it to the insurance companies". Another quote that summarizes this quite well is the following:

> "Psychotherapy does not profit from the disease mongering of groups".

In other words, even though the diagnosis of the algorithm might be regarded as valid by the interviewees, they see no real benefit in it. This dictates that it is crucial to provide a rational for why the technology is used and how it aids in practice. The use of the technology should be aimed to create information that cannot be produced with the current means and that creates an actual benefit for the patient and the therapeutic process.

When asked about what classification would be needed to create an additional benefit, information about certain personality structures and individual information processing were mentioned to be of interest.

> For myself, I profited much more from structure types to understand how people process different situations […] to understand what type of character the person has."

Three interviewees mentioned that a way to generate a potential benefit for the therapeutic process would be the use of the technology as a feedback mechanism. In addition to the 'subjective' perception of the patient and the therapist, the technology could provide objective means to test whether the therapeutic process is successful in the form of pre-post scans. In this context, another interviewee mentioned that it would be very beneficial if the technology could provide results that indicate the best therapeutic decision for the patient. The use of the technology as a feedback mechanism and as a direct decision aid for therapy both connect to the overall factor of creating an additional benefit for the patient.

The benefit factors that related to enhanced efficiency, namely financial benefit and faster turnaround time for tests were regarded less important, especially by the therapists in the sample. One of the interviewees stated that this is because:

> "Therapists, compared with other health practitioners, have much more time with the patients".

Therefore, saving time and enhancing efficiency would not result in an additional benefit as experienced by the practitioner. An increased benefit through an increase in the quality of work was not regarded as a potential benefit of the technology either. However, the interviewees did not negate that an increase in working quality would be a driver for the technology but questioned how the technology should increase their perceived quality of work. This might connect to the increased time that therapists can spend with patients compared to other health professions.

The topic of 'Increased Diagnostic Certainty' received somewhat ambivalent answers. One of the interviewees mentioned that an increase in the certainty of the results would benefit him. This individual however works in the field of neuropsychology and does not diagnose patients according to

DSM or ICD categories but in the context of neurological disorders. In the context of ICD based classification of patients, the factor of increased diagnostic accuracy was not mentioned to be of importance for the new technology. This might connect to the ease of ICD and DSM diagnosis and the missing benefit mentioned in this context. However, the creation of increased diagnostic accuracy in general was not denied of being a potential driver for the technology and it might be important in the context of diagnosis outside the DSM and the ICD classification system.

The point of increased quality of work for the practitioner was not mentioned to be an important driver for the implementation of the technology. However, it was not that an enhanced quality of working was denied being a potential driver for the technology but no direct connection from the technology towards this aspect could be drawn. It is also important that quality of work in this context is different from a better quality of the therapeutic process. The item will still be added to the list as potential enhancers of additional value for practice.

4.2.3 Finances

The most interviewees agreed that the financial part is crucial for the implementation. However, this aspect was less important for the practitioner per se. A quote that summarizes this quite well is the following:

> "Who is paying for it?" followed by the statement "It should not cost me anything".

This statement connects to another interviewee who stated that it is crucial to know that the insurance company pays for that. Therefore, the provision of this kind of information for psychotherapists might be a more dichotomous item of whether the insurance company pays or not.

Another interviewee, a psychiatrist who has more control and responsibility over the financial part due to his position stated that for him, the way the payment is conducted is very important.

> "It will probably be a software that you have to license with a certain amount of usages. From test-psychology I know that you normally buy a certain amount of analyses, let's say a hundred and if they are expired you have to buy the next. Therefore, I would like to know the price. […] I am not a marketing expert but for us as a big hospital it would probably be best to buy the test once and can then use it for a thousand patients or something in that line"

This indicates that finances are an important driver for the adoption of the technology into the mental health care setting, but not necessarily so for individual therapy and the therapist. The creation of a financial benefit is probably an item that is more important for higher hierarchies of the system, namely the health insurance companies and policy makers which was mentioned by one of the interviewees.

Creation of additional Value for clinical practice.

- Better Health Outcomes
- Evidence for therapeutic benefit of technology
- Patient Acceptance
- Increased Diagnostic Certainty
- *Better Quality of Work*

4.2.4 Generalizability

A potential concern that was mentioned in most interviews was the generalizability of the predictive model being developed by the algorithm. The main concern was that a huge number of confounding factors might exist that could contaminate the biomarker and therefore hinder the use of the algorithm for individual patients. This concern was especially highlighted by the two neuropsychologists in the sample. Since the algorithm is aimed to quantify differences in the neuronal correlates constituting certain disorders, every factor that might influence this neuronal correlate might also influence the predictive power of the decision function for single subject prediction. Medication was one example that could potentially have a significant influence on the biomarker. A comorbidity in the form of other mental, cognitive or neuronal disorders potentially distorting the biomarker was also mentioned. Other potential factors were age, sex or stress levels.

> "I would need to know on what basis the A.I. was developed. What patient groups were fed into it. You could probably give an unlimited amount of psychiatric diagnosis… you would probably also have to account for neuropsychiatric or neuronal disorders… Of course, you would also need to know the influence of medication on a scan, especially for fMRI, I could imagine that there are pharmacological effects that could distort it [the biomarker]. Can I differentiate an anxiety patient who takes medication from one that does not? To be honest I find it hard to imagine how to realize that".

To allow for the generalizability of the predictive model, it was regarded as very important to study literature and to test the influence of as many different factors on the biomarker as possible. The main concern in this regard is that due to different confounding factors being apparent in a single subject, the algorithm might give the wrong classification. Because the brain has a complexity that we do not fully understand (and some would say we are not even close to that), it might always be the case that some unknown independent variable is apparent in the single subject and distorts the validity of the predictive model. This asks for the generation of evidence for the validity of the predictive model for single subjects. Very detailed information of the training groups should be generated and their influence on the predictive features should be tested.

Gold Standard

The gold standard for diagnosis was regarded as less of a problem. While some participants stated that it would be enough for them to use a normal DSM or ICD checklist, other preferred multiple assessments of the same person to be sure about the validity of the diagnosis. However, as was indicated earlier, the validity of the missing gold standard appears to be a minor problem compared to the missing benefit of the diagnostic gold standard.

Image Acquisition

Related to the generalizability aspect was the use and provision of specific image acquisition protocols. Two participants who had experience with medical (brain) imaging technique stated that the scanner and the details of the image acquisition have a major impact on the results. One of the participants stated that during an active research they changed the scanner and that "even though the new scanner was structurally identical, you could see a significant difference in the statistical data". This indicates that it is important to provide detailed information on both the hardware and the software which is used. It might for example be a good idea to include calibration software for the scanner.

Cut-Off

Another factor relating to the generalizability aspect is the use of cut-off scores. It was mentioned by the interviewees that the nature of mental disorders would make it very hard to give specific classifications based on cut-off scores. Mental disorders were said to have "very soft borders compared to medical diagnoses". Even though cut-off scores exist in the scientific sense, the most interviewees regarded them as relatively useless for clinical practice. Again, the critique here was not directed on the validity of the scientific method but on the additional benefit that the results might bring for clinical practice. In this context, one of the participants mentioned the gap that is apparent between the scientific realm and the realm of therapy:

> "In science, I need to break down reality into practical variables, meaning I have to create a design out of it. In psychotherapy I try to perceive ramifications and connect this to the patients. For me it is another profession".

Another interviewee stated that there is a clear change in psychiatry that makes the use of cut-off scores appear quite outdated:

> "[…] One goes away from the division according to diagnosis. For example, today we have the autistic spectrum disorder. […] Today you see that different categories are actually very hard to differentiate. It is more like a cloud… a statistical probability the people belong to."

This statement connects to the fact that mental disorders experience a shift towards being seen dimensionally and that a lot of disorders are nowadays seen as an extreme form of traits that everybody possesses. Another interviewee also mentioned this:

> "[…] These are moods we all know. We all know a form of anxiety or a form of sadness. In some cases, they only get more extreme and reach a pathological significance".

The same interviewee also differentiated between spectrum disorders and biological disorders like Alzheimer.

An idea to overcome that potential barrier was a different form of presenting the information. Instead of using different features and biomarkers to give a classification of the subject in form of a cut-off score, the algorithm might present the predictive features to the practitioner and enable him to make his own decision. The interpretation of the results would then be handed over to the practitioner granting him the interpretational sovereignty. Again, the technology would be used as a decision aid, like other psychometric tests and would be used to create additional information to aid the decision of the practitioner. This would also help to decrease the concerns about the generalizability aspect because it would allow the practitioner to take a deeper look at the statistical model and to decide whether he/she can generalize the model on the patient.

<u>Generalizability</u>

Detailed analysis of Training data relation to potential confounders

1. Patients (E.g. Demographics)
2. Disorders (E.g. Comorbidities)
3. Assessment (E.g. Severity of Disorder)
4. Indication of Gold Standard (E.g. How was the diagnosis created?)

Provision of Image acquisition protocols and Scanner Data

Avoiding the use of cut-off scores.

4.2.5 Explainability

All participants regarded the interpretational sovereignty on the results of the technique as vital. This demands the algorithm to provide results in a way that the practitioner can use it to inform his decision making. Different ideas were proposed about how this might be done.

Most participants agreed that the algorithm should deliver some information about the features i.e.: the (neuronal-) correlates that the algorithm uses in its predictive model.

What was also regarded as crucial was the ability to compare the results of a single subject to different norm groups. On this way, it is possible to compare the subject to both healthy individuals and other groups of mental disorders.

> "It should be aimed specifically at the patient group that I treat and then […] it could for example provide some indication about the probability of the single subject belonging to one of those groups. It should provide a distinction to the healthy group but also some means about the specificity with which it belongs to one group compared with another one […] Someone who has schizophrenia might for example also have anxiety".

One of the interviewees also mentioned the creation of different standard brains or "Hirnmasken". This technique is already used in some programs and creates a norm brain based on the mean of the samples that it analyses. This technique allows a program to indicate single subject differences in brain structure or functionality towards a norm group. This can be done in both visual and statistical terms. It was for example proposed to provide the mean level of a group with its standard deviation to compare the single subject towards this. This also fits into the dimensional view of mental disorders. It was also deemed necessary to be able to compare different scans of a single subject over time in order to follow up potential changes and trends in the disorder.

