UNIVERSIDADE CATÓLICA PORTUGUESA

INVESTIGATING COGNITIVE NETWORK SIMILARITY IN BREAST CANCER DETECTION

Dissertation submitted to Universidade Católica Portuguesa to obtain a Master's Degree in Psychology in Business and Economics

By
Julian Berger

Faculty of Human Sciences

September 2021

UNIVERSIDADE CATÓLICA PORTUGUESA

INVESTIGATING COGNITIVE NETWORK SIMILARITY IN BREAST CANCER DETECTION

Dissertation submitted to Universidade Católica Portuguesa to obtain a Master's Degree in Psychology in Business and Economics

By
Julian Berger

Faculty of Human Sciences
Under the supervision of Professor Rui Gaspar

September 2021

Abstract

Identifying predictors of collective performance in medical decision-making, requiring diagnosticians to independently formulate judgements, is of key importance for effective care. The German Mammography Screening Program poses a prime example of such a situation in which diagnostic decisions are currently being made by at least two individuals. Given this importance, the present study combined insights from research in mental models and cognitive network science to study the cognition and diagnostic performance of experienced radiologists in their ability to correctly identify cancer presence in mammogram images. The study relies on a mixed-methods design, incorporating knowledge gained from interviewing experienced radiologists into a subsequent cross-sectional investigation on cognitive networks of cancer cues relevant in mammogram diagnoses. Cognitive networks were elicited from trained radiologists employed in the national screening program and compared based on their structure, path lengths, degree and clustering. Divergences between radiologists' networks were revealed through both visual and numerical analyses. However, results of generalized linear mixed modeling indicated divergences and similarities to be almost unequivocally as not being associated with both diagnostic performance and diagnostic similarity in dyadic decision making. The key finding of the present study suggests, however, that more precisely defined clusters within networks, represented via low clustering coefficient, were associated with correct classification of images to diagnostic categories, which warrants future research opportunities. The study concludes with the identification of three limitations present both in this inquiry as well as in prior research and calls for a renewed critical assessment of fundamental assumptions underlying current cognitive network studies.

Abstract

Identificar preditores de desempenho coletivo na tomada de decisão médica, que exigem que os diagnosticadores formulem julgamentos de forma independente, é de fundamental importância para um cuidado eficaz. O Programa Alemão de Rastreio por Mamografia representa um excelente exemplo de tal situação em que as decisões diagnósticas são atualmente feitas por, pelo menos, dois indivíduos. Dada a sua importância, o presente estudo combinou conclusões suscitadas por investigação sobre modelos mentais e sobre redes neuronais nas ciências cognitivas, para estudar a cognição e o desempenho diagnóstico de radiologistas experientes, na sua capacidade de identificar corretamente a presença de cancro em imagens de mamografia. O estudo baseia-se num desenho de métodos mistos, incorporando o conhecimento obtido a partir de entrevistas com radiologistas experientes, numa investigação transversal subsequente em redes cognitivas referentes a sinais de cancro relevantes em diagnósticos de mamografia. Redes cognitivas foram elicitadas em radiologistas treinados empregados no programa nacional de rastreio e comparadas, com base na sua estrutura, comprimentos da ligação, grau e agrupamento. Divergências entre as redes dos radiologistas foram reveladas por meio de análises visuais e numéricas. No entanto, os resultados de modelos lineares generalizados mistos indicaram quase inequivocamente que as divergências e semelhanças não estavam associadas ao desempenho diagnóstico e à semelhança diagnóstica na tomada de decisão diádica. O resultado central do presente estudo sugere, no entanto, que clusters mais precisamente definidos dentro de redes, representados por meio de baixo coeficiente de agrupamento, foram associados à classificação correta de imagens em categorias diagnósticas, o que apresenta futuras oportunidades de pesquisa. O estudo conclui com a identificação de três limitações presentes tanto nesta investigação quanto em investigações anteriores e apela a uma avaliação crítica renovada dos pressupostos fundamentais subjacentes aos estudos atuais de redes cognitivas.

Contents

1 Introduction

Human cooperation is a stable organizational feature in today's world. In organizations, people work in teams and in democracies, people come together and vote in elections. Usually, one expects that having more heads thinking and working on the same task increases performance and the likelihood of achieving success. Indeed, the phenomenon of collective intelligence or the so-called wisdom of the crowd can be observed in many domains such as economics (Arrow et al., 2008), political forecasting (Tetlock & Gardner, 2016; Tetlock, Mellers, Rohrbaugh, & Chen, 2014) and medicine (Kämmer, Hautz, Herzog, Kunina-Habenicht, & Kurvers, 2017; Kurvers, De Zoete, Bachman, Algra, & Ostelo, 2018; Kurvers, Krause, Argenziano, Zalaudek, & Wolf, 2015). Usually, the benefit of asking many people over asking one randomly chosen individual is explained by the diversity of judgement errors within a group (Davis-Stober, Budescu, Dana, & Broomell, 2014; Larrick & Soll, 2006). For example, when asking two people, one over- and one underconfident, to assess the future return on a past investment, the average of both is likely to be closer to the truth than just one of the singular estimates. However, in many real world situations, the wisdom of the crowd can be improved upon by choosing not to aggregate the intelligence of many but to aggregate between experts. In fact, Kurvers et al. (2019) and Prelec, Seung, and McCoy (2017) showed that even higher performance can be achieved if one pre-selects people based on how similar they behave, as it may signal expertise. The reasoning is that expertise not only leads to accuracy of the individual but to similar behaviors between individuals, because they reason in the same way.

Understanding expertise and selecting those who have the highest chance of cooperative success is paramount for organizational performance. One path of inquiry into cooperative success is offered by the research area of mental models, which has been previously linked to group performance (e.g. Edwards, Day, Arthur Jr, & Bell, 2006; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000; Mohammed, Ferzandi, & Hamilton, 2010). In the present study, several intersecting strands of study in the diverse area of mental model research were synthesized to investigate real-world, high-stake medical decision making. More specifically, this study examined the cog-

nition of expert radiologists reading mammography images in the search for breast cancer cues. The aim was twofold. First, an investigation into the cognition of expert medical diagnosticians took place, as represented by networks of concepts associated with detecting cancer in mammography images. This allowed to gauge similarities and differences in expert cognition. Second, the prior investigation of cognition was related to the diagnosticians' behavior of correctly or incorrectly diagnosing cancer in mammograms. Relating cognition and behavior allowed to gauge how successful the quantification of cognitive representations is able to predict overt behavior and, ultimately, whether this can help to make mammography screening in the real world to be more accurate and reliable. Alas, the association between cognition and behavior was weak in the present case but offers additional specific directions for future research.

This study proceeded as follows. First, several domains of the mental model literature are summarized and synthesized to derive a starting ground to investigate the relationship of mental models and diagnostic accuracy in breast cancer detection. Second, this synthesis is used to build a mixed-methods approach to eliciting and analyzing mental models of expert diagnosticians of the German Mammography Screening Program. Third, data analyses return insights into diverging cognitions and how differences relate to diagnosing breast cancer. Fourth, a critical reflection highlights methodological deficiencies of the literature and this study. Finally, a discussion summarizes the study and provides an outlook on what implications for research and practice were generated.

2 Literature Review of Mental Model Research

2.1 Mental Models

Rouse and Morris (1986, p.7) summarized the concept of mental models being "mechanisms whereby humans are able to generate descriptions of system purpose and form, explanations of system functioning and observed system states, and predictions of future system states." The concept thereby encompasses the notion of knowledge and its retrieval, being a storage of past observations and the process of recalling memories

to assess the current environments. Moreover, this definition also includes the mental models characteristic of being a starting point for the strategic planning of future actions based on trajectories derived from the environment and extrapolating from memories, which has been described as a key difference of human cognition compared to other animals (Seligman, Railton, Baumeister, & Sripada, 2016). Johnson-Laird (1983) emphasized the *model* aspect of mental models, arguing that human cognition is limited with regard to sensory input as well as memory retrieval and any mental operation that tries to assess current and future environmental conditions can only approximate the real world to a certain degree. In this regard, the concept of mental models compliments different theories of human cognition and behavior that go against a purely rational assumption underlying human thought, such as bounded rationality (Simon, 1955), ecological rationality (Gigerenzer & Goldstein, 1996) and the heuristics-and-biases program (Kahneman, 2011).

Conceptually, mental models are organizations of knowledge, often described as singular entities being more or less related to another inside an abstract space (Norman, 1983; Rouse & Morris, 1986). This compliments prominent theories of reasoning (Collins & Loftus, 1975) and memory (Anderson, 1983). Oftentimes, models are represented as domain specific networks of interconnected nodes. For example, the network of the domain *animals* might have several nodes such as *human*, *dog* and *cat*. Nodes are liked via edges that represent the strength of connections. For animals, an individual's mental model with the aforementioned nodes might have a strong link between *cat* and *dog* and a weak link of both nodes each with *human*, since the individual believes a human to be a different class of animal on their own. However, a second person might think differently and argue humans do not belong to the category animal, represented via an absence of edges (i.e., no connection) between the *human*-node and the animal nodes, or even firmly believe in the belonging of humans to the animal domain, represented via edge-strengths that are close to the edge between *dog* and *cat*. To answer the exemplary question whether humans are animals, people might fall back on their mental model and, depending on their presence and strength of the edges connecting the concept *human* to other animal concepts, reason that humans

do or do not belong to the wider category of animals. In essence, mental models are represented as domain specific networks of nodes and edges that can vary between individuals and result in possibly varying behavior.

The conceptualization of mental models as a cognitive representation of knowledge and environmental circumstances leading to differing actions, received a lot of attention in the domain of risk communication (e.g. Bostrom, Fischhoff, & Morgan, 1992; Bostrom, Morgan, Fischhoff, & Read, 1994; Bruine de Bruin & Bostrom, 2013; Jungermann, Schütz, & Thüring, 1988; Meyer, Leventhal, & Gutmann, 1985; Morgan, Fischhoff, Bostrom, Atman, et al., 2002; Wood, Bostrom, Bridges, & Linkov, 2012). A summary of the literature resulted in four main insights. First, mental models can be used to model peoples' knowledge about specific domains, such as climate change (Bostrom et al., 1994). This assessment happens on the level of nodes and edges. For example knowledge about both *carbon dioxide emissions* and *temperature* can be contained in a model as nodes in addition to their connection, e.g., the causal link of the greenhouse gas effect. Note that the connection between gas emissions and temperature can be a singular edge as well as a connection via the greenhouse gas effect as a node itself. People can differ here based on their knowledge, one person might know that emissions and temperature are generally related and another might know that the relation can be made via the greenhouse effect. Second, mental models can be used to infer what people should know. The literature assumes that for many domains a mental model which is highly representative of the real world exists. Finding this mental model is usually achieved by asking experts or scientists in a given domain, assuming they might know more than laypeople. For example, a political party worries about climate change and wants to incorporate mitigating actions into the program for the upcoming election. The party then seeks a scientist and lets him or her explain what climate change entails (i.e., nodes) and how the complex system of climate functions (i.e., edges). Third, the literature argues mental models can be used to make communication effective. This observation is claimed on comparisons between expert and lay mental models, highlighting gaps in which laypeople do not know of facts (i.e., missing nodes) or cases in which causal or correlational associations

between nodes are unknown (i.e., missing edges). Effective communication is said to take place when the message not necessarily entails the expert model but the parts of the expert model which laypeople were unaware of. Fourth and finally, the literature stresses the importance of integrating qualitative data collections in scientific and practical inquiries into expert and lay mental models. According to Bruine de Bruin and Bostrom (2013) and Morgan et al. (2002), the design of any message that intends to change behavior in a wider population begins with qualitative assessments of the literature and individuals via interviews or concept-mapping techniques (Wood et al., 2012). It is argued that the nature of more quantitatively oriented modes of data collection such as surveys are ill-suited for an open-ended generation of concepts and their relations that form mental models. Later, quantitative assessments may take place to verify mental model contents, but the detailed and comprehensive collection should take form in an interplay of researcher and subject.

2.2 Collective Mental Models

The study of collective mental models is an extension of the previously described insights, as evident by the definition of Smith-Jentsch, Mathieu, and Kraiger (2005, p.523) who referred to shared mental models as "an organized understanding or mental representation of knowledge shared by team members." The literature on *Team Mental Models* (e.g. Lim & Klein, 2006; Mohammed et al., 2010) and *Shared Mental Models* (e.g. Mathieu et al., 2000; Stout, Cannon-Bowers, Salas, & Milanovich, 1999) explicitly reference the literature on individual mental models cited earlier. However, the study of collective mental models is interested in the outcomes of diverging mental models and what determines said divergence. It is assumed, that by investigating situations in which the similarity of mental models affects group outcomes, such as performance, one can begin investigating mental models more qualitatively and derive training curricula to make the models between individuals more similar (Tannenbaum, Traylor, Thomas, & Salas, 2021).

This literature stream has seen widespread consideration in the study of organizational behavior and investigated shared cognition in laboratory settings in which

student as well as trained professionals play video-game simulations (Cooke, Gorman, Duran, & Taylor, 2007; Cooke et al., 2003; Mathieu, Heffner, Goodwin, Cannon-Bowers, & Salas, 2005) or perform real-world tasks (Kellermanns, Floyd, Pearson, & Spencer, 2008; Lim & Klein, 2006). To summarize the singular review of the field by Mohammed et al. (2010) while incorporating newer studies (e.g. Fisher, Bell, Dierdorff, & Belohlav, 2012; Gardner, Scott, & AbdelFattah, 2017; Gorman & Cooke, 2011; Santos, Uitdewilligen, & Passos, 2015; Tesler, Mohammed, Hamilton, Mancuso, & McNeese, 2018) the literature almost unequivocally found that mental model similarity was associated with better team performance, although effects sizes varied between tasks. This is often referred to the explanation, that (a) with increasing experience people gain more expertise in tasks and represent them more similarly in their mind and (b) maximizing performance often requires task-specific approaches, which peoples' mental models will approximate more accurately, and thereby similarly, over time.

When studying mental models in organizational settings, researchers assumed them to contain different dimensions with regard to the decision environments at hand, such as models of the temporal sequence of tasks, models of equipment use and models about the respective team members (Cannon-Bowers, Salas, & Converse, 1993; Santos et al., 2015). Nowadays, the two major contents that research differentiates are models of teamwork and taskwork (Mohammed et al., 2010). Teamwork models contain knowledge about interpersonal requirements of the team to cooperate as well as the strengths and weaknesses of the others. Taskwork models refer to the requirements to achieve the task at hand as well as steps to be taken during goal-pursuit.

The study of collective mental models exhibits a large heterogeneity in methods, conceptualizing models differently and varying greatly to how mental models are elicited as well as compared, as emphasized by Langan-Fox, Code, and Langfield-Smith (2000) and the meta-analysis by DeChurch and Mesmer-Magnus (2010). The elicitation of mental models residing within participants can rely on qualitative methods such as interviewing techniques or card sorting as well as quantitative approaches that have participants indicate the perceived similarity or relatedness between all pos-

sible pairwise combinations of relevant concepts, i.e., network nodes. In addition to the variance in elicitation techniques, techniques of analyzing the data vary as well. Qualitative approaches include less structured processes in which researchers infer differences via reasoning or visual inspection, if an elicitation method allows for such (e.g., card sorting). Quantitative approaches of pairwise ratings for example differ in the exact metrics and statistical analysis software they use. However, agreement is found in two aspects. First, most studies approach the collection of concepts relevant to task- and teamwork models qualitatively, for example by interviewing domain experts and summarizing their answers into nodes that may be represented in networks (Lim & Klein, 2006). Second, the literature concerns itself with analyzing and comparing mental model structures as a whole, meaning the analysis happens on the whole level of the network including all nodes and edges. While this allows for a great flexibility of the overall investigatory approach to elicit and analyse mental models in diverse settings, the question remains, whether all methods assess the same latent construct that is defined as shared or team mental model, complicating the comparison between studies and drawing inferences from the research domain as a whole (DeChurch & Mesmer-Magnus, 2010).