Another interviewee proposed the idea to also create a huge number of subgroups in the training sample to reduce the influence of potential confounding factors. For example, if the training sample would include different age groups, the practitioner could use the group that is suited for the patient and thereby reduce the potential influence of age on the biomarker. With this principle, multiple groups might be created. Because it is unlikely that all confounding factors can be found in the training groups, one should also provide very detailed information about the people in the training data. This might additionally help the practitioner to decide whether the biomarker can be translated towards the individual patient. One participant also stated that the definition and characteristics of the 'healthy' comparison group should be given which makes sense relating to the spectrum nature of mental disorders.

In comparison to that, no interviewee stated to be particularly interested in the way that deep learning algorithms work in general. However, it appears reasonable to provide some information about the functional aspects of the technology as it was also done at the beginning of the interviews.

<u>Explainability</u>
Provision of Training Sample Data
Comparison with different Norm Groups
    Visual Comparison
    Statistic Comparison
    Indication of belonging to these norm groups (e.g.: probability, percentile in distribution, etc.)

Ability to compare longitudinal subject development.
Indication of features being used by the model.

4.2.6 Validity

All participants agreed that the evaluation of the technique must withstand scientific criteria to be considered valid. In the context of the deep learning nature, participants did not see the necessity for additional validation outside the 'normal' spectrum. As one participant put it:

> "Of course, medication is completely different from an Artificial Intelligence! But the evaluation that I demand from medical methods is the same for me, I would have the same demands.".

In general, the participants stated that they would trust in the external evaluation of the technology if it is done by an "independent and trustworthy" organization. One interviewee stated the following:

> "With very few exceptions I never take the effort to look into those studies".

Asked about what would constitute these exceptions, the interviewee stated that it was mostly due to personal interest.

Another interviewee stated that information about "who finances the project" and about the "academic advisor of the study" would be important to have a better assessment of the context in which the results are produced. This might indicate a certain distrust towards the entry of big industry into the mental health sector. One participant recalled an EEG diagnosis technique which he experienced negative because it appeared to aim mainly at generating profit and another criticized the 'industrialization' of the sector. This indicates that some form of subjective assessment of the company developing the algorithm and the context of the evaluation might also play a role for the decision to implement the technology. Connecting to the other themes, a driving factor here might be whether the practitioner is convinced that the technology aims to help the patient (to create an additional benefit) or not. Like the financial evidence for the technology, the validation of it also appears to be a dichotomous and externalized item. However, the provision of information about this evaluative process and its context appear to be relevant to the practitioner and should be provided. This asks for a transparent evaluative process.

Trust/Convergent validity

Next to the official validation of the technology, all interviewees stated that they would also have to see that the technology works when deciding about whether to use it. Asking about what would qualify a technology as working properly, the most participants said that it must fit with their own perception. The following statement summarizes this quite well:

> "When the scan provides me with a pattern that comes to a similar conclusion as I do during my evaluation then it would increase my diagnostic certainty. I would have two different data sources that correspond with each other. This would somewhat represent the idea of convergent validity. If two different methods reach the same conclusion".

This indicates that the decision about whether to implement and use the technology in practice is a more personal choice which seems to be influenced by the subjective appraisal of the technology. Another interviewee mentioned the following:

> "It is important that the technology is valid and works as it should from the beginning because if the technology does not work well people will say it is nonsense and that mindset will then be confirmed. The technology should not be implemented to early".

This statement also indicates that the appraisal of new technology might be influenced by the subjective perception of it. Another interviewee stated that "trust building measurements" would be needed for him in order to implement the technology into practice. One should start to implement the technology for tasks that have a low probability of creating errors and that allow the practitioner to compare the results to his own perception.

Overall, the validation process can be segmented into two sets. One is the official validation which takes the form of a dichotomous yes/no item while the other one is the subjective evaluation of the practitioner that appears to be more dimensional.

External Validation

- Clinical Validation
- Test Performance
- Biological Validation

Internal Validation

- Convergent Validation
- Trust

4.2.7 Subjective Factors

The Evaluation of the subjective factors showed that there appears to be no concern about losing their job. Patient safety was also less of a concern. However, the data security which somewhat connects to patient safety was important for all of the participants. The data security aspect will be discussed in a section of its own. Three participants stated that they think that there will be a change in the provider/patient relationship relating to factors such as telemedicine and less direct contact with the patient. This change however was regarded as a general change in the healthcare system and not directly related to the implementation of the new technology.

Ethics

An important theme that emerged was the ethical consideration about the technology. In two interviews, it was mentioned that in some cases, people might not want to know whether they have a disorder. This relates somewhat to the factor of an increased benefit in therapy. If the categorization of a patient does not lead to any form of benefit in therapy, the categorization might result in adverse consequences on side of the patient. One of the interviewees gave the example of Chorea Huntington, for which there is no therapy yet that can cure the disease. Another theme that was mentioned by one of the interviewees is the connection of health insurance contributions to mental health.

> "[…] for example; insurance company XY says that if you make an insurance at our company and we make a brain scan of you and we see that there is no threat of developing a mental disorder, you will get a discount of 200 euros a month"

The interviewee stated that this would be something he would clearly see as an ethical barrier when thinking about whether to use the technology. An aspect that somewhat relates to that was the potential stigmatization of patients. This theme however received somewhat ambivalent answers. In general, the most interviewees thought that the provision of neuronal correlates for the pathology would help the most patients to better accept the disorder.

> "If people are provided with a physiological correlate, people might have the feeling that it is not only an imagination but that it actually exists that you can see. I think that would be better for them"

However, this is due to the appraisal of the patient and it was also mentioned that it might have adverse consequences. Three interviewees mentioned that this type of technology might also lead to disease mongering and stigmatization of patients. Again, the most important aspect for the ethical consideration was the patient. Most importantly, the patient must accept the technology and there seems to be no grey area in this regard. This might be influenced by that fact that therapy is very human centered and requires cooperation. The therapists and the patient must work together, and missing acceptance of the technology could lead to missing cooperation which would ultimately negate the benefit for the patient. It should therefore be aimed to create evidence for patient acceptance.

<u>Ethics</u>

No disease mongering
No connection of mental health and insurance contribution
Patient Acceptance

Data Security

The data security aspect also produced heterogenous information. What was crucial in all interviews was to comply with the official regulations. For some of the interviewees this was enough, and they had little personal concerns about data security themselves. However, four also indicated that it would be very important for them to guarantee the security of the data and two mentioned that they, subjectively, preferred the data in form of a physical Medium.

A relating theme that emerged was data anonymization. One of the interviewees who works in the field of interpreting medical brain imaging proposed the idea to remove the information from the radiologic image that would identify the patient:

> "Eventually, it is not a big problem to anonymize an MRT. You can anonymize certain DICOM entries [which is the data format in radiology]. Next to the image information there is additional information like for example about the examiner, the name of the patients, the date of birth, address, but also magnetic field strength and other stuff like that. And you can eliminate these fields in the data. […] Today you can even cut out the information about the face so that no one is able to reconstruct this information later."

This view was also shared by two other interviewees. Data anonymization could potentially help to reduce fears about data security without losing the ability to transfer data via networks. In general, the most participants expressed that they do not believe in something as absolute data security. One participant summarized it as follows:

> "It is a very sensible information that must be protected accordingly. But I do not know whether this is possible nowadays. I am more and more disillusioned in this way. Of course, we say that the security systems are isolated towards the outside and I think that you would not have any chance to access them as an individual person. But if the Chinese intelligence service tries to get the data, they will get it."

This statement should be interpreted as a functional example using a hyperbole, but it represents a mind setting about data security that was apparent between the lines every time the theme was talked about. Data anonymization therefore might be a more realistic way of reducing the barrier of data security. However, the data must still be in a format that can be used in the context of the single subject what asks for balance. It might be possible to have some sort of two-factor authentication in the future, similar to the way that cryptocurrencies work.

4.3 Synthesis of Literature Review and Interviews

The results of the interviews were used to adopt the already revised POCKET towards requirements for DLMSS in the field of mental disorders. Like the approach that was adopted when revising the POCKET with literature, the results were matched, and it was looked for similarities and potential differences. The final list of is presented in table 2.

Overall, it appears that the information gathered in the preliminary work can be translated towards the implementation of DLMSS into the mental health care setting. However, the qualitative nature of therapy dictates a somewhat different approach for the technology. The creation of efficiency and cut-off diagnoses results in relatively less experienced benefit for the therapeutic process which must be coped with when considering implementation of the technology and potential use-cases. This can be explained by differences in the outcome expectations which was apparent in the literature research. Due to this, the use-case of the technology should tailor to the expectation of the practitioners and evidence in this regard should be provided. Generally, the creation of information not yet producible and a direct connection to the therapeutic process and therefore the benefit of the patient can be used. It appears reasonable to integrate practitioners in the creation of potential use cases to guarantee that these two points are fulfilled to close the potential gap between science and practice.

Another result from the interviews is that some factors affecting the implementation of the technology appear to be dichotomous, such as data security and approval by official medical guidelines and institutions while some appear to be more ‘dimensional’ and subjective, such as convergent validity. Due to this, the themes of validation and data security are grouped into the category of ‘regulatory factors’. A factor that might explain this is whether the end-user can exert control over the relevant factor. A good example are the finances. The private psychotherapists in the sample are paid by the German health insurance and are therefore financially dependent on their decisions. This means that potential evidence requirements for the financial benefit must be presented to the insurance companies and not to the therapists. If the health insurance company does not pay for the use of the technology, the therapists cannot consider using it. Only if the company gives their okay and finances the use of the technology is the therapist able to use it. If these requirements are fulfilled, the second battery of requirements such as convergent validity and explainability gets important because the decisive power is translated towards the therapist. However, like so often in systems, there is a reciprocal connection between these two. The financial benefit for the insurance company is dependent on the effectivity of the therapist which is indeed dependent on his personal appraisal and use of the technology. The factor of ‘accountability’ appears to be non-relevant because the technology is not wanted to make decisions but to provide information that can inform them.

The subjective factors that emerged during the literature research appear to be less relevant. However, the themes of convergent validity, trust and ethics appear to be crucial and can be grouped into this

category. For the factor of ethics, it appears important that the results are not used in any other way than to inform and enhance the treatment of the patient. The factor of ‘Ethic’ is grouped into the category of ‘subjective factors’.

Two very important and interconnected factors are the generalizability-, and the explainability aspect. Very detailed information about the training sample should be provided in order to allow the practitioner to generalize the predictive model towards single-subject predictions. This asks for the generation of this information during the training phase and for the presentation of this information in clinical practice. This information should indeed be tailored to the dimensional and heterogenous nature of mental disorders and could for example be provided in visual and statistical terms. The adapted list of factors influencing the adoption of DLMSS in the context of the mental health setting can be seen in table 2. For a better overview, the external factors influencing the adoption of the technology are labeled as ‘Regulatory Factors’ and the factors that were important for the subjective appraisal of the technology are labeled ‘Personal Factors’.

Table 2

Overview of topics for evidence requirements.

Regulatory Factors	**Personal Factors**
Regulatory approval as medical product • Technical Validation • Clinical Validation • Biological Validation	Creation of additional Value for clinical practice • Better Health Outcomes • Better Quality of Care • Technology must create therapeutic benefit. • Patient must approve the technology. • *Increased Diagnostic Certainty* • *Better Quality of Work*

Data Security	Subjective Factors
	• Trust
	• Convergent Validation
	• Ethical considerations
	  o Prevention of disease mongering.
	  o No connection of mental health

	and insurance contribution.

Health Insurance Approval	Generalizability
	• Detailed analysis of Training data.
	  o Patient data (E.g. Demographics)
	  o Disorders (E.g. Comorbidities)
	  o Assessment (E.g. Severity of Disorder)
	  o Indication of Gold Standard (E.g. Who gave the diagnosis?)
	• Provision of image acquisition protocols and scanner related data.

	Explainability
	• Provision of Training Sample Data
	• Indication of Different Norm Groups
	  o Visual Comparison
	  o Statistical Comparison
	  o Indication of belonging to these norm groups (e.g.: probability, percentile in distribution, etc.)
	• Ability to compare longitudinal subject development.
	• Provision of features being used by the model.

Note. The factors of 'Increased Diagnostic Certainty' and 'Better Quality of Work' were both regarded as important in general. However, no direct link could be drawn to the technology.

5 Discussion

The presented work is highly qualitative and should be regarded in the context of preliminary work for the alignment of DLMSS towards the mental health sector. The strong points of the work are that it provides a good overview about the topic and that it was created in a multidisciplinary setting which allowed the creation of a very broad theme of topics. The qualitative nature of the work also helped to broaden the scope of this work. A short disclaimer that should be given here is that no technical experts in from the field of deep learning was included in the sample and that the results should be evaluated and eventually adapted towards what is possible.

The weak points of the work connect to the qualitative nature and the generalizability aspect of the results. The sample is limited to German health practitioners which might induce a bias on the data. There is a certain variance in user requirements (Wiegers, 2003) and due to the small size of the sample, the generalization of the data should also be regarded with caution.

Another important factor in this regard is that the interviews and the interpretation of these were conducted by a single researcher. This might impose somewhat of a bias on the results. However, it was tried to compensate for this potential lack of objectivity with the literature research and the integration of four different experts into the whole process.

Nevertheless, future work is needed in order to see whether the results can be translated towards the setting in general. In the following section, it will be tried to compare the gathered results in the current research that is available on the topic. After that, a section about the limitations of the study, the lack of participants, associated consequences for the results and potential reasons for it will be presented.

5.1 Integration of results with available research

Implementation into clinical practice

The biggest concerns seem to be the generation of a benefit for therapy and ultimately for the patient. An important factor that was mentioned was a general lack of additional benefit that is created through the diagnosis of mental disorders. In general, it fits in the belief that diagnostic codes are mainly used for administrative purposes (Reed et al., 2019). In a recent survey, 68.1% of mental health professionals reported that they use the classification system for administrative and billing purposes (First et al., 2018). However, 57.4 % of the respondents also reported that they use it in a systematic way (First et al., 2018), indicating that reality might not be as black and white as it might be indicated by this study (Reed et al., 2019). Nevertheless, the current diagnostic classification seems to need some overhaul and a new ICD version is about to get implemented in May 2019 (Reed et al., 2019). Congruent with the findings of this study, clinical utility is an important driver for the new classification systems of the ICD-11 (Reed et al., 2019). It indicates that the mentioned gap between psychological research and the therapeutic practice is an important consideration when designing technology for this setting. It also indicates the ‘valley of death’, which was used as a metaphor to describe the gap between medical research and practice in general (Beach, 2017), is translatable to the mental health care setting. It might be a good idea to integrate mental health practitioners in creation of such technology to guarantee clinical utility. A similar approach was also used for the creation of the ICD-11 (Reed et al., 2019). Due to these circumstances, the automatic classification of patient groups according to current ICD criteria appears to be a barrier for the integration into clinical

practice. Due to the central role of patient acceptance, evidence for this aspect should also be generated in the evaluative process of the technology.

Generalizability

Another argument against the automatic diagnosis through the system is the heterogeneity and the dimensionality of mental disorders which makes it difficult to generalize the predictive model of the algorithm towards single subjects. Especially the use of cut-of scores that would potentially constitute the automated diagnosis of a patient was regarded with caution. This finding is also congruent to the 'new' approach that is chosen for mental illnesses. For the ICD-11 CDDG, the use "arbitrary cutoffs and precise requirements related to symptom counts and duration are generally avoided" (Reed et al., 2019, p.4). A driving factor in this regard is the flexible nature of clinical judgements (Reed et al., 2019). Another potential concerns for the generalizability aspect was the dimensional nature of mental disorders that was mentioned by the interviewees. This is also concurrent to recent research on mental illnesses and to the developments in the new ICD-11 CDDG (Reed et al., 2019). Clark, Cuthbert, Fernández, Narrow & Reed, (2017), indicated the dimensional nature of mental disorders and the problematic nature of specific thresholds. Moreover, they even criticize the search for specific etiologic causes of psychopathology and propose a much more heterogenous view on mental disorders, also criticizing the simplified and dichotomous background of the whole 'nature vs nurture' debate (Clark et al., 2017).

Overall, it can be observed that there is an ongoing change in the way that mental disorders are seen. Modern research as well as therapeutic practice appears to be incongruent to the more dichotomous and cut-off-based view of mental disorders. Two of the interviewees mentioned the existence of underlying personality structures and data processing modalities of patients. This point can also be found in literature. Markon, (2010) indicated common features of psychopathological structures and gave new implications for the nature of mental disorders.

Eventually, the integration of DLMSS into the mental health setting must cope with the current changes that are ongoing in this setting and the classification of patients into diagnostic categories might soon be a relict. The ongoing changes also indicate the limitations about the knowledge we possess about the nature of mental disorders. Even though there is a constant increase of information on structural and functional correlates of mental disorders (Gong, 2017), the interactions and especially the causal mechanisms should be regarded with caution. Congruent to the interviews, there are numerous potential factors that might interact and distort the predictive validity of these correlates and thereby hinder the use of the predictive models for single subjects.

What is interesting in this regard was the mentioned interest in the integration of the technology into research. Based on this background, the properties of the technology might be able to create a benefit in therapeutic practice through the integration into the research setting. An interesting idea that also came up during the interviews was the use of completely unguided machine learning techniques for this setting. This technique might allow for the creation of new insights into cognitive structures and perceptual processing modalities of mental disorders and might be utilized as an active driver for the new discipline of 'psychoradiology' (Lui et al., 2016). However, whether or not this approach is realistic must be analyzed from a more technical machine learning perspective which exceeds the scope of this thesis.

Validation

The information obtained indicates that validation of the DLMSS can be categorized into two aspect. An external, regulatory validation through official medical guidelines and techniques, and an internal and personal validation of the product.

Regulatory validation

For the regulatory validation, the most important factor was the positive validation of the technology as a medical product. In general, there was little critique of this process and it appears that quality criteria that are used there are deemed acceptable. However, due to the novelty of the technique it appears to be very difficult for this type of technology to fulfill the strict criteria that are apparent in the evaluation process (Pesapane et al., 2018). A new regulatory framework is needed in order to allow the integration of the technology into the medical setting and there is a lot of effort to create these (FDA, 2019). The approach is still very recent, but the proposed categories of clinical evaluation seem to be concurrent to the ones found in the literature. The three types of validation include (FDA, 2019, p.9):

1. Valid Clinical Association, which describes a direct link between the output of the algorithm and the targeted clinical condition.
2. Analytical Validation, which describes the creation of 'accurate, reliable and precise output data'
3. Clinical Validation, which is accomplished when the intended purpose of the data is achieved in clinical practice.

The point of clinical validation is very similar to the one found during the literature research and analytical validation combines the aspects of 'Technical' and 'Biological' Validation. The point of 'valid clinical association' seems to fit into the category of an increased benefit for the technology.

Overall, the validation of the technology as a medical device is an inevitable barrier for the integration of the technology into the medical setting. Without advances on the regulatory side, it appears unrealistic for the technology to bridge the gap from science into medical practice. The concept of regulatory validation as a dichotomous evaluation variable seems to hold true.

Trust and convergent validity

The interviews indicated that in order to use the technology, it is necessary to trust it and that the results of the technology overlap with the personal assessment and perception of the interviewees. In the context of automation, trust can be an important predictor of usage behavior (Lee & See, 2004). Congruent to the interviews, trust into the organization who develops a technology also plays a role in the usage behavior of automation techniques (Lee & See, 2004). While the empiric literature on the acceptance of A.I. technology is still very sparse, early results indicate that trust in the innovating firm is important for the general trust that people put in A.I. technology (Hengstler, 2016). Especially transparent communication from the developing company appears to be an important driver for trusting the product (Hengstler, 2016). In this context, the communication of concrete information that is based on a specific application appears crucial and should be provided (Hengstler, 2016). This is an important aspect to integrate into the regulatory evaluation of such technology.