And yet, many publications over the years proliferated the practical relevance of similar mental models for organizations to achieve high performance (e.g. DeChurch & Mesmer-Magnus, 2010; Langan-Fox et al., 2000; Lim & Klein, 2006; Mathieu, Leslie, & Luciano, 2021; Tannenbaum et al., 2021). Thereby, the concept of mental models was not only regarded as an analytical indicator to measure peoples' behavior but has been advertised as an instrument to guide organizations towards better performance. Based on the track-record of mental model research in teams, such predictions seem to have high face validity. However, the verification of such may be less practical in reality. Verifying whether similarity of mental models are (a) predictive of task performance and (b) predictive with a meaningful intensity compared to different predictors of task performance would require organizations to follow the scientific literature on mental models. With ambiguous methodological approaches, there may be more practical ways to assess task and team performance, such as simply collecting data on past

performance, which is usually a very good predictor of future performance (Budescu & Chen, 2015; Mannes, Soll, & Larrick, 2014).

2.3 Cognitive Network Science

The study of cognitive network science shares its roots with the study of mental models but borrows its analytical approaches from the mathematical study of graph and network analysis (Baronchelli, Ferrer-i Cancho, Pastor-Satorras, Chater, & Christiansen, 2013; Castro & Siew, 2020; Siew, Wulff, Beckage, & Kenett, 2019). This means many studies in this domain consider objects in memory to be represented as networks of nodes and edges as well, which has received evidence from neuroscience (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013; Schapiro, Turk-Browne, Norman, & Botvinick, 2016; Schuck, Cai, Wilson, & Niv, 2016). Moreover, the elicitation of pairwise similarity ratings between nodes to form edges of individual network representations is shared with the study of collective mental models (e.g. Wulff, Hills, & Mata, 2021). However, the literature's interest typically does not lie in the question of how cognitive networks are related to the performance of interacting individuals in organizational settings. Additionally, the area of cognitive network science may consider networks' nodes to be semantic representations of language, whereas the study of mental models can include more abstract concepts as nodes in a network. To make this contrast more clear, Wulff et al. (2021) elicited similarity ratings of animals, whereas one exemplary pairwise comparison by Lim and Klein (2006) asked for the relatedness between team members' agreement on strategy and team members' understanding of the task. And yet, the domains assess identical network data structures to which cognitive network science supplies a more diverse set of quantitative measures. While mental models in the organizational research have so far focused on the comparison between network structures as a whole, cognitive network science allows for the assessment on three levels (Siew et al., 2019). Macroscopic measures quantify the organization of networks, which is akin to the analysis procedures used in collective mental model research (Santos et al., 2015). For example, Wulff et al. (2021) show that with increasing age, mental models become more distinct between

individuals. Additionally, on the level of mesoscopic analysis, network science is interested in finding communities within networks, searching for subsets of nodes that are more strongly connected with each other compared to all other nodes. For phonological networks, Siew (2013) finds that shorter words with less syllables were stored more closely to another compared to longer, multisyllabic words. On the microscopic level, analysis considers individual nodes and edges. Stella, De Nigris, Aloric, and Siew (2019) showed that students and professors connected different words to scientific areas of study such as *mathematics*; students associated this domain more with methods to learn, whereas professors related the area of study with research and words like *science*.

2.4 A Contemporary Synthesis to Investigate Cognition and Behavior

The following paragraphs synthesize the study of three distinct but related domains of cognitive science to inform how contemporary studies interested in organizational performance can make use of the relevant insights. Similarities are present across the domains investigating individuals and groups in the approach to approximate human cognition as networks, which lends network-specific methods to the analysis of individual networks and the comparison between the networks of various individuals.

Differences, however, are plentiful. First, if viewed on a continuum between solely qualitative and quantitative methods, the domains vary in their location. This applies both to the elicitation of mental models, as well as the analysis of collected data. Hence, whereas the research in the mental model approach to risk communication explicitly requires the qualitative consultation of experts, cognitive network science heavily relies on the quantitative analysis of larger data; while team mental model studies first detect important nodes from qualitative data to later quantitatively compare network structures.

Second, the domains have fundamentally different aspirations for applicability outside of academia. So far, cognitive network science has predominantly focused on informing academia how to model cognition and derive insights from said mod-

eling approach. In contrast, studies of mental models in risk communication and organizational considerations of shared mental models repeatedly lay claim to their practicality.

Altogether, there may be complementary benefits by selecting a combined approach. Following Mohammed, Klimoski, and Rentsch (2000, p.127) who referred that "the task context must be specified before the issue of appropriate measurement strategies can even be approached", it seems perfectly sensible to qualitatively consult expert decision makers in tasks that researchers have no training in. How else might a researcher infer what concepts represented as nodes are important for the mental model? Moreover, while quantitative approaches to network analysis have been present in both individual and collective mental models, the study of network science may supply additional metrics, which in turn create the possibility to investigate formerly unthinkable questions in relation to the comparison between cognitions. A combined application is therefore composed of a mixed-methods investigation into human cognition and behavior. In the following, the present study makes use of such a mixed method approach, which is explained in great detail after describing the overall decision context.

3 The Present Study

3.1 Breast Cancer and the German Mammography Screening

Breast cancer is the deadliest type of cancer for women worldwide (Key, Verkasalo, & Banks, 2001; Winters, Martin, Murphy, & Shokar, 2017) and the most common type of cancer among women in Germany (Bundesministerium für Gesundheit, 2016). Like the United Kingdom, Germany implemented a nation-wide Mammography Screening Program to detect breast cancer as early as possible, to allow for reliable prevention and alleviate the burden of many women and their families. Every woman above the age of 50 years receives a free mammography screening every two years. Every mammogram screening is performed within certified screening units and every image is assessed by two independent radiologists within a screening unit, who do not have

insight into each other's diagnosis. If at least one of the radiologists detects any cue of malignant deviations from healthy breast tissue, the pictures are assessed again in so-called consensus conferences between the radiologists and the leader of the screening unit, who are generally the most experienced diagnosticians and take responsibility for the decisions. If a consensus conference infers a risk of cancer, the woman is contacted for further assessment or biopsy.

Each radiologist within the German Mammography Screening Program undergoes extensive training beyond their medical study and assesses at least 5000 mammograms per annum. The study of mammography readings is of interest for academia, because it allows insights into highly specialized expert decision making (e.g. Carney et al., 2012; Litvinova, Kurvers, Hertwig, & Herzog, 2019) and it signifies a perfect example of making organizational use of collective intelligence via aggregation of independent judgements (e.g. Kurvers et al., 2016, 2019; Kurvers, Herzog, Hertwig, Krause, & Wolf, 2021). More importantly however, detecting reliable determinants of high performance in mammogram readings may have real-world impact and help diagnose cancer more accurately. To that end, the present study used mental models as cognitive networks to investigate mammography reading performance and gauge its ability to make the detection of cancer more accurate and, therefore, render more effective diagnoses.

3.2 Research Questions

The present study posed research questions from three perspectives. But to do so, one needs to first determine what outcomes of interest mental models could be associated with in the study of breast cancer detection. Of utmost importance is diagnostic accuracy. When a woman enters the screening unit, she expects the independent decision makers that view her mammogram to be as accurate as possible. Recently however, arguments have been raised, that accuracy is just one of two components in expert decision making that warrants attention (Kahneman, Rosenfield, Gandhi, & Blaser, 2016; Kahneman, Sibony, & Sunstein, 2021; Kurvers et al., 2021). For any medical diagnosis, one may be interested in the agreement between independent

judgements. The higher disagreement becomes, the less reliable a pair of diagnosticians appears to be. In other words, "good judges agree with each other" (Hodgson et al., 2008, p. 106). To the best of the authors knowledge, no one has yet assessed an association between mental models and judgement reliability between individuals, although the link appears to be close. If experts converge in their mental models over time, does this also relate to their similarity in behaviors? So, to address expert decision making more comprehensively, the following investigation examined both the outcomes of accuracy (i.e., making the correct diagnosis) and reliability (i.e., agreeing with other diagnosticians).

To heed the mixed-methods approach identified in the literature synthesis above, this study had to first answer the following research question: *What concepts are expert radiologists' mental models made off?* This question explicitly addressed what nodes are contained within the cognitive network of expert diagnosticians. Interviews with experienced radiologists from the German Mammography Screening Program yielded a set of 15 concepts that were used to create an online survey to share between expert radiologists. Answers derived from the survey were subjected to quantitative analysis.

In the first analysis, an investigation into the mental models of expert radiologists took place. This perspectives inquired into the extent to which experts' mental models differ and how this can be measured. This step was critically necessary to assess any associations between the outcomes of interest and the mental models of diagnosticians, since differences in networks need to be located first. Thus, the research question was partially exploratory and reads: *To what extent do mental models differ between expert radiologists?* Since quantitative methods within shared mental model research vary between studies (Langan-Fox et al., 2000) and cognitive network science supplies additional measures, this question was answered using two approaches.

The first approach regarded network structures. So-called Pathfinder (PF) networks have been previously validated for mental model comparisons in the study of shared mental models (DeChurch & Mesmer-Magnus, 2010; Langan-Fox et al., 2000; Schvaneveldt, 1990). PF networks reduce network to their sparsest possible connections between nodes, assuming that this reveals the true network structure. Networks

can then be compared based on the number of connections they share to derive a similarity measure. However, this method heavily disregards data as random noise, while cognitive network science has ways to use all of the data, as exemplified by Wulff et al. (2021). An additional measure of structural similarity was therefore derived as well based on all nodes and connections within networks. Based on these full networks, the secondary approach makes use of contemporary insights from cognitive network science. Following Siew et al. (2019), the mental models were therefore additionally analyzed using macro-, meso- and microscopic network measures. On the macroscopic level networks were compared based on their efficiency as Average Shortest Path Length (ASPL), connectivity as average degree ($\langle k \rangle$) and structuredness as average local clustering coefficient C (for a review, see Castro & Siew, 2020; Siew et al., 2019). For example, Vitevitch, Chan, and Roodenrys (2012) showed an association of structuredness with the ability to correctly recall memories. On the mesoscopic level, modularity (Q) assessed how easily networks break apart into subnetworks, which has been linked to easier information flow in networks (e.g. Marko & Riečanskỳ, 2021). The microscopic analysis focused on the importance of an individual node i, measured as closeness centrality (CC_i), which evaluates how easily node i can be accessed by all other nodes.

The second analysis tested whether network measures were related to behavioral outcomes of individual diagnostic accuracy as well as reliability between radiologists. To gauge whether mental models can practically inform breast cancer detection, the relationship between cognition and behavior needed to be established. Therefore, the following question was proposed: *How are measures derived from cognitive networks related to the behavioral outcomes of diagnostic accuracy as well as reliability between radiologists?* This in turn could inform how future diagnosticians might be trained and where common misconceptions within mental models occur. For example, high-performing individuals may agree on specific relationships between nodes, which separates them from low-performing individuals.

3.3 Methods

3.3.1 Pre Study: Qualitative Collection of Mammography Concepts

The following procedure ascertained relevant concepts of mental models in mammography readings and was based on the recommendations by Langan-Fox et al. (2000), Bruine de Bruin and Bostrom (2013) and DeChurch and Mesmer-Magnus (2010). The following pages recollect the process sequentially in a narrative style. The clear goal was to arrive at a set of relevant concepts that are part of diagnosticians' mental network during mammography screenings and allow for further quantitative assessment.

First, the researcher made himself acquainted with relevant studies of mammogram screening accuracy and facilitating conditions (e.g. Carney et al., 2012) as well as the setting of the German Mammography Screening Program by reading the information material freely available online. It became clear, that all relevant concepts would belong to the domain of taskwork mental models (Mathieu et al., 2005), as there is no interacting teamwork involved in the initial assessment of images. This step is done individually by the radiologist, uninfluenced by the diagnostic judgement that a second radiologist may have already made or may make in the future. Furthermore, the Mammography Screening Program assesses the likelihood of cancer in a mammogram in accordance to the American College of Radiology's (ACR) current and fifth version of the Breast Imaging Reporting and Data System (BI-RADS) (American College of Radiology, 2013). The BI-RADS is also used for training new radiologists and is the basis for any further on-the-job training. Therefore, the author also began to read the German and English versions of the BI-RADS atlases, as to translate insights from English studies into the German-speaking context of this study when necessary. The BI-RADS clearly delineates a list of visual cues for cancer detection, including pictures of how this can be represented in a mammogram. However, since the author has no practical experience in screening real mammograms, he was unable to decide which cues are relevant in practice.

Therefore, the author interviewed two screening unit leaders on two separate occasions. Each interview lasted for slightly more than one hour and was accepted on a voluntary basis, upon giving an informed consent. Initial contact was established

BI-RADS Concepts for Mental Models Analysis
The breast exhibits low density is almost entirely fatty.
The breast exhibits high density and has many fibroglandular mass.
Any type of mass.
Any type of calcification.
Any type of architectural distortion.
Any type of asymmetry.
Intramammary lymphnode.
Skin lesion.
Solitary dilated duct.
Skin retraction.
Nipple retraction.
Axillary adenopathy.
Skin thickening.
Malignancy: Decision for recall and further diagnosis
Benign: Decision for no recall; routine care

Table 1: Final set of concepts that was used to assess mental models of expert mammogram diagnoses.

by email, interviews were conducted via video-conference. Both interviews were semi-structured to allow for as much free recall of associative memory as possible by the interviewees. The interviewer only interrupted if clarification was necessary. The interview procedure began with open-ended questions and moved gradually to more specific questions. Important questions and answers of the interviews are supplied in Appendix A.

The synopsis revealed the following insights. First, individual cues are not assessed alone. Answers by the radiologists clearly connected any cue of cancer presence with either the outcome to keep women in the regular two year screening interval when no cancer indication is visible or to recall the patient for further inspection. Second, when specifically probed for the BI-RADS collection of cues, the interviewees stated to rely less on the system as they became more experienced. They said intuition takes place. However, they agreed that every radiologist is firmly acquainted with the system, as it is the foundation of all training material. Asking for the impact of breast density revealed that experts adjust their reading speed and attention to specific cues in light of high breast density, as this is a BI-RADS certified complication of cancer visibility.

In summary, the BI-RADS list of cues is not solely sufficient to build expert mental models. Mental models also include decision outcomes of clients' recall or no recall for additional cancer testing. Moreover, cue attention is moderated by the breast density. In principle, the clear distinction between cues validated future procedures to elicit and construct networks per pairwise relatedness ratings for further analysis of radiologists' cognition as found in the literature (DeChurch & Mesmer-Magnus, 2010; Langan-Fox et al., 2000). However, since all possible pairwise combinations of concepts need to be assessed by a participant, the completion time of such a task grows exponentially with increasing numbers of nodes. To represent this numerically, let N be the number of judgements $\binom{n}{k}$, whereas n is the number of available concepts and 2 is the number of objects that need to be judged in their relatedness. Varying n from 10 to 20 to 30 concepts results in $N = 45$, 190 and 435 judgements respectively. Since the author was unable to financially compensate participation in further data collections, a solution needed to be found to constrict the number of concepts so that participants would still participate voluntarily and not stop prematurely, while also ensuring a representative collection of nodes. This was achieved by constricting the BI-RADS classification of 4 levels of breast density into 2 levels of high and low density respectively as well as not asking for the 7 BI-RADS categories of malignancy likelihood but rather focusing on recall or no recall decisions. The latter step is also in line with the BI-RADS recommendation of auditing diagnostic accuracy, as this assesses only correctness of recall or no recall decisions. In addition to these concepts, all cues of the BI-RADS system were added without specifications of cue location in relation to the breast and without the cue of trabecular thickening, which is usually related with past radiotherapy and usually not caused by cancer (American College of Radiology, 2013). Morever, all concepts were simplified. For example, calcifications can have diverging visual representations. So, as not to ask for every subtype, the concept comprises all calcifications by explicitly stating *any type of calcification*. The final concept pool contains 15 concepts of cancer detection. Table 1 lists these concepts and how they were semantically operationalized for use in the subsequent quantitative cognitive network elicitation.

3.3.2 Design

The present study's design followed the typical mental model literature (e.g. Lim & Klein, 2006; Santos et al., 2015; Wulff et al., 2021) in that it was based on an cross-sectional approach in which participants perform a task of cognitive network elicitation, of which the results are then compared and possibly associated with dependent variables. While in general, studies comparing cognitive networks present stimuli to ascertain task performance themselves (e.g. Edwards et al., 2006), the practice of mammogram diagnoses constrained this possibility in the present case due to two aspects. First, radiologists of the German Mammography Screening Program evaluate at least 5000 mammograms per year. As someone not trained in mammogram diagnoses, selecting a representative sample of images while ensuring women's informed consent in the use of their image material for research purposes, was unfeasible. Second, mammography reading accuracy depends on the imaging technique employed (i.e., capturing the image) as well as the way in which images are displayed to diagnosticians. In Germany, the technical requirements are regulated by law (Kassenärztliche Bundesvereinigung, 2021), regularly maintained by trained technical personnel and therefore difficult to ensure for researchers outside of clinical settings. While previous studies interested in mammography reading accuracy let participants view images on their personal computers (e.g. Carney et al., 2012), advances in imaging as well as monitor resolution and contrast in clinical practice cannot be met by consumer hardware. Due to this, the screening unit leaders interviewed advised that professional radiologists would likely not participate in a data collection that measures reading accuracy based on cancer judgements made on regular consumer screens.

Within the Germany Mammography Screening Program, diagnosticians are examined on their diagnostic accuracy at least every two years as part of the screening program. So-called case collection exams are taken in the controlled environment of the Screening Program that ensure optimal technical requirements also used in the routine screening practice (Kassenärztliche Bundesvereinigung, 2021). The case collection exams therefore posed a great research opportunity for the scoring of individual accuracy as a dependent variable of cognitive network measures. Therefore, this study

collaborated with two of five Reference Centers, who are responsible for the training and evaluation of the program, and gained access to past case collection exam data. The process is delineated in the following section.

3.3.3 Procedure

The procedure to attain cognitive networks and case collection data differed between reference centers. At the Reference Center Nord in Oldenburg, the institution shared an email with their diagnosticians that invited them to participate in the study. The email included a link to the Qualtrics survey platform and notification that, after giving informed consent in the survey, allowed the reference center to share past case collection exam data with the study author. The survey can be found in Appendix B. To ensure anonymity, the reference center included a list of anonymized IDs for each participant in the email. Each participant used their personal ID at the beginning of the survey. That way, the study author was only able to see IDs in the Qualtrics platform and unable to associate individual responses with names. After the data collection was complete, the study author sent the list of IDs that participated to the Reference Center Nord, which returned the data of past case collection exams for each individual ID. So, the study author was able to match IDs between the data collected in the online survey environment and the performance data. Data collection began on the 28th of April 2021 and lasted until 10th June 2021. Final data were obtained by the author on 1st September 2021.

The same approach was suggested to the Reference Center Berlin. To comply with the reference center's specific conditions, the process was slightly altered. The study author contacted screening units by himself via email, which included the same link to Qualtrics and personalized IDs for the radiologists specific to a screening unit. As suggested by the Reference Center Berlin and to comply with data protection requirements, this process followed a pseudononymization procedure to not associate data from the online survey environment and case exam data in the same dataset. After receiving the case collection data, the study author deleted all digital information that allowed the reversal of the pseudononymization. The data collection began on

30th August and finished 10th September. Data were obtained on 10th September.

The online survey proceeded as follows. Participants were informed about the study intent and gave informed consent to use their answers for the study as well as consent for the respective reference centers to share the case collection exam data with the study author. Afterwards, participants were instructed to participate in a short trial instalment of the cognitive network elicitation. This trial used the same graphical user interface as the subsequent measures and was implemented so that participants could get acquainted with the task. For the trial as well as following measure, participants were reminded that there existed no correct response and they should answer according to their personal opinion. Next, the cognitive network elicitation followed. Afterwards, participants were again thanked for their time. Participation was not remunerated. Participation in the survey could be completed in more than one setting, as every click was saved intermittently, even if a web browser was closed. On average, participants took a little over 21 minutes to complete the survey ($M = 21.34$, $SD = 19.86$).

3.3.4 Measures

Cognitive Network Elicitation. The 15 concepts derived from the qualitative pre-study were used to elicit cognitive networks containing the concepts represented as nodes. Similar to previous studies in team mental models, as well as cognitive network domain (e.g. Edwards et al., 2006; Wulff et al., 2021), connections between nodes were based on participants' ratings of relatedness between all 105 unique possible pairings of concepts. Relatedness ratings were elicited on a 5-point Likert scale of 0 (= no relation) to 4 (= very strong relation). The sequence of concept pair-occurrence was randomized.

Cognitive Network Construction. In accordance with prior studies which associated similarity in mental models with performance (e.g. Lim & Klein, 2006), the so called Pathfinder (PF) algorithm (Quirin, Cordón, Guerrero-Bote, Vargas-Quesada, & Moya-Anegón, 2008; Schvaneveldt, 1990) was used to form PF networks for each participant based on the 105 pairwise relatedness-ratings between the 15 nodes. Some

authors argue that such relatedness-ratings contain measurement error and/or random noise and therefore a mental model can only be extracted by reducing the networks based on relatedness-ratings to the most essential components (Kenett, Ungar, & Chatterjee, 2021; Lim & Klein, 2006). In the present study, the PF algorithm kept only those connections that received the strongest relatedness-ratings while reducing the number of connections to a minimum that allows every node to have a connection to another node (i.e., 14 connections in the case of 15 nodes). PF networks are so-called unweighted networks as the relations between nodes are set to either one of the statuses present or absent.

Contrary to PF networks, measures employed in cognitive network science do not necessarily disregard connections and their respective connection strengths as noise (Siew et al., 2019; Wulff et al., 2021). So, to gauge the extent to which contemporary network measures can be associated with the outcomes of interest, additionally one full network per participant was constructed. These full networks are equally based on the 15 nodes but used all 105 pairwise relatedness-ratings to associate every node with every other node. The full networks are considered weighted networks as the connections between nodes are able to numerically vary in terms of their relatedness-strength in addition to being absent. In the end, two types of networks, PF networks and full networks, were constructed per participant and used to derive the independent variables described in the following sections.

Cognitive Network Similarity. This study considered two types of cognitive network similarity. First, PF similarity was calculated (Lim & Klein, 2006; Schvaneveldt, 1990) using the following formula for two networks:

$$\text{PF.sim} = \frac{I}{J - I}, \tag{1}$$

where I is the number of connections that two networks share and J is the number of all connections contained in both networks. For all possible pairwise combinations of participants, formula 1 was used to compare networks and averaged over all comparisons per individual to gain a PF similarity score resulting from comparing an individual to all others. PF similarity ranges between 0 and 1 while higher scores

indicate higher similarity between networks.

Additionally, similarity of the full networks was calculated. While Langan-Fox, Wirth, Code, Langfield-Smith, and Wirth (2001) suggest to treat the pairwise relatedness ratings as vectors within an Euclidean space and calculate the distance between vectors to gain a measure of similarity. Recent arguments in network studies have been raised to treat them within the framework of Likert and hence ordinal-scaled data. This approach generally offers to use correlation indices to measure the relational strength between sets of ratings. As generally emphasized by Liddell and Kruschke (2018) and taken into the network domain by Epskamp, Borsboom, and Fried (2018); Epskamp and Fried (2018), the use of Pearson's r or Spearman's *rho* correlation coefficients on ordinal data such as Likert scale responses is error prone and should therefore be substituted by the use of polychoric correlation (Olsson, 1979; Rigdon & Ferguson Jr, 1991). So, similarity of full networks was derived by computing polychoric correlations of ratings per possible pairwise combination of individuals and averaging over every individual's pairwise correlations.

Cognitive Network Measures. This study considered independent variables on the macro-, meso- and microscopic network level. On the macroscopic level, efficiency as Average Shortest Path Length (ASPL), connectivity as average degree ($\langle k \rangle$) and structuredness as average local clustering coefficient CC were of interest. Following Siew et al. (2019) and Rubinov and Sporns (2010), the measures were derived using the following formulae. The measures pertained to networks denoted as $G = (N; V)$, whereas N is the number of nodes and V is the number of connections between nodes. Connections have the strength w_{ij} between node i and j.

ASPL is a measure of how efficient information can move within a network, meaning stronger connections (i.e., shorter lengths) allow for easier traversal between nodes. It is computed as follows:

$$ASPL = \frac{1}{N(N-1))} \cdot \sum \alpha_{ij} \in g_{i \leftrightarrow j} f(w_{ij}), \tag{2}$$

whereas N is the number of nodes, α denotes whether there is a connection between i and j (0 if a participant rated the absence of a connection), $g_{i \leftrightarrow j}$ is the

path with the shortest connection between nodes and f is an inverse from high to low connection strength. ASPL is known to increase with age (Kenett et al., 2021; Wulff et al., 2021), but it is unknown how it relates to outcomes of mammogram readings.

Moreover, the average degree ($\langle k \rangle$) is considered. ($\langle k \rangle$) measures how connected nodes are within a network as it averages the numbers of connections between all nodes. The formula is as follows:

$$(\langle k \rangle) = \frac{1}{N} \cdot \sum_{n=1}^{N} k_i, \tag{3}$$

whereas k_i is the number of connections node i has. Differences in ($\langle k \rangle$) were again observed between age groups with younger adults having more connected networks compared to adults (Wulff et al., 2021). Kenett (2018) associated a more strongly connected network with higher creativity and less rigidity of thought. It has also been shown that accurate memory retrieval is associated with higher average degree (Vitevitch et al., 2012).

Last on the macroscopic level, the average clustering coefficient CC was derived. CC measures the extent to which nodes cluster together as triangles, allowing for a measure of cluster presence per node i. Higher clustering coefficients are present in denser networks with more strongly and numerously connected groups of nodes. For node i the local clustering coefficient is

$$c_i = \frac{2}{k_i(k_i - 1)} \cdot \sum_{j,k} (w_{i,j} w_{j,k} w_{k,i})^{1/3}, \tag{4}$$

whereas k is the degree of i and w is the intensity of the triangle between pairwise combination of nodes i,j and k and $w = w_{i,j}/max(w_{i,j})$ (Onnela, Saramäki, Kertész, & Kaski, 2005). For a complete network, the average clustering coefficient therefore is

$$CC = \frac{1}{N} \cdot \sum_{n=1}^{N} c_i. \tag{5}$$

In the realm of cognitive network science, higher CC has been observed in older compared to younger individuals (Kenett et al., 2021; Wulff et al., 2021) and faster

semantic retrieval (Siew, 2018).

On the mesoscopic level, modularity Q was measured. Q denotes how easily a network breaks apart into smaller sub-networks. For weighted networks, the formula is

$$Q = \frac{1}{V} \sum_{i,j \in N} \left[w_{ij} \frac{k_i k_j}{V} \right] \delta(s_1, s_2), \tag{6}$$

where V is the number of connections in the network, w is the connection strength between i and j, k is the number of connections per node (i.e., the degree) and the function δ denotes 1 if the communities (i.e., triangles) s_1 and s_2 of the nodes i and j are the same, else 0. Marko and Riečanskỳ (2021) showed that higher Q can be associated with more information flow in semantic networks and Kenett (2018) found higher Q in more creative individuals.

On a microscopic network level, individual local clustering coefficients are derived for the nodes *malign* and *benign*. This used formula 4. Associating the local clustering coefficient c_i with dependent variables was exploratory and was chosen since these nodes relate to the major decision outcomes of mammography diagnostics.

Similarity in Cognitive Network Measures. While the cognitive network measures themselves may have an association with dependent variables, the literature on team mental models is primarily interested in the *similarity* between networks (e.g. DeChurch & Mesmer-Magnus, 2010). To measure a similarity between these measures, the symmetric mean absolute percentage error (Tofallis, 2015) was used as a measure of relative variance between individuals' measures, following the recommendation by Kahneman et al. (2016). What henceforth is called the symmetric mean absolute percentage deviation (SMAPD) was conceptualized as a relative measure of numerical deviation invariant of scale. SMAPD takes a value between 0 and 1, whereas higher values delineate higher dissimilarity. It is defined as

$$\text{SMAPD} = \frac{100\%}{n} \sum^{n} \frac{|m_x - m_y|}{|m_x| + |m_y|}, \tag{7}$$

whereas n is the number of all possible pairwise combinations for the network

measure m of individual x which is compared to a different individual y per possible dyadic combination. Using a similarity measure that forces all inputs within the same bounds of 0 and 1, comparisons between measures of different sizes become more easily interpretable.

Mammogram Reading Accuracy. Each case collection exam consisted of 50 cases of which both breasts were shown each from the mediolateral-oblique (i.e., from the side, angled at 40° outwards from the body) and craniocaudal view (i.e., from above). Case collection exams contained a higher frequency of malign cases (between 21 and 29 out of 50) than naturally occurring in the general screening population (Kassenärztliche Bundesvereinigung, 2021).

For each breast, radiologists classified the case as belonging to one of 5 categories (1 = normal breast, 2 = benign lesion, 4a = suspect lesion (between 2 and 10 % percent of malignancy), 4b = suspect lesion (between 10 and 50 % percent of malignancy), 5 = malignancy (greater than 95 % percent of malignancy)). Each individual breast had a correct target category assigned by an independent panel of experienced radiologists with access to histological information, who update the pool of mammograms available for case collection exams on an annual basis (Kassenärztliche Bundesvereinigung, 2021).

Accuracy of mammogram diagnoses within the case collection exams were calculated based on American College of Radiology (2013) and the German Mammography Screening Program's performance audit criteria (Kassenärztliche Bundesvereinigung, 2021). Radiologists' category judgements and target categories were transformed to belong to either *benign* or *malign*, collapsing category 1 and 2 to *benign* and 4a, 4b and 5 to *malign*. In the screening practice, such binary judgements correspond to the the decision whether or not to recall a client for further investigation of actual cancer presence. In case the benign or malign judgement of a radiologist matched the target category, it was classified as *correct* or else *incorrect*, thereby creating a binary outcome measure of accuracy.

The German Mammography Screening Program additionally offers a more nuanced measure of accuracy by calculating deviation points (Kassenärztliche Bun-

desvereinigung, 2021). Deviation points assign a numerical value when the categorical judgements and targets differ, whereas higher points correspond to less accurate judgements. Table 2 depicts how deviation points were calculated for the present study.

Table 2: Matrix of deviation points per judgement and target category.

		Judgement			
		1	**2**	**4a/4b**	**5**
Target	**1**	0	1	5	6
	2	1	0	4	5
	4a/4b	6	5	0	1
	5	7	6	1	0

Mammogram Reading Similarity. To investigate the association between cognitive similarity and similarity in diagnoses, mammogram reading similarity was calculated. The measure is binary and corresponds to whether a pair of radiologists formulated the same or a different diagnoses of either *benign* or *malign* per singular mammogram, thereby using the transformed categorical judgements also used for the accuracy measure.

3.3.5 Participants

The sample size estimation of the present study was based on several criteria. Studies interested in determinants of mammogram reading accuracy typically consider *decisions* instead of *individuals* to be the sampling units of primary interest (e.g. Carney et al., 2012), since radiologists make thousands of breast cancer diagnoses per year. Analyses therefore result to multilevel models for which sample size estimations depend not just on the number of participants but also on the number of data points per participant (Brysbaert & Stevens, 2018; Harrison et al., 2018; Kumle, Vo, & Draschkow, 2021). Without a strong indication of expected effect size and reference data of prior studies with well powered designs, the estimation via simulation can be erroneous (Kumle et al., 2021). Therefore, this study's minimum sample size was estimated from a non-systematically rendered sample of past studies that investigated correlates of mammogram reading accuracy in high-quality journals, as ranked

by scimagojr.com (Elmore, Wells, & Howard, 1998; Esserman et al., 2002; Molins, Macià, Ferrer, Maristany, & Castells, 2008; Rafferty et al., 2013, 2014). Mean sample size were 3898 diagnoses ($SD = 1428$). To be cautious, the minimum sample size of the present study was therefore set at 5000.

In the end, 12 radiologists participated, resulting in a final $n = 7928$ after removing missing data of cases in which no definite correct answer was present. Radiologists that participated had on average $M = 5.78$ years of experience ($SD = 3.38$) in the German Mammography Screening Program as evidenced by the date of their first case collection exam. Data returned by the reference centers consisted of 81 case collection exams of which 32 were unique and of which 22 were taken by at least two of the 12 radiologists. The number of judgements made in case collection exams per radiologist varied between 100 and 1100, $Median = 1000$.