The same study also highlights the importance of regulatory approval and the creation of specific policies in order to guarantee trust in the performance of the technology (Hengstler, 2016). This connects to the mentioned factor of regulatory validation of the technology. It is also highlighted that these performance standards should include ethical considerations, congruent to the results of the interviews (Hengstler, 2016). Another potential concern from the interviews was the creation of data security, which can also be found as an important factor for the creation of trust in A.I. technology and information about it should be provided to the end-user (Hengstler, 2016). Somewhat congruent to the notion of 'convergent validation' is the concept of 'trialability', which is identified by the same study. Trialability refers to the direct experience with the technology and aims to let the end-user see that the technology works (Hengstler, 2016). This factors also connects to the outcome expectation of practitioners in that they must see by themselves that these are fulfilled which is indeed needed to make sure that the technology actually creates an additional benefit (Fleuren, Paulussen, Van Dommelen & Van Buuren, 2014; Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. 2018).

Explainability - Usability

The importance of usability for technical devices can be found in a substantial amount of literature. In the study of Hengstler, (2016), usability is additionally highlighted as an important determinant for

trust in the technology. However, the study of Hengstler focusses on A.I. in relation to automation. In the context of the interviews, the potential use case of the DLMSS appears to be to aid in clinical decision making and not to automate them. While the connection of trust and usability might therefore be somewhat different, the importance of usability, especially relating to the explainability aspect, is probably even higher.

Because the DLMSS is a software program, the creation of a user-interface to present the information to the visual system of the practitioner appears to be a valid medium to translate the information from the system to the practitioner. The advantage of the technology lies in the detection of patterns that are not detectable by humans. In the context of radiology, this is mostly due to limitations in the visual capabilities that we possess (Drew et al., 2013). Due to this, an interface must be created that enables the translation of these patterns to the practitioner without creating information overload. For this, it is important to reduce the complexity of the data patterns (Lurie & Mason, 2007).

A big advantage of humans over A.I. is that we can equip our associative system with meaning (Sloman, 1996). This connects back to the 'classic' two-system approach of Kahneman, (1974). We possess a fast and associative system that is based on similarity and temporal contiguity and a slower system which is more analytical (Kahneman, Slovic & Tversky, 1982; Sloman, 1996). In general, the pattern recognition of us humans is superior compared to even the most sophisticated A.I. systems. However, our pattern recognition is somewhat limited by our perception and data must be translated into some perceptual modality, for example visuals and numbers, in order to be comprehended by us. Deep Learning technology on the other hand can automatically extract very complex high-level patterns directly from the data that would be very hard for us to comprehend (Najafabadi, Villanustre, Khoshgoftaar, Seliya, Wald & Muharemagic, 2015). One of the (many) weaknesses that the systems have however is to interpret the results and to make sense of it (Najafabadi et al., 2015). This way of 'thinking' somewhat matches with our second system, which is rule based (Kahneman, 1974; Sloman, 1996), It might be a good idea to create a user interface that allows us to improve the limitations of our perceptual system and that utilizes the capabilities of our 'rule based' system. This fits in the notion of the provision with statistical data that was mentioned in the interviews. The most practitioners in the field of mental disorder have adequate knowledge about statistics, or at least obtained that during their study. Therefore, there is a rule-based pool of statistical knowledge that can potentially be utilized in order to make sense of this data. According to literature and the feedback of one of the interviewees, the provision of visual information might be a good way to augment the statistical data that is being presented. A classic approach would be the visualization of the data in form of distributions. Because the basis of the decision function are different features in brain structure and/or activation, it might also be a good idea to visualize this. This connects back to the signal detection theory and the reduction of visual complexity (Prins, 2016; Lurie & Mason, 2007). This would also allow to maximize the utilization of the capabilities in pattern recognition possessed

by the algorithm and the utilization of the higher cognitive functions of human beings. Eventually, visual information about the decisive features used by the algorithm in combination with statistical information about these features might a potential way to go. This approach also utilizes another advantage of our cognitive system. We can learn from reoccurring patterns and ascribe meaning to them over time if these patterns possess regularities which make them predictable and if there is enough opportunity to learn these (Kahneman & Egan, 2011). On this way, it might be possible to the practitioner to ascribe meaning to this information in relation to the patient over time. The provision of detailed patient information and the creation of different norm groups also fits in this approach by providing additional information which can be utilized in the learning process. This however must be balanced against the aspect of information overload.

Ethical Considerations

The relevance of ethical considerations for therapy can be found in numerous guidelines, like the American APA, the British BPS, the German DGP and so on. In all of them, the integrity and the rights of the patients are central. A consideration that should be made in the context of DLMSS is directed towards the use and the interpretation of neuronal correlates of the disorders. The technology is aimed to create different groups which are based on specific criteria. It was already mentioned during the interviews that a huge group of potential confounders might be present and that, due to the complexity of the brain, a large amount of them is probably still unknown. A potential confounder that was named was the age of the participants. It is likely that with an increase in the capabilities of both classifiers and imaging techniques, more and more factors will be discovered that can act as confounders on the predictive features and that enable the classifier to differentiate between groups. In accordance with the interviews, it appears very important to prevent the disease mongering of group differences in context of specific correlates of these differences. For example; it might be possible to detect significant group differences in neuronal correlates according to the age of patients. Similar findings might also be found concerning other characteristics of patients. It is important to acknowledge how little we still know about the brain and to differentiate between neuronal correlates and causal brain mechanism. This aspect was also highlighted by one of the interviewees. The misinterpretation of results might lead to stigmatization of patients and whole groups which would result in a barrier for the implementation of the technology into the mental health sectors. What comes to mind here is the emergence of phrenology in the 18th century.

5.2 Limitations

The lack of participants for the questionnaire was a problem in the study and led to a lack of expert evaluation regarding the questionnaire containing the evidence requirements. The questionnaire is however based upon a literature research and the POCKET and can be found in Appendix A. For future research, it might be of interest to validate these factors using expert feedback. In the following section, potential reasons for the lack of expert feedback will be elaborated.

First, there was no incentive given to the people who participated in the study. Especially in the health care setting, the most health-practitioners experience a high workload and have little time to spare. Because the research is part of a master thesis, the participants might simply not have experienced a high motivation to participate in a study that does not provide a form of perceived benefit for the. The lack of an incentive to participate in the study could potentially have increased the lack of a perceived benefit.

Another potential factor might be due to nature of the DLMSS. It is a new type of technology that is not yet implemented in practice and that combines different disciplines. Because of that, practitioners do not have hands-on expertise about the technology cannot talk from experience which asks for interference of their knowledge and experiences on the new technology. In the context of the healthcare setting, this might have been a potential barrier. Decisions that are made there result in direct consequences for the patient and most practitioners might be very careful to provide information on a topic in which they do not feel to have enough expertise. This was also apparent during the interviews. The interviewees were very careful to provide information about topics they did not feel to have enough expertise and verbalized these numerous times. If the emails that were send out received a response, the feedback also was directed towards the missing expertise on the specific topic. This aspect was also mentioned from the potential participants form the Kings College in London. What further complicated the subject was the fact that Philips did report to not yet have a network of mental health professionals who could be invited to participate in the questionnaire.

Due to the problems with finding participants, no sampling method was used except for the professional status of the participants. This method might also result in a potential bias during the interviews. Practitioners who participated might have felt an intrinsic motivation for the topic and might be more open to the general use of new technology. Additionally, all participants were German and were recruited from the social environment of the researcher. It is very important to see the results in the context of these aspects and as preliminary and qualitative work. Additionally, an overrepresentation of therapists was apparent in the sample which could potentially bias the results, especially in the context of what might be regarded as a potential benefit. Empirical validation of the results is needed, and it must be tested whether the results are generalizable to other fields of the mental health care setting like the field of psychiatry.

However, despite the small number of participants and even though the different professions were underrepresented in the sample, it was possible to include a radiologist, a psychiatrist, psychotherapists and neuropsychologists. Concerning the nature of the technology, this group allows for a relatively holistic evaluation of the technology because of their heterogenous expertise. Additionally, the research was guided by four experts in the field of human factors and health technology as well as health economy from the University of Twente and Philips Research in Eindhoven. Overall, this leads to heterogenous information synthesized from a multidisciplinary team of experts. In the context of the preliminary nature of the work, the participants of the interviews seem to be justified.

6 Conclusion

The presented work is a first step towards the alignment of DLMSS technology into the mental health care setting. The work presents a general outlook of future research directions and indicates that the integration of the technology must be tailored very specifically to the needs of the setting. A first overview about A.I. requirements for the field of mental disorders and about the potential that the technology might have for the new field of psychoradiology is provided. It appears that the general factors such as the creation of additional value for clinical practice can be translated towards DLMSS. However, the work also indicates that the use-case of the technology in the context of the mental health setting is somewhat different from the automation approach that is often aimed at in the field of medicine and that the needs of the mental healthcare setting indeed differ from the medical side. It is indicated that the diagnosis according to ICD or DSM categories generates less benefit and that more qualitative information about the patient and potential treatment decisions should be aimed at. The factor of explainability describes the need to translate this information into meaningful and understandable concepts for the practitioner which must indeed connect to an added benefit. More general, the technology should be used to create new information and not to make already existent practice more efficient. The work also indicates that the validation of the technology is not limited to official, institutional approval of it. In order to implement the technology into practice, a personal appraisal of it, termed ‘convergent validation’ is crucial and cannot be achieved through external validation in terms of studies or official approval. The practitioners must use the technology by themselves and it must produce results which converge with the personal perception of the practitioner.

6.1 Future Work

In the context of the research, there are different points that might be of interest to investigate further. The interviews delivered very broad information about higher-level requirements. More exploration is needed to discover more specific information and evidence that should be delivered for these higher-

level requirements, i.e.: ethical considerations. The questionnaire was an approach to validate these more specific requirements and should be evaluated with expert feedback. It might then be used as a first approach for more specific requirements.