3.3.6 Data Analysis

The present study analyzed multilevel data of diagnoses nested within case collection exam nested within individual radiologist and therefore relied on mixed effects models (Gelman & Hill, 2006). Moreover, Bayesian regression models via Stan in R (Bürkner, 2017) were used, due to their flexibility to analyze multilevel data. Additional benefits of Bayesian compared to regular statistical analysis (i.e., frequentist statistics) and null hypothesis significance testing (NHST) are discussed elsewhere (e.g. Gigerenzer, 2004; Kass, 2011; Kruschke, 2013; Kruschke & Liddell, 2018; McElreath, 2018; Wagenmakers et al., 2018) but the authors agree on the fact that p-values are no indicator of favoring H_1 over H_0 but instead give "the probability of obtaining results at least as extreme as those observed given that the null hypothesis is true" (Wagenmakers et al., 2018, p.35). Bayesian analysis, however, offers the quantification of evidence of one statistical model over a different model (i.e., H_1 compared to H_0) on a continuous scale via Bayes Factors (BF) (Kass & Raftery, 1995; Wagenmakers, 2007). For the remainder of this study, BF interpretation follows the guidelines by Jeffreys (1939) and Schad, Nicenboim, Bürkner, Betancourt, and Vasishth (2021), highlighting such associations that compared to the null model have moderate (BF > 3) or strong (BF

> 10) change evidence in favor of H_1, if an effect of an independent variable can be observed.

Analyses relied on Bayesian generalized linear mixed models (GLMM) using the software package *Bayesian Regression Models using 'Stan'* (*brms*) (Bürkner, 2017, 2018; Bürkner & Vuorre, 2019). All models used 2000 iterations for warm-up and 6000 sampling iterations per each of the four Markov chains. Weekly informative default priors included in *brms* were used. Model convergence was investigated via visual inspection of the Markov chains and the use of the Gelman-Rubin-Statistic ($\hat{R}$) (Gelman & Rubin, 1992). Goodness-of-fit between models was compared using the leave-one-out information criterion (LOOIC) (Bürkner, 2017). For ease of interpretation, additional Bayesian R^2 values are reported as well.

The modeling procedure investigating accuracy of judged cancer presence followed three steps. First, the distributions of the response variables were determined as to choose the most applicable type of regression analysis. Both response variables of accuracy, either correct/incorrect as well as deviation points, followed non-normal distributions. While the binary outcome's most appropriate regression type was binomial (i.e., logistic regression), the deviation points were right-skewed and consisted of mostly zeros, in other words few large deviations from the target category were present. Using the *check_distribution* function from the *performance* package (Makowski, Ben-Shachar, & Lüdecke, 2020) revealed a zero-inflated negative binomial distribution underlying the response variable. Model comparison based on LOOIC of the zero-inflated negative binomial against a gaussian model revealed the prior to exhibit a better fit to the data and was hence chosen for further analyses. Second, the fixed- and random effect structure of the regression models were assessed based on Barr, Levy, Scheepers, and Tily (2013) and Harrison et al. (2018). Comparing models with and without random slopes revealed model convergence failures in light of random slopes ($\hat{R} > 1.01$). All models reported in this study therefore include random intercepts to assess variations among radiologists and case collection exams. Furthermore, a maximal model structure including all network measures as predictors of accuracy led to model convergence failures ($\hat{R} > 1.01$). Therefore, one regression

model each was calculated per independent variable of interest. Third, in addition to the radiologists' diagnoses the data included information of each correct target category per singular mammogram. Inclusion of these categories as fixed-effect increased the model fit over models not including this fixed effect as evidenced by lower LOOIC values.

The models assessing similarity in mammogram readings followed the same procedure and showed the same results in terms of fixed- and random-effect structure, inclusion of correct target category as fixed-effect and model convergence failure when all predictors were entered at once. Compared to the dependent variable accuracy, the multilevel structure, however, differed. For similarity in diagnoses, all possible dyadic combinations of radiologists were created and searched for case collection exams, that both radiologists participated in. This resulted in 15 dyads, who shared at least one case collection exam. Per dyad, mammogram reading similarity (same versus different diagnoses) were calculated on the level of diagnoses and similarity in network structure as well as network measures were calculated per dyad. As a result, 15 unique dyads completed diagnosing a subset of 6892 mammograms. Models investigating mammogram reading similarity therefore included a random intercept to investigate variability between dyads.

Estimates reported from the GLMMs are means of the posterior distribution of the parameters of interest. Additionally, 95% highest density intervals (HDI) are reported akin to the frequentist use of confidence intervals (Kruschke, 2014). In case a HDI around an effect does not include zero, there is credible evidence of a positive or negative association between an independent and dependent variable within a regression model.

All data analyses and visualizations were done in RStudio and R version 4.0.2 (RStudio Team, 2021). Network measures were either derived from custom functions or made use of *igraph* (Csardi & Nepusz, 2006), *SemNet* (Christensen & Kenett, 2019), *NetworkToolbox* (Christensen, 2018) and *brainGraph* (Watson, 2020).

3.4 Results

3.4.1 Cognitive Network Analysis

The analyses focused on similarities or differences between radiologists cognitive networks based on 105 pairwise relatedness ratings of cancer presence cues relevant in mammogram diagnoses. Figure 1 presents a direct comparison between the two types of networks that were considered in the present study. PF networks, common in the team mental model literature, were derived from full networks, which included all pairwise-relatedness ratings.

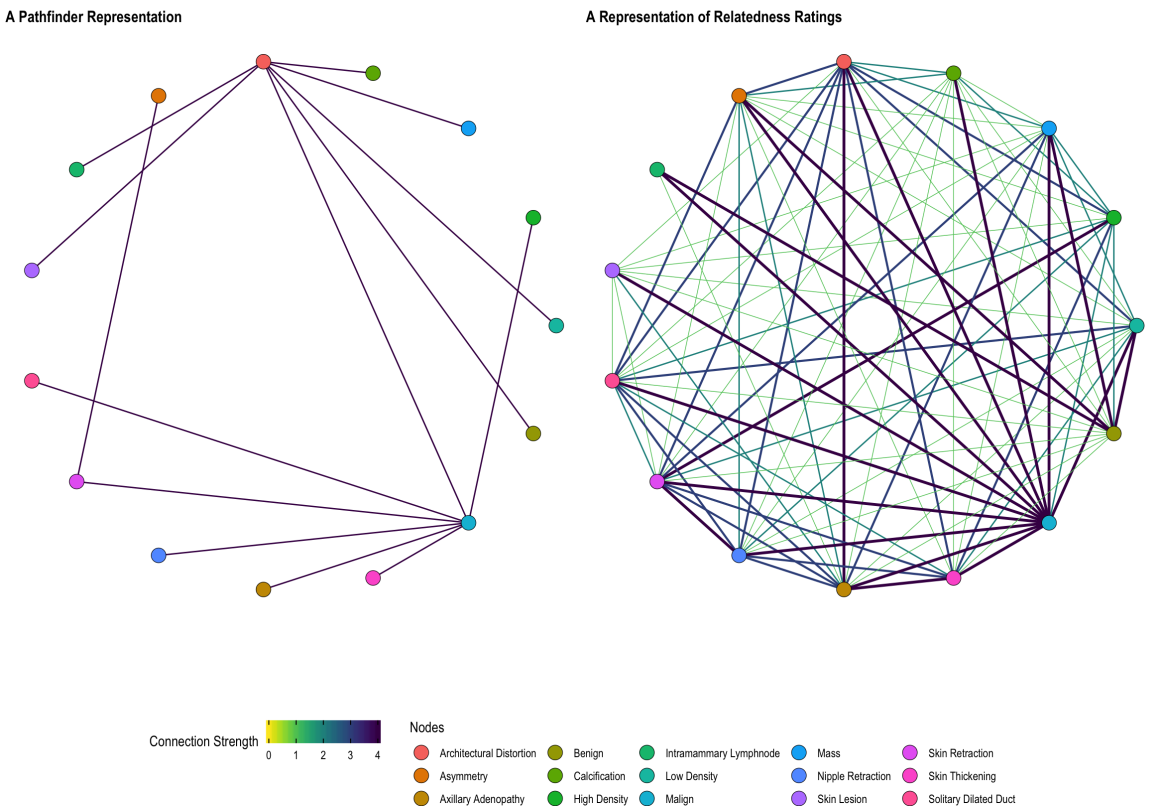

Figure 1: Visualization of network structures. The same network is represented as an unweighted network based on the PF algorithm (left) and as a network with weighted connections based on all 105 pairwise relatedness-ratings (right). Stronger connections (i.e., higher relatedness ratings) are indicated by darker colors in the right network.

To answer the research question, how cognitive networks differ, Figure 2 present the PF networks of all participants of this study. Visual interpretations suggests great

divergence. Networks differ in their most central node, i.e., the node which has the most connections. Although the node *malign* seems to be the most central node in 8 of 12 networks, any additional connections that do not include *malign* differ greatly between the 8 networks.

A similar pattern emerges when looking at the non-regularized networks based on the 105 pairwise relatedness-ratings (Figure 3). Not only differ networks in their most central node, but differences emerge on two more levels. First, networks differ in their connectedness, meaning they present diverging numbers of connections between nodes. Second, networks differ in their strengths of connections. For example, the first network in the second row shows few connections of low strength, indicated by a lighter color. In contrast, the network above (first network, first row) is much more connected and more connections are of higher strengths (i.e., darker color).

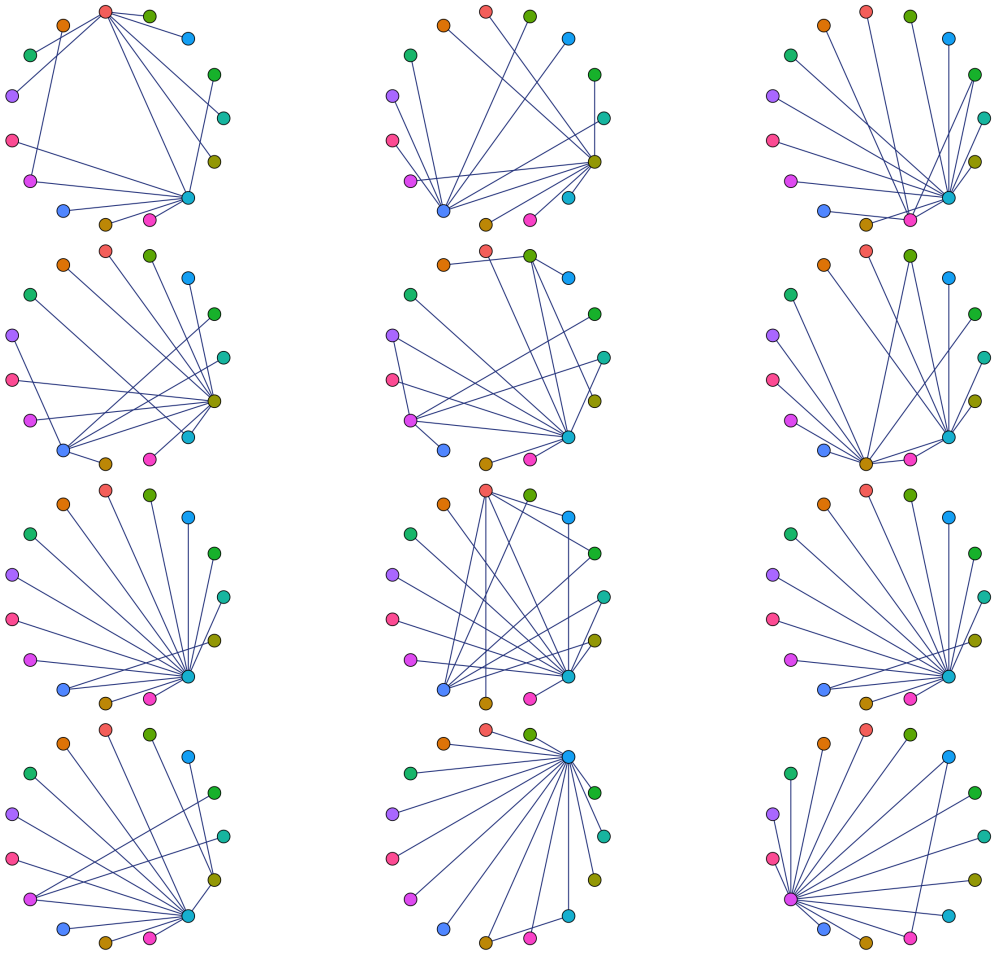

Figure 2: Visualizations of 12 radiologists' mental models based on the Pathfinder algorithm.

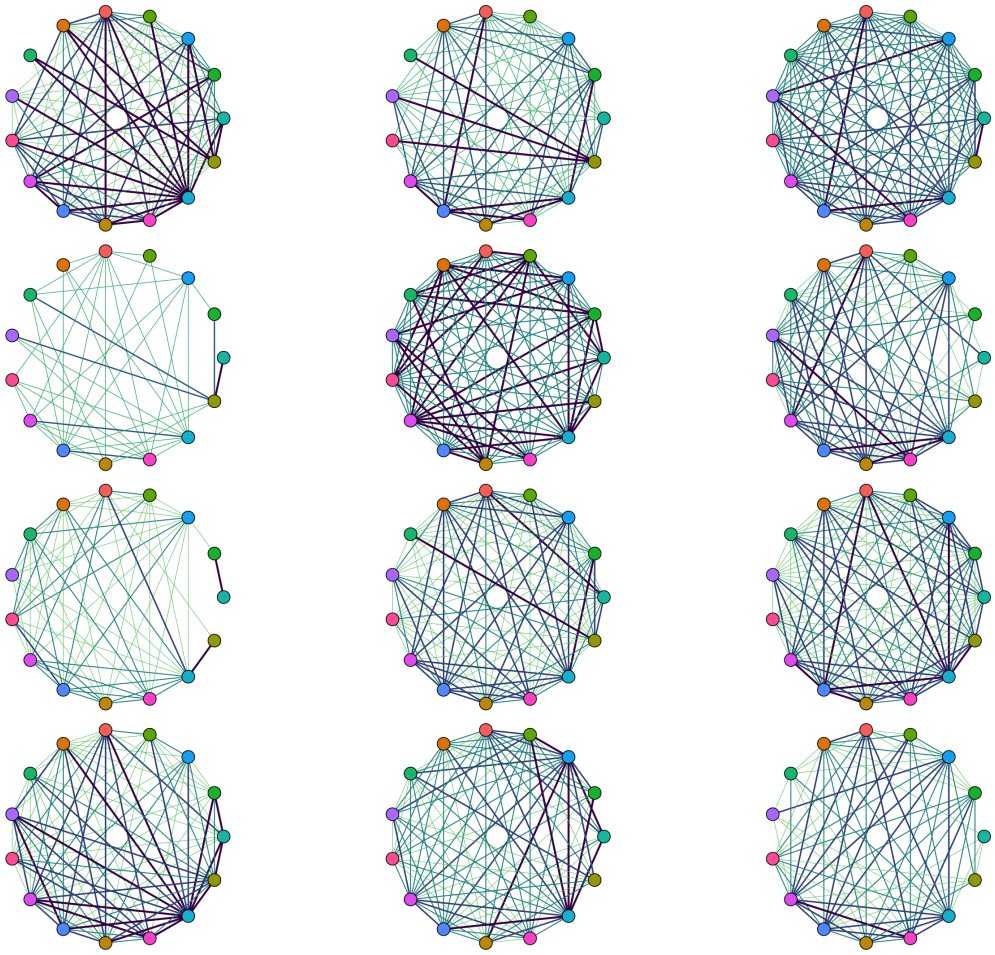

Figure 3: Visualizations of 12 radiologists' mental models based on the 105 pairwise-relatedness ratings.

Differences between networks can also be observed on a node level. Figure 4 presents the distributions across all 12 radiologists' networks of the measures path lengths (i.e., how far does one need to travel in a network to reach the node), degree (i.e., to how many nodes is a node connected) and clustering coefficient c_i (i.e., how embedded is the node across different sub-networks). Generally, the nodes of *high* and *low density* show the largest variance for all three measures. The important decision outcome of assigning *malign* to a diagnoses, also recalling patients for further examination, has the shortest path length and is very well connected to other nodes and shows among the highest clustering coefficient. *Benign*, however, shows much greater variation among all three measures. Whether the variance in network structures of PF networks and full networks as well as the variance in network measures share any

association with breast cancer diagnoses was tested in the following

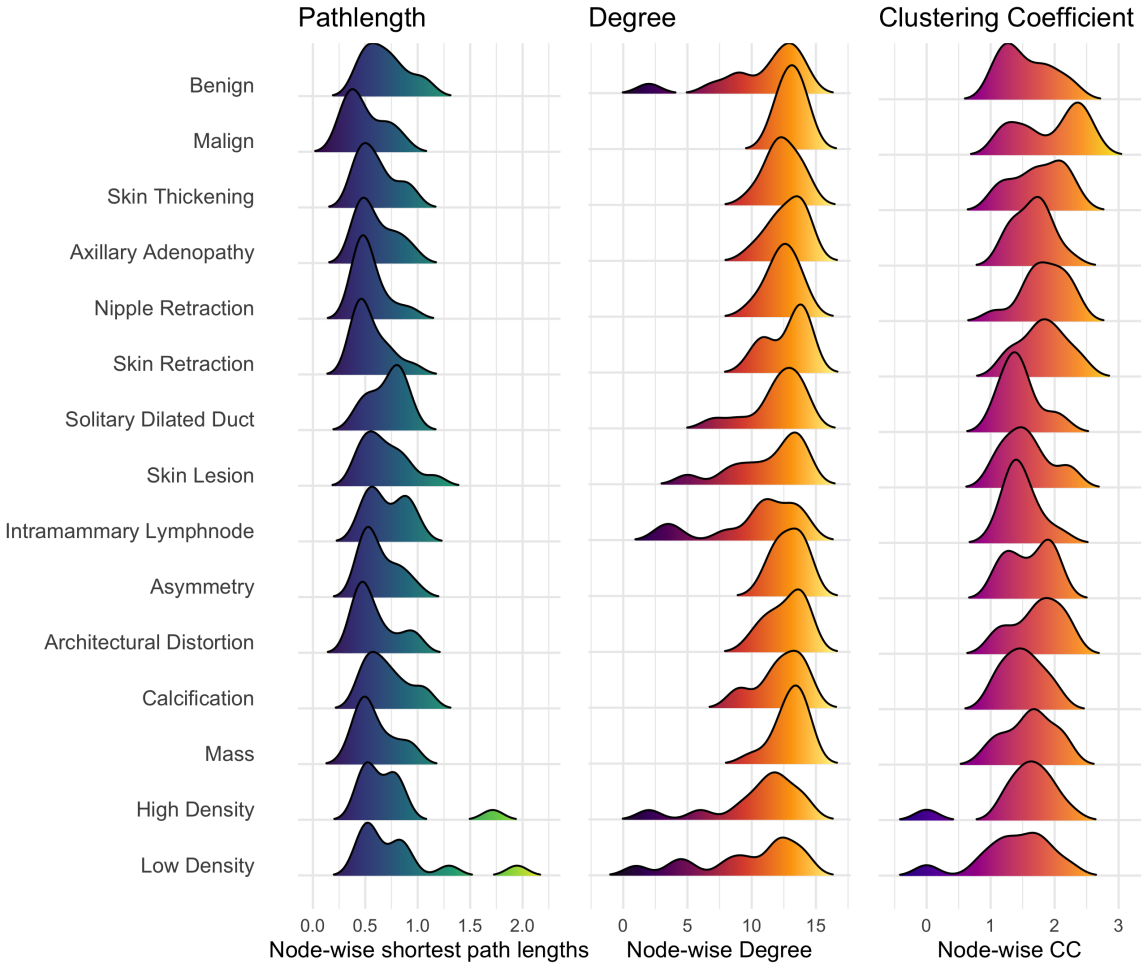

Figure 4: Distribution of path lengths, degree and clustering coefficient per node.

3.4.2 Mammogram Reading Accuracy

Table 3 presents the results of the GLMMs predicting diagnostic performance using the correct target category as fixed effect, as well as similarity in PF networks and full networks respectively. Neither PF similarity nor the polychoric correlation between the pairwise relatedness-ratings showed credible associations with the accuracy of cancer diagnoses. Credible effects, however, were observed for the correct target categories. While mammograms belonging to category 2 (i.e., benign lesions) were associated with decreased accuracy, category 5 (i.e., greater than 95 % percent of malignancy) was associated with greater accuracy in both models. This was a sensible result in light of the risks associated with wrong decisions in routine care. If a diag-

nostician errs on side of recalling a patient for further examination (i.e., assigning a malignant category), the psychological distress of a possible malignant finding as well as quick and minimally invasive biopsies may in the end cause much less harm than wrongfully diagnosing the absence of cancer. Similar results are present when using the network measures as predictors (Table 4) or similarity in network measures as predictors (Table 5). Across the models, there is mostly moderate change in evidence to belief that the inclusion of variables based on cognitive networks allows to better model the diagnostic accuracy of breast cancer screenings (BF < 5).

The literature considers cognitive networks to evolve with time (e.g. Santos et al., 2015; Wulff et al., 2021). Although there is no evidence at what rate and to what extent cognitive networks of radiologists develop, this point of view enables one to criticize testing the association between cognitive networks elicited via pairwise-relatedness ratings in 2021 and case collection exams dating back as far as 2009. To alleviate this concern, all GLMM analyses were repeated, although at reduced sample size ($n = 1100$), using only the most recent case collection exam data per participant. Results are presented in Table 11, 12 and 13 in Appendix C. The results paint a similar picture: no credible effects of the measures derived from cognitive networks were observed in the models, although the models showed better fit to the data compared to the models using all performance data as indicated by lower LOOIC values.

Table 3: GLMMs predicting diagnostic accuracy using measures of similarity in network structures.

	Model 1	Model 2
Intercept	**3.567**	**3.437**
	[2.763; 4.346]	[2.402; 4.562]
Category 2	**−0.501**	**−0.501**
	[−0.750; −0.237]	[−0.750; −0.244]
Category 4a	**−0.675**	**−0.678**
	[−1.228; −0.116]	[−1.229; −0.126]
Category 4b	−0.239	−0.239
	[−0.605; 0.139]	[−0.616; 0.128]
Category 5	**0.811**	**0.811**
	[0.263; 1.373]	[0.259; 1.373]
PF.sim	−0.585	
	[−3.552; 2.319]	
cor_{poly}		−0.022
		[−3.111; 2.889]
SD: ID	0.38	0.38
SD: Exam	0.68	0.68
R^2	0.084	0.084
Num. obs.	7928	7928
LOOIC	2836.8	2836.6
BF	3.64	3.31

Bold Null hypothesis value outside the confidence interval.

When using GLMMs to predict deviation points as an alternative measure of diagnostic accuracy using target categories as well as the cognitive network measures, one exception was observed (Tables 6, 7 and 8).

CC was credibly associated with less deviation points (β_{CC} = -.533; HDI =

Table 4: GLMMs predicting diagnostic accuracy using network measures.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**3.821**	**3.493**	**2.815**	**3.643**	**2.933**	**3.476**
	[2.807; 4.863]	[1.143; 5.787]	[1.437; 4.198]	[2.973; 4.337]	[1.849; 4.009]	[2.304; 4.863]
Category 2	**−0.502**	**−0.501**	**−0.501**	**−0.502**	**−0.503**	**−0.502**
	[−0.751; −0.241]	[−0.750; −0.250]	[−0.747; −0.234]	[−0.756; −0.248]	[−0.758; −0.247]	[−0.746; −0.247]
Category 4a	**−0.676**	**−0.676**	**−0.676**	**−0.676**	**−0.680**	**−0.678**
	[−1.214; −0.140]	[−1.221; −0.124]	[−1.207; −0.126]	[−1.208; −0.106]	[−1.216; −0.133]	[−1.216; −0.119]
Category 4b	−0.241	−0.239	−0.239	−0.240	−0.240	−0.238
	[−0.600; 0.145]	[−0.595; 0.149]	[−0.594; 0.146]	[−0.609; 0.137]	[−0.617; 0.138]	[−0.595; 0.143]
Category 5	**0.808**	**0.808**	**0.811**	**0.810**	**0.807**	**0.806**
	[0.280; 1.392]	[0.243; 1.380]	[0.268; 1.376]	[0.247; 1.374]	[0.255; 1.381]	[0.271; 1.388]
ASPL	−0.596					
	[−2.100; 0.843]					
degree		−0.005				
		[−0.198; 0.179]				
CC			0.384			
			[−0.450; 1.188]			
Q				−3.714		
				[−13.926; 5.289]		
c_{malign}					0.262	
					[−0.236; 0.815]	
c_{benign}						−0.027
						[−0.827; 0.714]
SD: ID	0.37	0.39	0.35	0.38	0.35	0.38
SD: Exam	0.67	0.68	0.67	0.67	0.67	0.68
R^2	0.084	0.084	0.084	0.084	0.084	0.084
Num. obs.	7928	7928	7928	7928	7928	7928
LOOIC	2836.3	2836.7	2836.4	2835.9	2836.2	2836.8
BF	2.52	0.21	1.79	15.7	1.21	0.90

Bold Null hypothesis value outside the confidence interval.

Table 5: GLMMs predicting diagnostic accuracy using similarity in network measures.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**3.438**	**3.573**	**3.423**	**3.324**	**3.572**	**3.379**
	[2.315; 4.643]	[2.776; 4.342]	[2.199; 4.585]	[2.425; 4.255]	[2.308; 4.931]	[2.039; 4.806]
Category 2	**−0.500**	**−0.501**	**−0.500**	**−0.502**	**−0.502**	**−0.501**
	[−0.751; −0.250]	[−0.749; −0.242]	[−0.759; −0.246]	[−0.763; −0.253]	[−0.756; −0.240]	[−0.763; −0.247]
Category 4a	**−0.674**	**−0.677**	**−0.676**	**−0.679**	**−0.678**	**−0.677**
	[−1.221; −0.136]	[−1.220; −0.141]	[−1.208; −0.133]	[−1.218; −0.125]	[−1.239; −0.149]	[−1.188; −0.110]
Category 4b	−0.239	−0.241	−0.239	−0.241	−0.241	−0.239
	[−0.613; 0.127]	[−0.611; 0.132]	[−0.616; 0.133]	[−0.614; 0.137]	[−0.614; 0.132]	[−0.594; 0.152]
Category 5	**0.812**	**0.811**	**0.813**	**0.812**	**0.811**	**0.812**
	[0.250; 1.373]	[0.251; 1.385]	[0.251; 1.394]	[0.245; 1.374]	[0.265; 1.369]	[0.267; 1.376]
$ASPL_{SMAPD}$	−0.035					
	[−3.331; 3.154]					
$degree_{SMAPD}$		−0.860				
		[−4.928; 3.319]				
CC_{SMAPD}			0.037			
			[−4.701; 4.226]			
Q_{SMAPD}				0.194		
				[−1.244; 1.657]		
$c_{malign;SMAPD}$					−0.399	
					[−4.219; 3.048]	
$c_{benign;SMAPD}$						0.176
						[−4.354; 4.203]
SD: ID	0.39	0.39	0.38	0.39	0.39	0.39
SD: Exam	0.68	0.68	0.68	0.68	0.68	0.68
R^2	0.083	0.084	0.084	0.084	0.084	0.084
Num. obs.	7928	7928	7928	7928	7928	7928
LOOIC	2836.3	2836.0	2836.8	2836.6	2836.5	2836.8
BF	3.63	4.75	5.02	1.69	4.09	4.83

Bold Null hypothesis value outside the confidence interval.

[1.103;0.031], BF = 6.23). This association held when analyses were repeated using only the most recent data (Tables 14, 15 and 16 in Appendix C) and increased in strength (β_{CC} = -.841; HDI = [-1.443; -0.225], BF = 29.09). Moreover, when using only the recent data, the local clustering coefficient c_i of node *malign* also exhibited a credible negative association with the number of deviation points (β_{ci} = -.500; HDI = [-0.899; -0.091], BF = 11.76).

Table 6: GLMMs predicting deviation points using measures of similarity in network structures.

	Model 1	Model 2
Intercept	**−1.261**	**−1.242**
	[−1.817; −0.688]	[−1.808; −0.656]
Category 2	**0.657**	**0.658**
	[0.550; 0.768]	[0.549; 0.767]
Category 4a	0.126	0.127
	[−0.174; 0.408]	[−0.151; 0.414]
Category 4b	0.156	0.157
	[−0.012; 0.315]	[−0.006; 0.317]
Category 5	0.063	0.064
	[−0.107; 0.231]	[−0.102; 0.227]
PF.sim	0.723	
	[−1.321; 2.811]	
cor_{poly}		0.697
		[−1.566; 3.020]
SD: ID	0.33	0.34
SD: Exam	0.38	0.38
R^2	0.126	0.125
Num. obs.	7928	7928
LOOIC	14314.2	14315.3
BF	3.17	3.16

Bold Null hypothesis value outside the confidence interval.

Table 7: GLMMs predicting deviation points using network measures.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**−1.614**	−1.136	−0.226	**−1.300**	−0.445	**−1.020**
	[−2.309; −0.978]	[−2.960; 0.737]	[−1.125; 0.668]	[−1.787; −0.804]	[−1.153; 0.283]	[−1.963; −0.043]
Category 2	**0.658**	**0.658**	**0.658**	**0.658**	**0.658**	**0.658**
	[0.552; 0.768]	[0.548; 0.767]	[0.545; 0.763]	[0.550; 0.767]	[0.552; 0.767]	[0.554; 0.771]
Category 4a	0.126	0.125	0.128	0.127	0.126	0.124
	[−0.165; 0.404]	[−0.157; 0.419]	[−0.154; 0.429]	[−0.170; 0.406]	[−0.165; 0.417]	[−0.167; 0.421]
Category 4b	0.157	0.157	0.159	0.158	0.158	0.158
	[−0.004; 0.318]	[−0.008; 0.320]	[−0.010; 0.312]	[−0.011; 0.315]	[−0.002; 0.321]	[−0.003; 0.325]
Category 5	0.062	0.064	0.063	0.063	0.062	0.064
	[−0.108; 0.229]	[−0.101; 0.236]	[−0.100; 0.228]	[−0.101; 0.227]	[−0.104; 0.225]	[−0.101; 0.231]
ASPL	0.806					
	[−0.109; 1.815]					
degree		0.004				
		[−0.141; 0.157]				
CC			**−0.533**			
			[−1.103; −0.031]			
Q				3.802		
				[−3.680; 11.256]		
c_{malign}					−0.334	
					[−0.696; 0.001]	
c_{benign}						−0.044
						[−0.613; 0.546]
SD: ID	0.27	0.35	0.25	0.32	0.26	0.34
SD: Exam	0.38	0.38	0.38	0.39	0.38	0.38
R^2	0.126	0.125	0.125	0.125	0.126	0.125
Num. obs.	7928	7928	7928	7928	7928	7928
LOOIC	14314.1	14314.6	14314.4	14314.6	14313.8	14314.8
BF	6.18	0.16	6.23	16.8	3.71	0.68

Bold Null hypothesis value outside the confidence interval.

Table 8: GLMMs predicting deviation points using similarity in network measures.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**−1.371**	**−1.179**	**−1.400**	**−0.949**	**−1.598**	**−1.303**
	[−2.227; −0.529]	[−1.740; −0.647]	[−2.284; −0.527]	[−1.614; −0.268]	[−2.473; −0.808]	[−2.388; −0.207]
Category 2	**0.658**	**0.658**	**0.659**	**0.660**	**0.658**	**0.658**
	[0.551; 0.770]	[0.552; 0.770]	[0.549; 0.766]	[0.555; 0.773]	[0.550; 0.769]	[0.551; 0.771]
Category 4a	0.126	0.125	0.127	0.126	0.127	0.124
	[−0.158; 0.419]	[−0.160; 0.418]	[−0.171; 0.412]	[−0.147; 0.424]	[−0.164; 0.420]	[−0.168; 0.408]
Category 4b	0.158	0.158	**0.159**	0.159	0.157	0.157
	[−0.002; 0.321]	[−0.004; 0.317]	[0.001; 0.320]	[−0.007; 0.318]	[−0.004; 0.315]	[−0.002; 0.326]
Category 5	0.063	0.063	0.065	0.066	0.064	0.064
	[−0.103; 0.229]	[−0.101; 0.227]	[−0.098; 0.237]	[−0.101; 0.238]	[−0.104; 0.224]	[−0.098; 0.229]
$ASPL_{SMAPD}$	0.842					
	[−1.569; 3.226]					
$degree_{SMAPD}$		0.555				
		[−2.452; 3.673]				
CC_{SMAPD}			1.216			
			[−1.966; 4.617]			
Q_{SMAPD}				−0.251		
				[−1.429; 0.811]		
$c_{malign;SMAPD}$					1.466	
					[−0.790; 3.778]	
$c_{benign;SMAPD}$						0.699
						[−2.936; 3.939]
SD: ID	0.34	0.35	0.33	0.34	0.31	0.34
SD: Exam	0.38	0.38	0.38	0.38	0.38	0.38
R^2	0.126	0.126	0.125	0.125	0.125	0.125
Num. obs.	7928	7928	7928	7928	7928	7928
LOOIC	14314.3	14314.6	14314.3	14314.5	14314.2	14314.7
BF	4.03	3.71	5.25	1.44	6.90	4.32

Bold Null hypothesis value outside the confidence interval.