A factor that emerged from the interviews indicated that an additional benefit must be created through the technology which must connect to the therapeutic process. Examples included the use as a treatment decision aid or pre-post measurements to create clinical endpoints to evaluate therapy. Future research might explore whether these points are generalizable and other ways to utilize the technology in order to create a benefit for the patient and the therapeutic process. In this regard, it might also be important to observe how useful the standard classification systems really are and to what extent the upcoming classification systems might be utilized since research about this topic appears to deliver somewhat incongruent results (Reed et al., 2019, First et al., 2018).

Another factor for future research includes the convergent validation of the technology through overlap with the perception of the practitioner. It might be of interest to see to what extent this overlap influences the appraisal of the technology. If the concept and the role of convergent validation for the subjective approval of the technology holds true, more research might be conducted to see to what extent the technology must overlap with the own perception and/or expectation and to what extent it might diverge from it in order to be judged positively by a practitioner. This factor also appears to overlap with research about trust and usage behavior in automation technology and it might be of interest to see if and how it fits in there.

The theme of “regulatory factors” might also be a point for future work. While it can be expected that practitioners have general knowledge about these kinds of factors, they do not represent this target group very well which decreases the validity of these aspects. Regulatory organizations like health insurance companies and federal agencies must be included in this process to make sure that evidence for the requirements that these organizations have for the technology is generated. At the moment, guidelines about requirements for A.I. technology are still in the making. Due to the mentioned ‘valley of death’ and the principle of enhanced clinical utility which is apparent both in the new ICD and the new DSM, this might be a good chance to include a multidisciplinary team of experts and real-life prototype experiments with the technology to create guidelines that allow to utilize the whole potential of the technology while also guaranteeing the safety of the patient. The results presented in this work might be used to inform future research concerning the creation of such guidelines.

Due to the novelty of the technology and the lack of experience that can be expected from most practitioners, it might be best to validate at least some requirements like the explainability aspect of the technology in form of prototypes and in a real-life setting.

6.2 Outlook

The results of this work indicate that the implementation of the technology towards being used in clinical practice still needs to overcome some obstacles. However, there are some important aspects that might indicate that the implementation of Deep Learning technology or even more sophisticated A.I. applications into the setting is only a matter of time. There are some important enablers for the technology non-specific to the medical or the mental health care setting. We live in a time of 'Big Data', meaning that nearly all information that exists is saved in some form of digital medium (Najafabadi et al., 2015). Additionally, the constant increase in the capabilities of computer hardware, a point which was one of the initial enablers of deep learning in general, pave the way for the use of the technology outside from highly advanced supercomputers. Another important factor to consider when thinking about the future of the technology in the mental health sector is the financial burden that is created by it (Olesen et al., 2012). In the year 2010, the total cost of brain disorders in Europe equaled 798 billion euros (Olesen et al., 2012). This cost is not only due to therapeutic measures but also due to indirect costs by a loss of working power (Olesen et al., 2012). This is an important incentive for the integration of measurements that might be able to reduce these costs. Because A.I. programs are basically a software program that can utilize already existing technology like brain scanners, the implementation of these into the setting might not result in adverse consequences regarding the costs (Goyen & Debatin, 2009). These aspects produce a positive background for the implementation of the technology into the field of mental healthcare.

Next to these influencers is the way the technology connects to the nature of our human way of thinking. The way we grasp our world is through the filter of our perception. A somewhat simplified elaboration of this is the concept of 'imaging a new color'. Since we do not possess the sensory structures to see outside of the wavelength of roughly 380 nm to 780 nm, we do not have any physical constitute to represent these. We can create gadgets like infrared optics in order to translate longer wavelengths to our perceptions but without changes on our sensory organs we will never be able to perceive them as they are. The realm of physics for example uses the same principle working mechanism and is nowadays more or less completely 'filtered' by formulas in order to understand and 'perceive' our reality. Our sensory organs can no longer provide us with enough information to understand it and therefore we mathematically deduce reality from the observations that we can make. Since the foundation of modern science is the measurement of observable and measurable endpoints, it asks for the translation of these constructs to our perception. Due to this, the creation of means which help us to observe and measure these endpoints can be regarded as very important drivers for all scientific disciplines.

Since psychology is a discipline from the healthcare sector, the need to generate measurable endpoints is very important both for the clinical and the scientific sector. We need these observable and 'objective' endpoints in order to create and test hypothesis and potential success of treatment

decisions. With the implementation of Deep Learning technology into the setting for the interpretation of brain imaging data, it might be possible to create new clinical endpoints that indeed allow us to create new theory or throw away old one by creating concepts that we can perceive and measure. Driven by the financial need to improve the mental healthcare setting and the age of digital data we live in, these points provide positive prospects for the integration and the success of the technology into mental healthcare, especially the scientific setting of psychology. These might then be translated towards clinical practice and used to create new ways of therapy and help to bridge the gap that appears to be apparent between science and practice.

Sources

Abi-Dargham, A., & Horga, G. (2016). The search for imaging biomarkers in psychiatric disorders. *Nature medicine*, *22*(11), 1248.

Arbabshirani, M. R., Plis, S., Sui, J., & Calhoun, V. D. (2017). Single subject prediction of brain disorders in neuroimaging: promises and pitfalls. *NeuroImage*, *145*, 137-165.

Bergsland, J., Elle, O. J., & Fosse, E. (2014). Barriers to medical device innovation. *Medical devices (Auckland, NZ)*, *7*, 205.

Beach, T. G. (2017). A review of biomarkers for neurodegenerative disease: will they swing us across the valley?. *Neurology and therapy*, *6*(1), 5-13.

Borsci, S., Uchegbu, I., Buckle, P., Ni, Z., Walne, S., & Hanna, G. B. (2018). Designing medical technology for resilience: integrating health economics and human factors approaches. *Expert review of medical devices*, *15*(1), 15-26.

Clark, L. A., Cuthbert, B., Lewis-Fernández, R., Narrow, W. E., & Reed, G. M. (2017). Three approaches to understanding and classifying mental disorder: ICD-11, DSM-5, and the National Institute of Mental Health's Research Domain Criteria (RDoC). *Psychological Science in the Public Interest*, *18*(2), 72-145.

Cresswell, K., & Sheikh, A. (2013). Organizational issues in the implementation and adoption of health information technology innovations: an interpretative review. *International journal of medical informatics*, *82*(5), e73-e86.

Crommelin, D. J., Broich, K., Holloway, C., Meesen, B., Lizrova Preiningerova, J., Prugnaud, J. L., & Silva-Lima, B. (2016). The regulator's perspective: How should new therapies and follow-on products for MS be clinically evaluated in the future?. *Multiple Sclerosis Journal*, *22*(2_suppl), 47-59.

Cuckler, G. A., Sisko, A. M., Poisal, J. A., Keehan, S. P., Smith, S. D., Madison, A. J., ... & Hardesty, J. C. (2018). National health expenditure projections, 2017–26: despite uncertainty, fundamentals primarily drive spending growth. *Health Affairs*, *37*(3), 482-492.

Donawa, M. E. (2011). European medical device usability requirements. *European Medical Device Technology*, *2*(6).

Davis, A. M. (1998). The harmony in rechoirments. IEEE Software, 15(2), 6–8

Drew, T., Evans, K., Võ, M. L. H., Jacobson, F. L., & Wolfe, J. M. (2013). Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images?. *Radiographics*, *33*(1), 263-274.

Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. *American journal of community psychology*, *41*(3-4), 327.

Eurostat. (2019). *Baseline projections: demographic balances and indicators.* Retrieved from http://appsso.eurostat.ec.europa.eu/nui/submitViewTableAction.do

Fan, W., Liu, J., Zhu, S., & Pardalos, P. M. (2018). Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS). *Annals of Operations Research*.

FDA, (2019). Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. Retrieved from: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device

First, M. B., Rebello, T. J., Keeley, J. W., Bhargava, R., Dai, Y., Kulygina, M., ... & Reed, G. M. (2018). Do mental health professionals use diagnostic classifications the way we think they do? A global survey. *World Psychiatry*, *17*(2), 187-195.

Fixsen, D. L., Naoom, S. F., Blase, K. A., & Friedman, R. M. (2005). Implementation research: a synthesis of the literature.

Fleuren, M. A. H., Paulussen, T. G. W. M., Van Dommelen, P., & Van Buuren, S. (2014). Measurement instrument for determinants of innovations (MIDI). *Leiden: TNO*.

Gillies, R. J., Kinahan, P. E., & Hricak, H. (2015). Radiomics: images are more than pictures, they are data. *Radiology*, *278*(2), 563-577.

Goyen, M., & Debatin, J. F. (2009). Healthcare costs for new technologies. *European journal of nuclear medicine and molecular imaging*, *36*(1), 139-143.

Hollis, C., Sampson, S., Simons, L., Davies, E. B., Churchill, R., Betton, V., ... & Kabir, T. (2018). Identifying research priorities for digital technology in mental health care: results of the James Lind Alliance Priority Setting Partnership. *The Lancet Psychiatry*, *5*(10), 845-854.

Hampel, H., Lista, S., & Khachaturian, Z. S. (2012). Development of biomarkers to chart all Alzheimer's disease stages: the royal road to cutting the therapeutic Gordian Knot. *Alzheimer's & Dementia*, *8*(4), 312-336.

Huddy, J. R., Ni, M., Misra, S., Mavroveli, S., Barlow, J., & Hanna, G. B. (2018). Development of the Point-of-Care Key Evidence Tool (POCKET): a checklist for multi-dimensional evidence generation in point-of-care tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*.

Isaac, M., & Gispen-de Wied, C. (2015). CNS biomarkers: Potential from a regulatory perspective: Case study–Focus in low hippocampus volume as a biomarker measured by MRI. *European Neuropsychopharmacology*, *25*(7), 1003-1009.