3.4.3 Mammogram Reading Similarity

The present study was additionally interested in the hitherto untested associations between cognitive networks and the similarity in decisions. Therefore, GLMMs tested associations between similarity in network structures of PF and full networks as well as similarity in network measures for each decision of a dyad of radiologists that passed one or several same case collection exams. Results are presented in Tables 9 and 10.

Neither similarity in cognitive network structure nor similarity in network measures exhibited credible associations with mammogram reading similarity. Moreover, including the measures as a predictor within a GLMM did not provide any evidence that instances in which two radiologists independently reached the same diagnostic conclusion were better explained with than without the framework of cognitive network science (all $BF < 3$).

Table 9: GLMMs predicting diagnostic similarity using measures of similarity in network structures.

	Model 1	Model 2
Intercept	**2.779**	**2.689**
	[2.503; 3.072]	[2.324; 3.046]
Category 2	−0.138	−0.140
	[−0.358; 0.067]	[−0.360; 0.068]
Category 4a	−0.281	−0.283
	[−0.785; 0.264]	[−0.805; 0.256]
Category 4b	**0.686**	**0.689**
	[0.270; 1.109]	[0.269; 1.133]
Category 5	**1.837**	**1.843**
	[1.160; 2.520]	[1.177; 2.586]
PF.sim	−0.428	
	[−1.266; 0.358]	
cor_{poly}		−0.028
		[−0.865; 0.814]
SD: Dyad	0.18	0.21
R^2	.011	.011
Num. obs.	6892	6892
LOOIC	3148.0	3148.9
BF	1.99	0.94

Bold Null hypothesis value outside the confidence interval.

Table 10: GLMMs predicting diagnostic similarity using similarity in network measures.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**2.661**	**2.657**	**2.567**	**2.518**	**2.613**	**2.612**
	[2.270; 3.077]	[2.333; 2.975]	[2.141; 2.998]	[2.185; 2.881]	[2.192; 3.035]	[2.250; 2.987]
Category 2	−0.140	−0.141	−0.139	−0.145	−0.140	−0.142
	[−0.356; 0.070]	[−0.348; 0.075]	[−0.353; 0.077]	[−0.357; 0.068]	[−0.354; 0.068]	[−0.353; 0.070]
Category 4a	−0.280	−0.283	−0.283	−0.284	−0.284	−0.284
	[−0.813; 0.244]	[−0.810; 0.250]	[−0.790; 0.260]	[−0.805; 0.252]	[−0.814; 0.255]	[−0.809; 0.270]
Category 4b	**0.688**	**0.689**	**0.690**	**0.684**	**0.689**	**0.689**
	[0.260; 1.112]	[0.259; 1.110]	[0.260; 1.113]	[0.273; 1.119]	[0.262; 1.123]	[0.252; 1.095]
Category 5	**1.840**	**1.841**	**1.841**	**1.837**	**1.837**	**1.844**
	[1.188; 2.566]	[1.186; 2.565]	[1.176; 2.546]	[1.149; 2.508]	[1.157; 2.518]	[1.147; 2.550]
ASPL_{SMAPD}	0.032					
	[−0.818; 0.765]					
degree_{SMAPD}		0.098				
		[−0.952; 1.137]				
CC_{SMAPD}			0.307			
			[−0.865; 1.342]			
Q_{SMAPD}				0.229		
				[−0.186; 0.669]		
$\text{c}_{malign;SMAPD}$					0.142	
					[−0.700; 0.938]	
$\text{c}_{benign;SMAPD}$						0.208
						[−0.704; 1.194]
SD: Dyad	0.21	0.22	0.18	0.18	0.20	0.21
R^2	0.011	0.011	0.010	0.011	0.010	0.011
Num. obs.	6892	6892	6892	6892	6892	6892
LOOIC	3149.2	3148.7	3149.1	3148.2	3149.0	3148.7
BF	0.96	1.20	1.82	1.11	1.12	1.24

Bold Null hypothesis value outside the confidence interval.

3.5 Discussion

This study aimed to investigate differences between cognitive networks of expert radiologists and how these relate to real-world diagnoses. It made use of a mixed-methods design based on best practices derived from the literature on team mental models and recent investigations into cognitive network science. In a two step process, experienced radiologists from the German Mammography Screening Program were interviewed to derive decision criteria relevant to mammography reading. Next, these criteria were used to build an online-survey which trained radiologists participated in, deriving cognitive networks from 105 pairwise relatedness-ratings. Survey data and past performance evaluation data, provided by two reference centers of the screening program, were matched and investigated for associations.

Based on a mixture of established quantitative methods from the literature on team mental models (e.g. Lim & Klein, 2006) as well as contemporary measures from cognitive network science (Siew et al., 2019), networks were investigated regarding their similarities and differences. Visual interpretation of network structure showed networks to differ in their most connected nodes as well as number of connections and their strengths. Variances were also observed when examining metrics pertaining to individual nodes. *High* as well as *low density* exhibited the largest variances in the three network measures of path lengths, degree and local clustering. Whether this was due to varying attention to cues between radiologists to consider breast density in diagnostic judgements or due to the reduction from 4 down to 2 levels of density as opposed to the BI-RADS criteria (American College of Radiology, 2013) remains unknown. Comparing metrics between *malign* and *benign*, however, might offer a speculative answer, as *benign* exhibited a similar pattern to the density nodes. The answer depends on how radiologists might consider a mammogram image *a priori* to a classification. Given the fact that the natural frequency of cancer in the general population is much lower than *benign* findings, a radiologist might assume an image to be cancer free and look for cues of malignancy, instead of assuming a neutral image and searching for either cues of *benign* or *malign* findings to formulate a final diagnosis. If the first situation is the case in which the *benign* status-quo is assumed

and *malign* impressions have to reach a certain threshold to change the diagnosis, greater similarity in node metrics between radiologists might be the case, as individuals were more similar in their interpretation of *malign*, which relates more to the practice, compared to relatedness-ratings of *benign*, resulting in different node metrics. In other words, the ratings for the *malign* node might have been more easy and sensible in the eyes of the radiologists, whereas for *benign* the interpretation was less clear, as it may be one of two diagnostic outcomes, just not the outcome radiologists are searching cues for. Should such a hierarchy of node importance be corroborated with the possibility to differ between individuals, the variances of the density nodes might be explained because of varying cue attention between individuals. Further interviews with the sample used in the present study might shed more light onto this result. Still, the present study showed that in addition to previously established methods relying on regularizing network structures to its seemingly most central connections (Langan-Fox et al., 2000), the inclusion of newer methods that do not disregard data as random noise or measurement error is able to provide a more comprehensive picture of how the cognition of radiologists vary. Variances, however, were largely unrelated to actual behavior in diagnostic practice.

Neither the similarity in network structure nor measures derived from cognitive networks were associated with correct diagnoses of mammograms. When using an alternate measure, the average clustering coefficient CC and the local clustering coefficient c_{malign} showed a credible negative association with the amount of deviation points, a count measure of how far off a categorical judgement for an image was compared to the correct target. For studies that associate mental models and task performance as the dependent variable, network clustering remains a yet understudied property. One interpretation of this finding in relation to mammogram diagnostics could be found in the spreading-activation-theory of memory (Anderson, 1983; Collins & Loftus, 1975; Patterson, Nestor, & Rogers, 2007). When encountering a cancer cue (e.g., an architectural distortion) that is represented as a node with connections to other nodes, said cue trigger is able to activate different nodes in a network (e.g., malignant finding, recall the patient). Clustering coefficients describe the degree to

which a network is made out of smaller, more denser sub-networks. If high clustering (i.e., many sub-networks that are densely connected) is negatively associated with a correct decision, it might be due to a person's ability to differentiate between sensible associations between nodes, meaning the decision outcomes to assign benign or malign findings are pertained within their respective networks instead of being triggered by all possible nodes of cancer cues within a larger network, thereby reducing the discriminant ability to differentiate between a benign or malignant observation. In other words, lower clustering might be beneficial, as it stands for the presence of more precisely crafted sub-networks. How this could benefit the screening practice or training of future radiologists, however, remains questionable, as the effects were observed for the non-practically relevant measure of deviation points in addition to being more credibly observed in models with reduced sample size. Future research could investigate this link between clustering and accuracy of diagnoses, regardless of looking at mammograms or different medical diagnostics.

The null-effect of PF similarity on diagnostic performance deserves highlighting as prior mental model studies interested in performance regularly highlighted significant associations with the measure (e.g. Mathieu et al., 2000; Mohammed et al., 2010). Yet, the present study's design and mode of analysis differed from prior installments in several regards. Many of the studies that found significant associations did so with lay people such as students playing video games (e.g. Cooke et al., 2003), while only few investigated real world experts (e.g. Lim & Klein, 2006) like the present study did. The positive evidence may therefore be partly related to differences in tasks and samples, as well as differences in dimensions of cognitive models that were examined. Mental models between individuals may differ in what they contain. While this study exclusively looked at the task related nodes of cancer cues, studies finding significant effects additionally investigate models of teamwork which can relate to the interaction between team members (e.g. Lim & Klein, 2006). Moreover, analyses of multilevel data allow for more robust individual assessments of performance (Harrison et al., 2018), as variability is assessed and controlled for on a per-person level. However, these methods remain unused in the team mental model literature so far. Also, it

should be noted that greater performance variability of individuals in different settings could be an explanation for effects, while the sample used in this study made the correct decision in 94.9% of the cases and the worst performer in 91%.

The poor predictive performance of cognitive networks and associated measures could also be observed when modeling between-diagnostician reliability in dyads since no credible associations were present. In light of the predominance of null results, one sentence by Mathieu and colleagues (2005, p. 40) stands out: "Rather than assume one correct model, we encourage further exploration of the notion that experts may possess multiple heterogeneous models of team-relevant knowledge". While the citation references mental models of teamwork, this might also apply for task related models. If several correct models reside within individuals in a study sample, a singular measure of similarity like PF similarity will not account for the intra- and inter-individual variance. Even though Mohammed and colleagues (2010) declared measurement debates within the literature stream of shared and team mental models to be largely settled, arguments can be made that a renewed discussion concerning measurement and statistical analysis should take place due to two reasons. First, variability in measurement allows for cherry picking methods which are most likely to show an effect as to increase chances of publication (Ritchie, 2020) while also artificially inflating the explanatory prowess of the concept *mental model research*, if a multitude of methods can be arbitrarily said to concern themselves with *mental models*. Second, the literature that compares individuals' models and investigates the association between (dis-)similarities with task performance has so far not designed any robustness checks with regard to the structure of networks and associated measures, although such starting points can be found in cognitive network science. To offer a beginning for a renewed discussion, the following limitation section highlights key aspects for future research using the present study as a cautionary example.

3.6 Limitations of the Literature and the Present Study

The limitations of the present study represent possible shortcomings in the fundamental literature, which provide ample possibilities for future research projects.

For example, future studies could critically examine the selection of nodes considered for statistical analyses. Since this study aimed to associate differences or similarities in networks with the dependent variable of diagnostic performance, its design and procedure followed the best practices in shared and team mental model literature (Langan-Fox et al., 2000; Mohammed et al., 2010). So, a task analysis and interviews with experts revealed a specific set of 15 nodes which are relevant cues in the diagnoses of mammograms. The selection of nodes, however, might have had an effect on the (dis-)similarities between cognitive networks and thus said measures' relation with diagnostic behavior. How might the results have changed, due to either the inclusion or exclusion of nodes? Certainly, this question can be addressed manually, by for example removing the diagnostic outcome categories of *benign* and *malign* and re-running the models. And yet, what rational validates the inclusion or exclusion of specific nodes? In case of the present study, the feasibility to collect data without the option of paying participants was the guiding principle, limiting the selection of nodes to just 15. This also meant disregarding the location of cancer cues within the breast as a possible criterion for malignancy (American College of Radiology, 2013), which would have compounded the number judgements further. Moreover, manual selection of inclusion or exclusion offers the possibility to search for desirable results, p-hacking or hypothesizing after the results are known (Kerr, 1998). Instead, the robustness of results due to varying node selections could be analyzed via simulating varying sizes of networks. Using the example of the 15 nodes used in this study, such a simulation procedure could rely on the following process. The first step would be to vary the selection of nodes in an increasing manner, starting in increments of 1 from 5 to 15. For each of the numbers between 5 and 15, an according number of nodes would be selected from the 15 possibilities at random and transformed into networks using the relatedness-ratings as connections. In the second step, independent variables would be derived to then, in a third step, calculate models to test for associations with dependent variables of interest. This three-step process would be repeated several thousand times and confidence intervals for each network size between 5 and 15 as well as regression estimates would be saved. Averaging results from these repeated runs would

then inform us about the underlying uncertainty associated with every possible network size. In case one detects a region of more stable results, i.e., a network size that is associated with credible results (or significant results, this simulation procedure is impartial to frequentist versus Bayesian statistics) researchers can investigate which selection of nodes within a specific network size is responsible for the effect and, more importantly, use subject matter expertise to interpret whether this specific selection of nodes is sensible or if one is possibly looking at spurious correlations. A similar procedure has been used in cognitive network studies, though it focused on removing connections above or below varying strength-thresholds (Cosgrove, Kenett, Beaty, & Diaz, 2021; Kenett, 2018). Given the fact that a singular model reported in the present study took on average 20 minutes to compute on a system with a 4-core, 2,8 GHz processor with 16 GB working memory, this procedure was out of question without access to more compute power. Switching to frequentists analyses and parallelizing calculations on computer systems with more processing units is likely to reduce calculation times significantly to less than a day.

Limitations should moreover be noted in the treatment of connections between nodes. The notion that a connection contains measurement error or random noise (e.g. Kenett et al., 2021; Lim & Klein, 2006) might not be unwarranted. Partly, this might result from the way they are inferred, if done so using single item Likert-scales. Common practice in psychology often tries to assess latent constructs and reducing measurement error by asking multiple questions on n-point Likert scales that all try to address closely related dimensions of a construct using different phrasings and averaging the measure over the items. Doing so when asking for the relatedness between the nodes in the case of networks such as the ones used in this study would double the burden for participants to partake in such a study and effectively reduce the total number of nodes that become testable within a data collection as time and monetary reimbursements increase. Additional noise between individuals' answers might be introduced with the type of question that is being asked to infer the connection strength between nodes. Most commonly the type of questions ask for either the *similarity* or *relatedness* between nodes (Langan-Fox et al., 2000). While, as in the case of

most kinds of self-reported questions in psychological research, the wording in itself can be understood ambiguously, research could venture onto the inquiry of causal relationships between nodes. To the best of the author's knowledge, this has seen no precedence and was therefore disregarded for the present study, although arguments can be raised in favor of using causal judgements. First, mammogram diagnostics rely on causal relationships between cues. For example, the presence of a tumor of a given size can cause an architectural distortion (American College of Radiology, 2013). Consequently, a question trying to assess the relationship between nodes could therefore read *Does X cause Y?* for a binary measure or even *How strongly does X cause Y?* for approaches of ordinal scaled data and above. Second, asking for a direction of effect (e.g. $X \rightarrow Y$?) allows to view networks with *directed* connections, further increasing the information a network is able to capture from an individual and therefore also generating more statistical ways to interpret and compare cognitive networks (Christensen, 2018; Rubinov & Sporns, 2010). Of course, asking a question that requires one click within a survey per pairwise association between nodes, whether pertaining to causality or not, requires time. An alternative approach, although used to collect participants' perception of relationships between people within a social network, has recently been employed by Son, Bhandari, and FeldmanHall (2021): individuals represented as nodes had to be arranged on the screen, whereas closer arrangements between nodes represented a stronger relationship. Such an operationalization might allow future studies in the cognitive network domain to possibly assess more nodes within a shorter period of time.

Finally, a point of debate concerning both nodes and connections should be confronted regarding possible moderation effects. Synonymously to moderations or interactions between variables in statistical models, conditional hypotheses for the association between nodes are possible. In the realm of mammogram diagnoses, for example, one could hypothesize that *the relatedness between* architectural distortions *and a* malign *diagnoses depends on* breast density (American College of Radiology, 2013). However, the typical mode of eliciting association between nodes relies on bivariate judgements of association strengths and is therefore unable to capture two-

way or even higher-order interaction effects, as seen in the present study. Trying to accommodate for such interactions by creating more questions is likely to cannibalize a questionnaire by increasing attrition and participants might drop out because answering many questions of the following type is dull: *How strongly are X and Y related if A is high? If A is low? If B is met but not A?* It should be of note that, mathematically speaking, the tools for analyzing networks containing interactions are readily being explored in the domain of hypergraph analyses, in which nodes do not relate in pairs but in groups (Neuhäuser, Mellor, & Lambiotte, 2020). As to how far incorporating hypergraph analyses into cognitive network science enriches our understanding of cognition and subsequent behavior remains to be seen but surely presents an opportunity for future research to explore, once network elicitation methods cope with the increase of data an individual has to contribute.

4 Conclusions

The present study set out to model the cognition and diagnostic behavior of expert radiologists screening mammogram images for breast cancer using complementary methods and measures derived from three distinct but related streams of the literature that all fashion human memory to contain networks of connected nodes which are used to guide behavior. A mixed-methods approach aimed to inform the study's design to be ecologically valid in the area of mammogram diagnoses research by interviewing experienced radiologists and building a contemporary quantitative cross-sectional data collection and analyses with the radiologists' knowledge at its foundation. Contrary to previous research, most of the measures derived from comparisons between radiologists' cognitive networks were not associated with diagnostic performance or diagnostic similarity in dyads. The credible negative association between network clustering and deviation points of correct diagnostic category judgements, however, presents an opportunity for future research. Even so, the use of cognitive network science did not advance our understanding of radiologists' cognition and behavior in a meaningful way, as the prevalence of null-results do not warrant a more detailed

and qualitative investigation into specific networks or parameters associated with diagnostically relevant outcomes such as recall decisions or reliability between decision makers.

The major contributions the present study offers for future research projects therefore resides in providing a mixed-methods study design and statistical analysis framework based on a complementary treatment of related literature streams. Additionally, a call for a renewed critical examination of network elicitation methods and conceptualization of cognitive networks both in theory as well as in analyses is presented as a closing note, so that future investigations into complex multi-cue judgement tasks can draw more robust inferences.

5 References

American College of Radiology. (2013). *Bi-rads atlas: Breast imaging reporting and data system - fifth edition.* American College of Radiology.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, *22*(3), 261–295.

Arrow, K. J., Forsythe, R., Gorham, M., Hahn, R., Hanson, R., Ledyard, J. O., ... Zitzewitz, E. (2008). The promise of prediction markets. *Science*, *320*(5878), 877–878.

Baronchelli, A., Ferrer-i Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348–360.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Bostrom, A., Fischhoff, B., & Morgan, M. G. (1992). Characterizing mental models of hazardous processes: A methodology and an application to radon. *Journal of Social Issues*, *48*(4), 85–100.

Bostrom, A., Morgan, M. G., Fischhoff, B., & Read, D. (1994). What do people know about global climate change? 1. mental models. *Risk Analysis*, *14*(6), 959–970.

Bruine de Bruin, W., & Bostrom, A. (2013). Assessing what to address in science communication. *Proceedings of the National Academy of Sciences*, *110*(Supplement 3), 14062–14068.

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1).

Budescu, D. V., & Chen, E. (2015). Identifying expertise to extract the wisdom of crowds. *Management Science*, *61*(2), 267–280.

Bundesministerium für Gesundheit. (2016). *Mammographie-screening.* Retrieved 2021-05-30, from https://www.bundesgesundheitsministerium.de/

service/begriffe-von-a-z/m/mammographie-screening.html

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bürkner, P.-C. (2018). Advanced Bayesian Multilevel Modeling with the R Package brms. *The R Journal*, *10*(1), 395–411. Retrieved from https://doi.org/10.32614/RJ-2018-017 doi: 10.32614/RJ-2018-017

Bürkner, P.-C., & Vuorre, M. (2019). Ordinal regression models in psychology: A tutorial. *Advances in Methods and Practices in Psychological Science*, *2*(1), 77–101.

Cannon-Bowers, J. A., Salas, E., & Converse, S. (1993). Shared mental models in expert team decision making. *Individual and Group Decision Making: Current Issues.*, 221–246.

Carney, P. A., Bogart, T. A., Geller, B. M., Haneuse, S., Kerlikowske, K., Buist, D. S., ... others (2012). Association between time spent interpreting, level of confidence, and accuracy of screening mammography. *American Journal of Roentgenology*, *198*(4), 970–978.

Castro, N., & Siew, C. S. (2020). Contributions of modern network science to the cognitive sciences: revisiting research spirals of representation and process. *Proceedings of the Royal Society A*, *476*(2238), 20190825.

Christensen, A. P. (2018). NetworkToolbox: Methods and measures for brain, cognitive, and psychometric network analysis in R. *The R Journal*, 422–439. doi: 10.32614/RJ-2018-065

Christensen, A. P., & Kenett, Y. N. (2019). Semantic network analysis (SemNA): A tutorial on preprocessing, estimating, and analyzing semantic networks. *PsyArXiv*. doi: 10.31234/osf.io/eht87

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407.

Cooke, N. J., Gorman, J. C., Duran, J. L., & Taylor, A. R. (2007). Team cognition in experienced command-and-control teams. *Journal of Experimental Psychology: Applied*, *13*(3), 146.

Cooke, N. J., Kiekel, P. A., Salas, E., Stout, R., Bowers, C., & Cannon-Bowers, J. (2003). Measuring team knowledge: A window to the cognitive underpinnings of team performance. *Group Dynamics: Theory, Research, and Practice*, *7*(3), 179.

Cosgrove, A. L., Kenett, Y. N., Beaty, R. E., & Diaz, M. T. (2021). Quantifying flexibility in thought: The resiliency of semantic networks differs across the lifespan. *Cognition*, *211*, 104631.

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal*, *Complex Systems*, 1695. Retrieved from https://igraph.org

Davis-Stober, C. P., Budescu, D. V., Dana, J., & Broomell, S. B. (2014). When is a crowd wise? *Decision*, *1*(2), 79.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). Measuring shared team mental models: A meta-analysis. *Group Dynamics: Theory, Research, and Practice*, *14*(1), 1.

Edwards, B. D., Day, E. A., Arthur Jr, W., & Bell, S. T. (2006). Relationships among team ability composition, team mental models, and team performance. *Journal of Applied Psychology*, *91*(3), 727.

Elmore, J. G., Wells, C. K., & Howard, D. H. (1998). Does diagnostic accuracy in mammography depend on radiologists' experience? *Journal of Women's Health*, *7*(4), 443–449.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212.

Epskamp, S., & Fried, E. I. (2018). A tutorial on regularized partial correlation networks. *Psychological Methods*, *23*(4), 617.

Esserman, L., Cowley, H., Eberle, C., Kirkpatrick, A., Chang, S., Berbaum, K., & Gale, A. (2002). Improving the accuracy of mammography: volume and outcome relationships. *Journal of the National Cancer Institute*, *94*(5), 369–375.

Fisher, D. M., Bell, S. T., Dierdorff, E. C., & Belohlav, J. A. (2012). Facet personality

and surface-level diversity as team mental model antecedents: implications for implicit coordination. *Journal of Applied Psychology*, *97*(4), 825.

Gardner, A. K., Scott, D. J., & AbdelFattah, K. R. (2017). Do great teams think alike? an examination of team mental models and their impact on team performance. *Surgery*, *161*(5), 1203–1208.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: models of bounded rationality. *Psychological Review*, *103*(4), 650.

Gorman, J. C., & Cooke, N. J. (2011). Changes in team cognition after a retention interval: the benefits of mixing it up. *Journal of Experimental Psychology: Applied*, *17*(4), 303.

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E., ... Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ*, *6*, e4794.

Hodgson, R. T., et al. (2008). An examination of judge reliability at a major us wine competition. *Journal of Wine Economics*, *3*(2), 105–113.

Jeffreys, H. (1939). *Theory of probability.* Oxford Univ. Press.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness* (No. 6). Harvard University Press.

Jungermann, H., Schütz, H., & Thüring, M. (1988). Mental models in risk assessment: Informing people about drugs. *Risk Analysis*, *8*(1), 147–155.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kahneman, D., Rosenfield, A., Gandhi, L., & Blaser, T. (2016). Noise. *Harvard Business Review*, 38–46.

Kahneman, D., Sibony, O., & Sunstein, C. (2021). *Noise: a flaw in human judgment.*

William Collins Publishers.

Kämmer, J. E., Hautz, W. E., Herzog, S. M., Kunina-Habenicht, O., & Kurvers, R. H. (2017). The potential of collective intelligence in emergency medicine: pooling medical students' independent decisions improves diagnostic performance. *Medical Decision Making*, *37*(6), 715–724.

Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science*, *26*(1), 1.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kassenärztliche Bundesvereinigung. (2021). *Vereinbarung von qualitätssicherungsmaßnahmen nach § 135 abs. 2 sgb v zur kurativen mammographie (mammographie-vereinbarung)*. Kassenärztliche Bundesvereinigung. Retrieved from https://www.kbv.de/media/sp/Mammographie.pdf

Kellermanns, F. W., Floyd, S. W., Pearson, A. W., & Spencer, B. (2008). The contingent effect of constructive confrontation on the relationship between shared mental models and decision quality. *Journal of Organizational Behavior*, *29*(1), 119–137.

Kenett, Y. N. (2018). Investigating creativity from a semantic network perspective. In Z. Kapoula, E. Volle, J. Renoult, & M. Andreatta (Eds.), *Exploring transdisciplinarity in art and sciences* (pp. 49–75). Springer Verlag.

Kenett, Y. N., Ungar, L., & Chatterjee, A. (2021). Beauty and wellness in the semantic memory of the beholder. *Frontiers in Psychology*, *12*.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Key, T. J., Verkasalo, P. K., & Banks, E. (2001). Epidemiology of breast cancer. *The Lancet Oncology*, *2*(3), 133–140.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573.

Kruschke, J. K. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, *25*(1), 155–177.

Kumle, L., Vo, M. L.-H., & Draschkow, D. (2021). Estimating power in (generalized) linear mixed models: An open introduction and tutorial in r. *Behavior Research Methods*, 1–16.

Kurvers, R. H., De Zoete, A., Bachman, S. L., Algra, P. R., & Ostelo, R. (2018). Combining independent decisions increases diagnostic accuracy of reading lumbosacral radiographs and magnetic resonance imaging. *PloS one*, *13*(4), e0194128.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Carney, P. A., Bogart, A., ... Wolf, M. (2016). Boosting medical diagnostics by pooling independent judgments. *Proceedings of the National Academy of Sciences*, *113*(31), 8777–8782.

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., Moussaid, M., Argenziano, G., ... Wolf, M. (2019). How to detect high-performing individuals and groups: Decision similarity predicts accuracy. *Science Advances*, *5*(11).

Kurvers, R. H., Herzog, S. M., Hertwig, R., Krause, J., & Wolf, M. (2021). Pooling decisions decreases variation in response bias and accuracy. *iScience*, 102740.

Kurvers, R. H., Krause, J., Argenziano, G., Zalaudek, I., & Wolf, M. (2015). Detection accuracy of collective intelligence assessments for skin cancer diagnosis. *JAMA Dermatology*, *151*(12), 1346–1353.

Langan-Fox, J., Code, S., & Langfield-Smith, K. (2000). Team mental models: Techniques, methods, and analytic approaches. *Human Factors*, *42*(2), 242–271.

Langan-Fox, J., Wirth, A., Code, S., Langfield-Smith, K., & Wirth, A. (2001). Analyzing shared and team mental models. *International journal of Industrial Ergonomics*, *28*(2), 99–112.

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology*, *79*,

328–348.

Lim, B.-C., & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior*, *27*(4), 403–418.

Litvinova, A., Kurvers, R. H., Hertwig, R., & Herzog, S. M. (2019, Aug). *When do experts make inconsistent decisions?* PsyArXiv. Retrieved from psyarxiv.com/dtaz3 doi: 10.31234/osf.io/dtaz3

Makowski, D., Ben-Shachar, M. S., & Lüdecke, D. (2020). The easystats collection of r packages. *GitHub*. Retrieved from https://github.com/easystats/easystats

Mannes, A. E., Soll, J. B., & Larrick, R. P. (2014). The wisdom of select crowds. *Journal of Personality and Social Psychology*, *107*(2), 276.

Marko, M., & Riečanskỳ, I. (2021). The structure of semantic representation shapes controlled semantic retrieval. *Memory*, 1–9.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Cannon-Bowers, J. A., & Salas, E. (2005). Scaling the quality of teammates' mental models: Equifinality and normative comparisons. *Journal of Organizational Behavior*, *26*(1), 37–56.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*(2), 273.

Mathieu, J. E., Leslie, J. B., & Luciano, M. M. (2021). Wrapping team members' heads around managing virtual team-related paradoxes. In M. McNeese, E. Salas, & M. Endsley (Eds.), *Fields of practice and applied solutions within distributed team cognition* (pp. 1–20). CRC Press.

McElreath, R. (2018). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman and Hall/CRC.

Meyer, D., Leventhal, H., & Gutmann, M. (1985). Common-sense models of illness: the example of hypertension. *Health Psychology*, *4*(2), 115.

Mohammed, S., Ferzandi, L., & Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, *36*(4),

876–910.

Mohammed, S., Klimoski, R., & Rentsch, J. R. (2000). The measurement of team mental models: We have no shared schema. *Organizational Research Methods*, *3*(2), 123–165.

Molins, E., Macià, F., Ferrer, F., Maristany, M.-T., & Castells, X. (2008). Association between radiologists' experience and accuracy in interpreting screening mammograms. *BMC Health Services Research*, *8*(1), 1–10.

Morgan, M. G., Fischhoff, B., Bostrom, A., Atman, C. J., et al. (2002). *Risk communication: A mental models approach*. Cambridge University Press.

Neuhäuser, L., Mellor, A., & Lambiotte, R. (2020). Multibody interactions and nonlinear consensus dynamics on networked systems. *Physical Review E*, *101*(3), 032310.

Norman, D. A. (1983). Some observations on mental models. *Mental models*, *7*(112), 7–14.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443–460.

Onnela, J.-P., Saramäki, J., Kertész, J., & Kaski, K. (2005). Intensity and coherence of motifs in weighted complex networks. *Physical Review E*, *71*(6), 065103.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987.

Prelec, D., Seung, H. S., & McCoy, J. (2017). A solution to the single-question crowd wisdom problem. *Nature*, *541*(7638), 532–535.

Quirin, A., Cordón, O., Guerrero-Bote, V. P., Vargas-Quesada, B., & Moya-Anegón, F. (2008). A quick mst-based algorithm to obtain pathfinder networks (, n-1). *Journal of the American Society for Information Science and Technology*, *59*(12), 1912–1924.

Rafferty, E. A., Park, J. M., Philpotts, L. E., Poplack, S. P., Sumkin, J. H., Halpern, E. F., & Niklason, L. T. (2013). Assessing radiologist performance using combined digital mammography and breast tomosynthesis compared with digital

mammography alone: results of a multicenter, multireader trial. *Radiology*, *266*(1), 104–113.

Rafferty, E. A., Park, J. M., Philpotts, L. E., Poplack, S. P., Sumkin, J. H., Halpern, E. F., & Niklason, L. T. (2014). Diagnostic accuracy and recall rates for digital mammography and digital mammography combined with one-view and two-view tomosynthesis: results of an enriched reader study. *American Journal of Roentgenology*, *202*(2), 273–281.

Rigdon, E. E., & Ferguson Jr, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in confirmatory factor analysis with ordinal data. *Journal of Marketing Research*, *28*(4), 491–497.