Joint Report prepared by the European Comission and the Economic Policy Committee. (2015). *The 2015 Ageing Report: Underlying Assumptions and Projection Methodologies.*

Jack, A. (2018). Neuroimaging in neurodevelopmental disorders: focus on resting-state fMRI analysis of intrinsic functional brain connectivity. *Current opinion in neurology*, *31*(2), 140-148.

Kahneman, D., & Egan, P. (2011). *Thinking, fast and slow* (Vol. 1). New York: Farrar, Straus and Giroux.

Kahneman, D., Slovic, S. P., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. Cambridge university press.

Karsh, B. T. (2004). Beyond usability: designing effective technology implementation systems to promote patient safety. *BMJ Quality & Safety*, *13*(5), 38

Kip, M. M. A., IJzerman, M. J., Henriksson, M., Merlin, T., Weinstein, M. C., Phelps, C. E., … Koffijberg, H. (2018). Toward Alignment in the Reporting of Economic Evaluations of Diagnostic Tests and Biomarkers: The AGREEDT Checklist. *Medical Decision Making*, *38*(7), 778–788. https://doi.org/10.1177/0272989X18797590

Klein, G. (2008). Naturalistic decision making. *Human factors*, *50*(3), 456-460.

Lindsay, R. K., Buchanan, B. G., Feigenbaum, E. A., & Lederberg, J. (1980). Applications of artificial intelligence for organic chemistry: the DENDRAL project. *New York*.

Lui, S., Zhou, X. J., Sweeney, J. A., & Gong, Q. (2016). Psychoradiology: the frontier of neuroimaging in psychiatry. *Radiology*, *281*(2), 357-372.

Lurie, N. H., & Mason, C. H. (2007). Visual representation: Implications for decision making. *Journal of marketing*, *71*(1), 160-177.

Nichols, L. M. (2002). Can defined contribution health insurance reduce cost growth?. *EBRI issue brief*, (246).

Norman, G. (2005). Research in clinical reasoning: past history and current trends. *Medical education*, *39*(4), 418-427.

McBee, M. P., Awan, O. A., Colucci, A. T., Ghobadi, C. W., Kadom, N., Kansagra, A. P., ... & Auffermann, W. F. (2018). Deep learning in radiology. *Academic radiology*, *25*(11), 1472-1480.

Mcknight, D. H. (2005). Trust in information technology. The Blackwell encyclopedia of management. Vol. 7 management information systems. Malden: Blackwell Publications.

Medeiros, F. A. (2017). Biomarkers and surrogate endpoints: lessons learned from glaucoma. *Investigative ophthalmology & visual science*, *58*(6), BIO20-BIO26.

Muszyńska, M. M., & Rau, R. (2012). The old-age healthy dependency ratio in Europe. *Journal of population ageing*, *5*(3), 151-162.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

O'connor, J. P., Aboagye, E. O., Adams, J. E., Aerts, H. J., Barrington, S. F., Beer, A. J., ... & Buckley, D. L. (2017). Imaging biomarker roadmap for cancer studies. *Nature reviews Clinical oncology*, *14*(3), 169.

Olesen, J., Gustavsson, A., Svensson, M., Wittchen, H. U., Jönsson, B., CDBE2010 Study Group, & European Brain Council. (2012). The economic cost of brain disorders in Europe. *European journal of neurology*, *19*(1), 155-162.

Patel, V. L., Shortliffe, E. H., Stefanelli, M., Szolovits, P., Berthold, M. R., Bellazzi, R., & Abu-Hanna, A. (2009). The coming of age of artificial intelligence in medicine. *Artificial intelligence in medicine*, *46*(1), 5-17.

Petkovic, D., Altman, R., Wong, M., & Vigil, A. (2018). Improving the explainability of Random Forest classifier–user centered approach. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (Vol. 23, pp. 204-215). NIH Public Access.

Pesapane, F., Volonté, C., Codari, M., & Sardanelli, F. (2018). Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights into imaging*, 1-9.

Prins, N. (2016). *Psychophysics: a practical introduction*. Academic Press.

Hampel, H., Lista, S., & Khachaturian, Z. S. (2012). Development of biomarkers to chart all Alzheimer's disease stages: the royal road to cutting the therapeutic Gordian Knot. *Alzheimer's & Dementia*, *8*(4), 312-336.

Huddy, J. R., Ni, M., Misra, S., Mavroveli, S., Barlow, J., & Hanna, G. B. (2018). Development of the Point-of-Care Key Evidence Tool (POCKET): a checklist for multi-dimensional evidence generation in point-of-care tests. *Clinical Chemistry and Laboratory Medicine (CCLM)*. Randall, M. (2017). *Overview of the UK population: July 2017*. Office for National Statistics. 1–17.

Reed, G. M., First, M. B., Kogan, C. S., Hyman, S. E., Gureje, O., Gaebel, W., ... & Claudino, A. (2019). Innovations and changes in the ICD-11 classification of mental, behavioural and neurodevelopmental disorders. *World Psychiatry*, *18*(1), 3-19.

Roberts, E. B. (1988). Technological innovation and medical devices. *New medical devices: invention, development, and use*, 35-51.

Roberts, S. F., Fischhoff, M. A., Sakowski, S. A., & Feldman, E. L. (2012). Perspective: Transforming science into medicine: How clinician-scientists can build bridges across research's "valley of Death." *Academic Medicine*, *87*(3), 266–270.

Rogers, E. M. (1995). Diffusion of Innovations (4th Eds.) ACM The Free Press (Sept. 2001). *New York*, 15-23.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Scher, H. I., Morris, M. J., Larson, S., & Heller, G. (2013). Validation and clinical utility of prostate cancer biomarkers. *Nature reviews Clinical oncology*, *10*(4), 225.

Shah, S. G. S., & Robinson, I. (2007). Benefits of and barriers to involving users in medical device technology development and evaluation. *International journal of technology assessment in health care*, *23*(1), 131-137.

Si, B., Yakushev, I., & Li, J. (2017). A sequential tree-based classifier for personalized biomarker testing of Alzheimer's disease risk. *IISE Transactions on Healthcare Systems Engineering*, *7*(4), 248-260.

van Ginneken, B., Schaefer-Prokop, C. M., & Prokop, M. (2011). Computer-aided diagnosis: how to move from the laboratory to the clinic. *Radiology*, *261*(3), 719-732.

Wiegers, K. E. (2003). Software requirements (2nd ed.). Redmond, WA: Microsoft Press.

Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in cognitive sciences*, *15*(2), 77-84.

Woods, D. D., & Hollnagel, E. (2006). *Joint cognitive systems: Patterns in cognitive systems engineering*. CRC Press.

Woolf, S. H. (2008). The meaning of translational research and why it matters. *Jama*, *299*(2), 211-213.

World Health Organization. (2010). *Clinical evidence for medical devices: regulatory processes focussing on Europe and the United States of America: background paper 3, August 2010*(No. WHO/HSS/EHT/DIM/10.3). Geneva: World Health Organization.

World Health Organization. (2010). *Medical devices: managing the mismatch: an outcome of the priority medical devices project*.

World Health Organization. (2018). *Total Health expenditure as % of GDP*. Retrieved from https://gateway.euro.who.int/en/indicators/hfa_566-6711-total-health-expenditure-as-of-gdp/

Zaharchuk, G., Gong, E., Wintermark, M., Rubin, D., & Langlotz, C. P. (2018). Deep learning in neuroradiology. *American Journal of Neuroradiology*, *39*(10), 1776-1784.

Appendix A - In Silicio Evidence Tool (ISET)

Factors	Sub-Factors	Evidence Requirements
Medical Pathway	Description of new Pathway	Test Indication and Function (Eg.: diagnosis/risk prediction/monitoring)
		Clinical need for test.
		Description of indicated population.
		Description of the intended user
		Rationale for the strategy.
		Description of how the clinical pathway will change by incorporating the DLMSS.
		Stakeholders of DLMSS (i.e., people affected by the clinical use of this tool
	Financial consequences of new technology	Evaluation that compares costs before and after introduction of a test to a clinical pathway.
		Ability to incorporate local population data into economic model/analysis.
		Indication of costs per test
		Advantage in terms of cost per quality adjusted life year (QALY)
		Cost of test Including: cost of device, cost of extra, equipment needed to perform, any other costs.
	Quality of work	Indication of subjective well-being of medical staff (e.g., measurement of stress level).

		Information of potential change in face-to-face time with patients.
		Advantages and disadvantages of new test pathway at an institutional or regional level.
		Indication of working hours by staff.
	Quality of care	**Indication of change in subjective well-being of patient (E.g.: health-related quality of life)**
		Consequences of the test result to patient (e.g., potential treatment and its impact on the patient)
		Consequences of incorrect test result to patient.
		Advantages and disadvantages of new test pathway to the patient.
		Potential risks of test procedure to patient.
Regulations	**Regulations**	**Provision of official guidelines for the use of the technology.**
		Description of Data Sharing Procedure.
		Regulatory approval obtained (e.g., FDA, NICE)
		Indication of how data security is achieved.
		Indication of how data anonymity is guaranteed.
Technical Description of	**Technical Description of Test**	**Associated equipment required to perform test (device/computer hardware/other consumables/power**

Test		**source/batteries).**
		Detailed description of the test process including image acquisition protocols.
		Description of the image modality being used (e.g., MRI, PET.).
		Description of Turnaround time for a single test.
		Description of how the results are presented to user.
		IT system interoperability.
Usability	**Usability**	**Suggested standard operating procedure for test device and process**
		Identification of operator dependent steps
		Training requirements for undertaking procedure
		Training requirements for using the test device
		Training requirements to interpret the results
		Suggested method(s) for competency assessment
		Test device calibration procedure and internal quality control protocol including level of expertise required to perform
		Test device maintenance required and level of expertise necessary to perform
		Description of robustness of test device

		Provision of support infrastructure with the device (E.g.: service, agreements, helpline, website)
	Explainability	**Description of Risk-Factors that are being used by the algorithm.**
		Indication of Features that contribute to the decision.
		Direction of the Features that contribute to the decision.
		Potential interaction of features that contribute to the decision.
		Indication of Cut-Off Score from the features.
		Loss of accuracy if some features are not used.
		A visual presentation of the decision function.
		Indication of percentile in distribution (e.g., 86th percentile in the distribution of patients with schizophrenia.
Validity of Biomarker	**Statistical Measurements**	**Sensitivity and specificity of test device in an optimized or laboratory setting (Sensitivity – proportion of people with disease who have a positive test result; specificity - proportion of people without disease who have a negative test result).**
		Sensitivity and specificity of test device in a real world or clinical setting.
		Negative and positive predictive value of test results.

		Receiver Operating Curve (ROC) and Area Under Curve (AUC) analysis of continuous diagnostic test results.
	Biological Validation	**Test Performance in Multisite studies (Reproducibility).**
		Test Performance in Longitudinal Studies with the same subject (Repeatability).
		Assessment of Potential Biases in the training data.
		Assessment of other factors that can influence the performance of a Biomarker (E.g.: other pathologies).
		Definition of Right outcome for training and test sample (How was the right outcome (ill/not ill) defined, especially in case no gold standard is available).
		Correlation of Biomarker and clinical outcome variable (E.g.: symptoms, treatment response, etc.).
		Connection of Biomarker to existing theory (E.g.: Features used by the Biomarker can also be found in literature).
		Indication of the Standard Variation for a Biomarker.
		Correlation of new test results with results of already existing tests.
	Clinical Validation of Biomarker	**Results of clinical trials.**
		Evidence that test results provide additional information to what is

		available at the time.
		Linked evidence approach (the synthesis of acquired evidence on test accuracy, impact of decision making and effectiveness of consequent treatment to evaluate overall test effectiveness.)
		Detailed information of the data set (number of tests performed, incidence/prevalence of disease or outcome, etc.).
		List of relevant publications of clinical trials.
		Systematic review/meta-analysis of clinical trials.
		Timeline of any modification to the device since the evidence is obtained and justification that evidence remains reliable.
		Local pilot or case study.

Appendix B – Interview schedule

Data (Sorted by Subject):

1. Audio Data
2. Informed Consent
3. Summary of interview topics

Background Information:

1. Age
2. Professional Status
3. Years of Experience
4. Area of Specialty (If applicable)

Interview:

Introduction

Summary of the interview Process.

Explanation of the technology and the goal of the interview.

Open Questions

If you would have to decide whether to implement and use the technology, what kind of information would you like to receive?

1. What do you think about the application of the technology in the field of mental disorders in general?

2. What are the criteria that the technology must fulfill before you would deem it applicable for use (facilitators)?

3. Do you have specific concerns regarding this type of technology / Do you see potential barriers?

Evaluation of Literature research – Semi Structured

Facilitators

Added Benefit of Technology

1. Quality of Work (How could the technology enhance the quality of your work? What evidence do you want?)
    a. Prompts:
        i. Decreased Turnaround time for tests, Creation of additional knowledge, Increase of Face-to-Face contact

2. Quality of Care (How could the technology enhance the quality of care for the patient? What evidence do you want?)
    a. Prompts:
        i. Increase of Subjective wellbeing, Consequences of test result to patient (effectiveness of potential treatments, etc.)

3. Financial Aspects (If important, what evidence should be provided for a financial benefit?)
    a. Prompts:
        i. Costs per test, Comparison before/after the implementation of the technology

4. Are there other potential benefits the technology could bring to your work?

Barriers

Validity

Training Data

1. Do you have concerns about the gold standard for the diagnosis of the training sample? What information would you like to have in this regard?
    a. Prompts:

i. Use of existing test batteries, multiple assessments by professionals

2. Would you like to have information on the training sample? What information?
    a. Prompts:
        i. Number of participants, Characteristics of participants, Assessment of potential biases

Validation

What evidence would you want to have?

1. Technical Validation.
    a. What kind of information would you like to receive concerning the neurological correlate of the biomarkers being used?
        i. Prompts:
            - Multisite studies (Reproducibility), Longitudinal studies (Repeatability), Correlation of Biomarker and outcome variable.

2. Test Performance.
    a. What kind of information would you like to receive about the statistical performance of the test?
        i. Prompts:
            - Sensitivity, Specificity, Accuracy, External Validation.
3. Clinical Validation.
    a. What kind of information would you like to receive concerning the medical trials?
        i. Prompts:
            - Meta-Review of Clinical Trials, Consequences of incorrect test results, additional information for clinical decision making.

Explainability

Would you like to receive more information about deep learning and how it works before using the technology? (Model Explainability). What type of information?

When working with the actual device, what kind of information should it provide to you in order to inform your decision making?

Prompts:

- Risk Criteria.
- Decision Features.
- Probabilities.

Regulations

1. Do you have concerns about the confidentiality and security of the patient data?

2. Do you have concerns about the accountability of the actions that are taken on basis of the technology?

Subjective Factors

1. Do you have specific concerns about patient safety for this type of technology?
2. Do you have concerns that the technology can change the provider/patient relationship?
3. Do you have concerns relating to your job and a potential loss due to the technology?
4. Do you have any other concerns about the implementation of the technology into this setting?

End of the interview.

1. Short rehearsal of agreed consent and provision of contact details in case additional ideas might come after the interview or the person might know someone else who could be interested or anything else requiring contact.
2. Ending the interview and giving thanks to the participant.

Appendix C – Literature Table

Reference	Title	General content	Translatable Factors and Potential Evidence Requirements	Comments/Conclusion
Arbabshirani, Plis, Sui & Calhoun, (2017)	Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls.	The source deals with the use of brain imaging for single subject prediction and gives a general overview about important factors. The most factors relate to aspects of the technology and not requirements (e.g.: overfitting)	Data anonymity, differential diagnosis, heterogeneity of mental disorders, generalizability towards single subjects.	Gives a comprehensive overview about the use of neuroimaging techniques for mental disorders.

Huddy, Ni, Misra, Mavroveli, Barlow & Hanna, 2018	Development of the Point-of-Care Key Evidence Tool (POCKET): a checklist for multi-dimensional evidence generation in point-of-care tests	Opinion of different groups of stakeholders about what should be integrated in the evaluation of a point-of-care devices in order to use the technology.	Technical Description of Test, Clinical Pathway, Stakeholders, Economic Evidence, Test Performance, Usability and Training, Clinical Trials	The list is created in the context of Point-Of-Care Devices. However, most factors appear to be generalizable. There are 65 specific evidence requirements in the paper.
Groen, 2011	Variance in User-Requirements	The source deals with the origin of variance in user-requirements.	Provision of Information, Safety, Finances, Juridical aspects, Usability	
Gallego, Casey, Norman & Doodall, 2011	Introduction and uptake of new medical technologies in the Australian health care system: A qualitative study	Expert opinions on an already existing medical evaluation process.	Transparency, Interactivity, Fairness	The requirements are not based on medical devices per se, but on the evaluation process of them. Connects loosely to evidence requirements.

Ludwick & Doucette, 2009	Adopting electronic medical records in primary care: Lessons learned from health information systems implementation experience in seven countries.	Different Factors influencing the implementation and adoption of health technology systems	Privacy, Patient Safety, Provider/Patient Relations, Staff Anxiety, Time Factors, Quality of Care, Finances, Efficiency, Liability	Literature based risk factors for implementation.
Hengstler, Enkel & Duelli, 2016	Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices	Description of factors influencing the adaption of these systems.	Predictability, Dependability, Faith, Trust in the innovating firm, Communication	The factors are based on literature and interviews with companies selling A.I. technology. The factor of trust is based mainly on automation literature.

Petkovic, Altman, Wong & Vigil, 2018	Improving the explainability of Random Forest classifier–user centered approach	How to enhance the explainability of Classifiers	Model Explainability, Sample Explainability, Features contributing to decsions (E.g.: MRI voxels), Direction of Features, Interaction of Features, Presentation of analysis, Loss of accuracy in case not all features are used.	Model explainability refers to the way why and how the algorithm works while sample explainability connects to how the algorithm made a specific decision for a specific sample.
Fu & Costafreda, 2013	Neuroimaging-Based Biomarkers in Psychiatry: Clinical Opportunities of a Paradigm Shift Cynthia	Prognosis and Diagnosis of different psychiatric disorders using neuroimaging biomarkers.	Diagnostic Uncertainty, Creation of Biomarkers in the absence of diagnostic gold-standard.	The article focusses more on the technical aspects and indicates the importance of understanding the reasoning process. It connects to Explainability, Predictability and Transparency: How can the expert trust the biomarker

				to be valid?
Kip et al., 2018	Toward Alignment in the Reporting of Economic Evaluations of Diagnostic Tests and Biomarkers: The AGREEDT Checklist	The article provides a checklist of items that should be integrated and tested when conducting a health economic evaluation of tests and biomarkers.	Evaluation of Tests and Biomarkers, Use of Diagnostic Tests, Test Performance and Characteristics, Patient Management Decisions, Impact on Health outcomes and costs, Social Impact	The items should be part of a health economic evaluation process, for example of a new device. It is based on literature and expert opinions and is aimed at the transparency and completeness of evaluations for tests and Biomarkers. It contains 43 items.

Greenhalgh et al., 2017	Beyond Adoption:A New Framework for Theorizing and Evaluating Nonadoption, Abandonment, and Challenges to the Scale-Up, Spread, and Sustainability of Health and Care Technologies	Provision of a general framework for the integration of Health Technologies.	Condition of Illness, Nature of Technology, Value Proposition of Technology, Adopters of the Technology, Organizational Context, Wider system Context	The source deals with general factors that influence the implementation of Health Technology which indirectly relates to requirements. 24 factors in the source.
WHO Background Paper 6, 2010	Barriers to innovation in the field of medical devices.	Provides a general overview about this topic.	Costs, Regulations, Fit of technology (Pathways), Emotional and subjective factors of personell.	The Exploration of subjective factors should be integrated in the interviews.
Park, Chang & Nam, 2018	A Bayesian Network Model for predicting post-stroke outcomes with available Risk Factors	Describes machine learning models to predict post-stroke outcomes with risk-factors	Interpretability	Understanding the reason behind the action of the classifier appears crucial → connects to explainability

Keenan et al., 2018	Quantitative Magnetic Resonance Imaging Phantoms: A Review and the Need for a System Phantom	Enhancing the validity of imaging biomarkers and creating quantitative MRI biomarkers.	Analysis of measurements across systems and longitudinal studies. creation of a system phantom (standard reference structure) and standardized protocols.	Connects to the problem of validity without existence of a gold standard; By enabling the comparison of different MRI studies and Biomarkers, these can be quantified and validated. For this, standardization is needed.

O'Connor et al., 2017	Imaging biomarker roadmap for cancer studies	Describes different steps that are required to qualify an imaging biomarker as suitable for cancer.	Technical Validation: Repeatability and Reproducibility, Sensitivity and Specificity, Assessment of Potential Biases (Reference Phantom), Feasibility (E.g.: Time, Setting, etc.), Safety, Toleration of treatment, Regulatory and Ethical Approval. Biological Validation: Biomarker - Biology - Outcome Variable - Value in decision making (E.g.: Treatment). Cost-Effectiveness: Advantage in terms of cost per quality adjusted life year (QALY) Qualification: Regulatory qualification, Linking Biomarker to biological processes.	The Biomarker connects to cancer. It indicates that it is essential to connect the Biomarker to analyzable, observable and known information

Scher, Morris, Larson & Heller, 2013	Validation and clinical utility of prostate cancer biomarkers	Overview of how to create and verify Biomarkers for prostate cancer.	Analytical Validation: Data about Device, The imaging modality, Conditions for reproducible and accurate results. Clinical Validation: Results can inform medical decision making, Results provide additional information to what is available at the moment.	Copes with criteria of the Quantitative Imaging Biomarker Alliance (QIBA)
Medeiros, F. A. 2017.	Biomarkers and Surrogate Endpoints: Lessons Learned from Glaucoma	Investigates how Biomarkers can be validated as surrogate endpoints for medicine in the context of glaucoma.	Validation of Biomarker as surrogate endpoint.	The source deals specifically with surrogate endpoints for glaucoma treatment. It is important to connect the biomarker to already existing knowledge or measurable endpoints to verify them.

Algarni & Stoessl, 2016.	The role of biomarkers and imaging in Parkinson's disease	Explains the potential role of imaging biomarkers for the Parkinson disease.	Validation of Biomarker, Combining Different Biomarkers into a test battery.	Because there is only sparse theoretical validation for the validity of the new Biomarkers, combining it with already known tests and assessments might be a good way to help people trust in the Biomarker.
Pesapane, Codari & Sardanelly, 2018	Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine	A general estimation about the role of A.I. and important factors for its use in this context.	Independent validation (E.g.: External institute), Standardized acquisition protocols, Anxiety on side of Radiologists, Defined Evaluation Criteria & Reporting Guidelines.	In order to enable the use of A.I. in medical imaging, certain requirements have to be met, the most important one being again the validation of them and processes leading to the clinical validation.

Crommelin et al., (2016)	The regulator's perspective: How should new therapies and follow-on products for MS be clinically evaluated in the future?	Copes with different topics relating to MS and especially the creation of new therapies and treatments based on imaging biomarkers.	Protocol for Standardization, use of clinical endpoints: Change of Feature connects to Change in outcome.	The article copes with the standpoint of regulatory institutions. Standardization is a potential requirement for multisite studies and not for the technology per se.
Van Ginneken, Schaefer-Prokop & Prokop, 2011.	Computer-aided Diagnosis: How to Move from the Laboratory to the Clinic	A general outlook on computer aided diagnosis and associated requirements based on 2011. Does not cope with A.I. and specific problems.	Sufficient Performance (E.g.: Sensibility and Specificity), No increase in reading times, Integration in Workflow (Pathway), Regulatory approval, Cost efficiency, Large- and high-quality databases	Very general requirements.

Hampel, Lista & Khachaturian, 2012	Development of biomarkers to chart all Alzheimer's disease stages: The royal road to cutting the therapeutic Gordian Knot	Important Factors in the creation and validation of Biomarkers specifically for Alzheimer	Analytical: Outcome studies, quality assessment, Accuracy, Repeatability, Reproducibility, Other Factors influencing the detection of Biomarker, Biological Variation within and between subjects. Clinical Validation: Phase I-IV human trials,	The source deals specifically with Alzheimer.

Beach, 2017	A Review of Biomarkers for Neurodegenerative Disease: Will They Swing Us Across the Valley?	A general outlook about the topic; Potential advantages and ways to implement new Biomarkers.	Correlation with clinical endpoints and qualification as surrogate endpoints. Neuropathological Validation. Overcoming Inaccurate Diagnostic gold standards with huge sample sizes.	Validation of the Biomarker is only one part of the study.
Isaac & Gispen-de Wied, 2015	CNS biomarkers: Potential from a regulatory perspective: Case study – Focus in low hippocampus volume as a biomarker measured by MRI.	Indication of different steps needed for the qualification as a Biomarker.	Qualitative and Quantitative Evaluation approach (Qualitative analysis of potential Biomarker and Quantitative analysis about the functioning of theses Biomarkers [ROC curve, etc.])	Hippocampus atrophy is a prominent finding in AD patients. How is it possible to achieve validation if findings are completely based on algorithm (no connection to known biology)?

Si, Yakushev & Li, 2017	A sequential tree-based classifier for personalized biomarker testing of Alzheimer's disease risk.	Development of a testing procedure to use Biomarkers and other risk criteria in practice.	The article creates a decision tree for different biomarkers found in literature. The decision tree indicates the cut-off-scores for the biomarkers and combines it with other risk factors. Connects to explainability.	The article fits only partly with the inclusion criteria. However, the general idea of combining different criteria and creation of a decision tree seems promising for the factor of explainability.

Obuchowski et al., 2016	Statistical Issues in Testing Conformance with the Quantitative Imaging Biomarker Alliance (QIBA) Profile Claims	The article copes with certain issues that exist in the profile claims of the quantitative biomarker alliance.	Clinical Context, Technical Performance Claims, list of actors (Device, Software, Person), technical and performance requirements for the actors (activities), summary of scientific studies supporting performance claim, procedures to test conformity with technical and performance requirements	It appears that multicenter studies are essential in the validation of biomarkers and that standardization is a necessary step to enable this.

Pesapane, Volonté, Codari & Sardanelli, 2018	Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States	A general estimate about A.I. in medicine focused on ethical and legal aspects.	Ethics, Regulations, Accountability, Data protection	Illuminates the more subjective factors that might influence the adoption of A.I. technology into the field of medicine.
Lambin et al., 2017	Radiomics: the bridge between medical imaging and personalized medicine	Describes necessary step for the creation of new biomarkers and proposes a model for this.	Combining different features, Correlation of Biomarker features with clinical endpoints, creation of robust biomarkers, calibration and discrimination, Statistical consistency between training and validation set, Independent replication of results (multisite studies)	

Thrall, Li, Li, Cruz, Do, Dreyer & Brink, 2018	Artificial Intelligence and Machine Learning in Radiology: Opportunities, Challenges, Pitfalls, and Criteria for Success	Factors potentially Influencing implementation of A.I. in Radiology	<u>The technology must create an additional value for clinical practice:</u> Increased Diagnostic Certainty, Faster Turnaround Time, Better outcomes for patients, better quality of work life, algorithm being tolerant of different data acquisition protocols and it should work in diverse patient populations.

Miller & Brown, 2018	Artificial Intelligence in Medical Practice: The Question to the Answer?	General Overview and Potential Jeopardies of A.I. technology in medical practice.	Technological Barrier to Patient Care? Potential Non-Medical Barriers in direct patient care (Subjective). Loss of physician's work or reduction of value (could be implemented as evidence requirement in pathway).	The paper gives a general overview and then formulates 'Potential Jeopardies'. These are not empiric but result from the inherent logic of the paper.

Fan, Liu, Zhu & Pardalos, 2018	Investigating the impacting factors for the healthcare professionals to adopt artificial intelligence-based medical diagnosis support system (AIMDSS)	Factors that influence the intention to adopt this sort of technology. Based on Technology Acceptance Models.	Initial Trust: Propensity to Trust, Performance Expactancy, Social Influence, Effort expactancy. Performance Expactancy: Task Complexity and Technology Characteristics. Technology Characteristics: Diagnosis Capacity, Interpretability of results and interoperability between systems.	Trust in the technology is considered the key factors (initial trust and how to influence that). = Detailed description of decision and leading role during of physician during assistant diagnosis.

Thessen, 2016	Adoption of Machine Learning Techniques in Ecology and Earth Science	A very broad outlook about A.I. in the context of Ecology. Potential Barriers and Facilitators.	User Friendly Interfaces, Training Data, Lack of Tools and Service (Barrier)	The training data set is of very high importance, especially considering generalizability).
Lui, Zhou, Sweeney & Gong, 2016.	Psychoradiology: The Frontier of Neuroimaging in Psychiatry	Provides a general overview about the use of quantitative imaging biomarkers in the field of mental disorders	Use of standard imaging protocols, reduction of potential confounding variables on the biomarker.	The article provides fewer specific requirements but a very detailed framework of how imaging based biomarkers might revolutionize the mental healthcare setting.

Patel, Shortliffe, Stefanelli, Szolovits, Berthold, Belazzi, Abu-Hanna, 2008	The coming of age of artificial intelligence in medicine	Integration into workflow, Alignment with cognitive nature of practitioners, Data confidentiality, Creation of benefit for medical practice	Provides a very early outlook into the field of A.I. implementation into medicine. Less specific evidence requirements but comprehensive explanation about general aspects of A.I. and how it relates to the field of medicine.
Zaharchuk, Gong, Wintermark, Rubin & Langlotz, (2018).	Deep learning in neuroradiology	Image acquisition protocols, prevention of overfitting the model, exposure to the technology and peer approval, Transparency of predictive model.	The article provides a general outlook into the application of deep learning techniques into neuroradiology. It copes mainly with technical aspects, but it is possible to deduce potential requirements for the mental healthcare setting.