Ritchie, S. (2020). *Science fictions: Exposing fraud, bias, negligence and hype in science*. Random House.

Rouse, W. B., & Morris, N. M. (1986). On looking into the black box: Prospects and limits in the search for mental models. *Psychological Bulletin*, *100*(3), 349.

RStudio Team. (2021). Rstudio: Integrated development environment for r [Computer software manual]. Boston, MA. Retrieved from http://www.rstudio.com/

Rubinov, M., & Sporns, O. (2010). Complex network measures of brain connectivity: uses and interpretations. *Neuroimage*, *52*(3), 1059–1069.

Santos, C. M., Uitdewilligen, S., & Passos, A. M. (2015). A temporal common ground for learning: The moderating effect of shared mental models on the relation between team learning behaviours and performance improvement. *European Journal of Work and Organizational Psychology*, *24*(5), 710–725.

Schad, D. J., Nicenboim, B., Bürkner, P.-C., Betancourt, M., & Vasishth, S. (2021). Workflow techniques for the robust use of bayes factors.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature neuroscience*, *16*(4), 486–492.

Schapiro, A. C., Turk-Browne, N. B., Norman, K. A., & Botvinick, M. M. (2016). Statistical learning of temporal community structure in the hippocampus. *Hippocampus*, *26*(1), 3–8.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*(6), 1402–1412.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Ablex Publishing.

Seligman, M. E., Railton, P., Baumeister, R. F., & Sripada, C. (2016). *Homo prospectus.* Oxford University Press.

Siew, C. S. (2013). Community structure in the phonological network. *Frontiers in Psychology*, *4*, 553.

Siew, C. S. (2018). The orthographic similarity structure of english words: Insights from network science. *Applied Network Science*, *3*(1), 1–18.

Siew, C. S., Wulff, D. U., Beckage, N. M., & Kenett, Y. N. (2019). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, *2019*.

Simon, H. A. (1955). A behavioral model of rational choice. *The Quarterly Journal of Economics*, *69*(1), 99–118.

Smith-Jentsch, K. A., Mathieu, J. E., & Kraiger, K. (2005). Investigating linear and interactive effects of shared mental models on safety and efficiency in a field setting. *Journal of Applied Psychology*, *90*(3), 523-535.

Son, J.-Y., Bhandari, A., & FeldmanHall, O. (2021). Cognitive maps of social features enable flexible inference in social networks. *Proceedings of the National Academy of Sciences*, *118*(39).

Stella, M., De Nigris, S., Aloric, A., & Siew, C. S. (2019). Forma mentis networks quantify crucial differences in stem perception between students and experts. *PloS one*, *14*(10), e0222870.

Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors*, *41*(1), 61–71.

Tannenbaum, S. I., Traylor, A. M., Thomas, E. J., & Salas, E. (2021). Managing teamwork in the face of pandemic: evidence-based tips. *BMJ Quality & Safety*, *30*(1), 59–63.

Tesler, R., Mohammed, S., Hamilton, K., Mancuso, V., & McNeese, M. (2018). Mirror, mirror: guided storytelling and team reflexivity's influence on team mental models. *Small Group Research*, *49*(3), 267–305.

Tetlock, P. E., & Gardner, D. (2016). *Superforecasting: The art and science of prediction*. Random House.

Tetlock, P. E., Mellers, B. A., Rohrbaugh, N., & Chen, E. (2014). Forecasting tournaments: Tools for increasing transparency and improving the quality of debate. *Current Directions in Psychological Science*, *23*(4), 290–295.

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, *66*(8), 1352–1362.

Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of Memory and Language*, *67*(1), 30–44.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, *14*(5), 779–804.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... others (2018). Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57.

Watson, C. G. (2020). braingraph: Graph theory analysis of brain mri data [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=brainGraph (R package version 3.0.0)

Winters, S., Martin, C., Murphy, D., & Shokar, N. K. (2017). Breast cancer epidemiology, prevention, and screening. In R. Lakshmanaswamy (Ed.), *Approaches to understanding breast cancer* (Vol. 151, p. 1-32). Academic Press.

Wood, M. D., Bostrom, A., Bridges, T., & Linkov, I. (2012). Cognitive mapping tools: review and risk management needs. *Risk Analysis*, *32*(8), 1333–1348.

Wulff, D. U., Hills, T., & Mata, R. (2021, June). *Structural differences in the semantic networks of younger and older adults*. PsyArXiv. Retrieved from psyarxiv.com/s73dp doi: 10.31234/osf.io/s73dp

6 Appendix

6.1 Appendix A: Interviews

Date: 25.03.2021 (second interview)
Interviewees

- Name redacted
- Name redacted

Intro:
This interview is conducted as part of my master's thesis. The working hypothesis to be verified is that similarity of mental models is associated with accuracy in mammography findings. Mental models represent structured knowledge in mind of each individual person. The assumption associated with similarity is that as expertise increases, expert perceptions become more similar and, accordingly, so does accuracy.

Mental models can be elicited by having study participants rate the similarity or coherence of individual concepts. In the context of mammogram readings, we have decision making based on visual indicators, which may be more or less decisive for a cancer diagnosis. One such visual indicator would be, for example, an architectural distortion.

1. Question
The purpose of this interview is to collect such visual indicators. In the following, I would like to ask you to list very simply which indicators you look for when making a diagnosis.

- Update on the preliminary scan, if there is something new then it is always to be noted
- Changes behind or above the glandular tissue
- Focal finding, visible on two levels
- Compression
- Region of the finding where glandular tissue may not be expected
- Individual appearance of lesions
- What does the focal finding look like, contour criteria / structure
- Number of foci
- The smaller the more likely not to be seen in the second plane / image
- Any lesion with mircocalcification, microcalcification in adipose tissue
- Microcalcification associated with density elevation
- Microcalcification unilateral
- Suspicions surrounding milk ducts is pulled out immediately
- Depending on breast density, look at other areas, structures more difficult to recognize at high density, rather wave to because dense lesions can be hidden
- Contour analysis of the glandular body, longer observation time, the denser the more dependent on indirect signs
- Cysts
- Age of patient, although not very important
- The more experience a radiologist has, the less he/she is to classify findings in accordance with the BI-RADS system. Intuition takes over.
- Smooth boundary focal lesion is clearly a carcinoma
- BIRADS is not everything, with BIRADS suspicions would also fail, strict interpretation of BIRADS is biased towards benign findings and hence no recall of patients
- Experts prefer to be hypersensitive.
- In case of false positive: minimally invasive procedures are not dramatic.

- Finders who also perform biopsies have a higher specification and sensitivity.
- Decision is not between BIRADS categories but simply between inviting or not inviting. BIRADS however is well known and training focusses a lot on BIRADS

2. Question
Thank you very much. Now I would like to talk to you about individual findings that are recorded in the BI-RADS (Breast Imaging Reporting & Data System of the ACR. These include: Are these indicators important for reporting? (importance is denoted with an x, if mentioned by the interviewees)

- Parenchymal density / breast density x
- Mass x
- Calcifications x
- Architectural disturbance x
- Asymmetries x
- Intramammary lymph node x
- Skin lesion x
- Solitary dilated duct x
- Skin retraction x
- Mammillary retraction x
- Skin thickening x
- Axillary lymphadenopathy / enlargement of axillary lymph nodes x

More comments on BIRADS by the interviewees:

- Diagnosticians should look very sensitively, BIRADS 2 is also sometimes pushed into the consensus conference, 50 to 80 images in the consensus conference, of which 60 to 70 percent are not recalled
- Depending on the prevalence of the last days, different decisions are made in conferences and findings are made until there are (statistically) sufficient cases.
- BIRADS not everything but more structured than before, clear progress
- At the beginning of *name redacted* training, BIRADS was almost exclusively about the locality of an abnormality.

3. Question
Thank you very much. At this point, we can conclude the part about the indicators and there remains one last question that I would like to discuss with you. It concerns the survey itself, in which the concepts we just discussed are to be evaluated in terms of their association.

Test 1
Which question is easier for you to answer?
Either: How similar are masses and calcifications?
Or: How related are masses and calcifications related? (easier)

Test 2 (NOTE: "RELATED" has contextually dependent meanings. In the German original of this interview, the verbs that were compered were "verwandt" and "zusammenhängen". The latter was chosen for the survey.
Either: How related are masses and calcifications?
Or: How related are masses and calcifications?

6.2 Appendix B: Survey

Herzlich Willkommen!

Danke, dass Sie an dieser Umfrage im Rahmen meiner Masterarbeit teilnehmen. Die Studie befasst sich mit Determinanten von diagnostischer Genauigkeit im Mammographiescreening.

Ich weise darauf hin, dass alle Ihre Angaben streng vertraulich verwendet werden. Alle Angaben werden ausschließlich für wissenschaftliche Zwecke verwendet. Sollten die Ergebnisse publiziert werden, werden die Daten **ausschließlich pseudonymisiert verwertet** und auf **aggregierter Ebene ohne Rückschlussmöglichkeit auf Individuuen** ausgewertet.

Ich verpflichte mich nach dem Matching der IDs mit den Daten der Fallsammlungsprüfung aufseiten des Referenzzentrums alle Daten zu löschen, welche die Pseudonymisierung rückgängig machen könnten.

Die Umfrage wird ca. 15 Minuten in Anspruch nehmen.

Herzlichen Dank und viel Spaß,

Julian Berger

Wenn Sie dem zustimmen und zur Teilnahme an der Umfrage fortfahren möchten, geben Sie bitte Ihre **persönliche ID** ein, die Sie der Email erhalten haben.

Im Sinne der Pseudonymisierung bitte ich Sie, an dieser Stelle explizit ihr Einverständnis zu geben:

Ich bestätige, dass das Referenzzentrum Berlin die Daten der Einzelentscheidungen meiner zuletzt abgelegten Fallsammlungsprüfung (Bewertungskategorien mit Richtigkeitsabgleich) inklusiv der von mir oben angegebenen ID an Herrn Julian Berger senden kann, damit er diese im Rahmen der vorliegenden Studie untersuchen darf.

Ich stimme zu. (Hier bitte klicken.)

Weiter

Wie angekündigt, geht es in dieser Studie um Determinanten diagnostischer Genauigkeit im Mammographiescreening. Auf der folgenden Seite werden Sie dazu gebeten, Zusammenhänge zu bewerten. Die genaue Fragestellung wird auf der nächsten Seite erläutert.

Um sich kurz mit der Umgebung vertraut zu machen, finden Sie hier drei Zusammenhänge zum Test. Diese werden nicht in der Studie verwertet.

Wie beurteilen Sie den folgenden Zusammenhang?

Kein Zusammen-hang	Geringer Zusammen-hang	Moderater Zusammen-hang	Starker Zusammen-hang	Sehr starker Zusammen-hang

Weiter

Bei der Befundung von Mammographien lassen sich unterschiedliche visuelle Merkmale erkennen und für Handlungsempfehlungen nutzen. Im Folgenden sind Sie gebeten, den Zusammenhang solcher Merkmale untereinander und mit Handlungsempfehlungen zu bewerten. Hier gibt es **kein richtig und kein Falsch**, beantworten Sie einfach, wie es sich für Sie am besten anfühlt.

Wie beurteilen Sie den folgenden Zusammenhang?

Kein Zusammen-hang	Geringer Zusammen-hang	Moderater Zusammen-hang	Starker Zusammen-hang	Sehr starker Zusammen-hang

Weiter

6.3 Appendix C: Statistical Tests

Table 11: GLMMs predicting diagnostic accuracy using measures of similarity in network structures (only most recent case collection exams).

	Model 1	Model 2
Intercept	**3.453**	**2.974**
	[2.089; 4.722]	[1.496; 4.630]
Category 2	−0.356	−0.356
	[−1.037; 0.320]	[−1.037; 0.284]
Category 4a	292.296	275.544
	[−1.455; 1086.906]	[−1.917; 989.100]
Category 4b	0.871	0.848
	[−0.641; 2.569]	[−0.650; 2.564]
Category 5	**34.942**	**32.073**
	[0.298; 122.081]	[0.384; 109.541]
PF.sim	−0.648	
	[−5.352; 3.855]	
cor_{poly}		1.024
		[−3.082; 4.906]
SD: ID	0.35	0.35
SD: Exam	0.43	0.40
R^2	0.017	0.017
Num. obs.	1100	1100
LOOIC	326.7	325.9
BF	6.57	6.25

Bold Null hypothesis value outside the confidence interval.

Table 12: GLMMs predicting diagnostic accuracy using network measures (only most recent case collection exams).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**3.773**	**4.341**	2.227	**3.314**	**2.596**	**2.280**
	[2.215; 5.400]	[0.908; 8.160]	[−0.005; 4.469]	[2.279; 4.445]	[0.790; 4.326]	[0.228; 4.323]
Category 2	−0.360	−0.351	−0.357	−0.360	−0.356	−0.380
	[−1.018; 0.339]	[−1.046; 0.323]	[−1.012; 0.347]	[−1.038; 0.305]	[−1.031; 0.323]	[−1.073; 0.306]
Category 4a	281.043	393.758	496.343	287.936	295.587	298.791
	[−1.658; 1011.036]	[−1.547; 1255.286]	[−1.485; 1328.447]	[−2.234; 1051.112]	[−1.669; 1071.489]	[−1.712; 1068.442]
Category 4b	0.887	0.868	0.874	0.886	0.870	0.865
	[−0.687; 2.580]	[−0.657; 2.557]	[−0.713; 2.486]	[−0.700; 2.543]	[−0.638; 2.477]	[−0.693; 2.514]
Category 5	**32.350**	**47.553**	**46.586**	**34.480**	**36.357**	**34.543**
	[0.502; 108.290]	[0.366; 141.449]	[0.147; 159.533]	[0.281; 118.871]	[0.309; 127.882]	[0.347; 120.863]
ASPL	−0.725					
	[−2.949; 1.350]					
degree		−0.085				
		[−0.375; 0.209]				
CC			0.662			
			[−0.650; 1.934]			
Q				−0.218		
				[−14.670; 15.687]		
c_{malign}					0.359	
					[−0.439; 1.206]	
c_{benign}						0.678
						[−0.573; 1.925]
SD: ID	0.34	0.33	0.31	0.35	0.32	0.30
SD: Exam	0.42	0.44	0.41	0.43	0.43	0.41
R^2	0.017	0.017	0.018	0.017	0.018	0.018
Num. obs.	1100	1100	1100	1100	1100	1100
LOOIC	325.9	326.4	325.3	326.3	325.5	325.0
BF	3.91	0.49	1.56	21.7	1.75	3.10

Bold Null hypothesis value outside the confidence interval.

Table 13: GLMMs predicting diagnostic accuracy using similarity in network measures (only most recent case collection exams).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**3.244**	**3.269**	**3.307**	**3.056**	**3.364**	**2.593**
	[1.385; 4.995]	[2.061; 4.593]	[1.653; 5.084]	[1.550; 4.569]	[1.392; 5.378]	[0.541; 4.687]
Category 2	−0.367	−0.357	−0.363	−0.362	−0.352	−0.360
	[−1.016; 0.338]	[−1.084; 0.299]	[−1.029; 0.338]	[−1.066; 0.314]	[−1.032; 0.335]	[−1.055; 0.323]
Category 4a	284.889	284.681	310.102	283.575	290.295	289.333
	[−1.324; 1041.851]	[−1.262; 968.612]	[−1.639; 1151.218]	[−1.555; 971.363]	[−1.778; 1067.548]	[−1.331; 1031.736]
Category 4b	0.851	0.869	0.875	0.905	0.896	0.866
	[−0.611; 2.563]	[−0.626; 2.598]	[−0.661; 2.568]	[−0.654; 2.644]	[−0.621; 2.581]	[−0.607; 2.556]
Category 5	**38.429**	**32.362**	**48.175**	**34.387**	**36.296**	**33.567**
	[0.267; 138.176]	[0.533; 109.630]	[0.308; 157.723]	[0.582; 113.923]	[0.222; 123.011]	[0.247; 114.029]
$ASPL_{SMAPD}$	0.173					
	[−4.705; 5.289]					
$degree_{SMAPD}$		0.227				
		[−6.253; 6.967]				
CC_{SMAPD}			0.000			
			[−6.556; 6.235]			
Q_{SMAPD}				0.447		
				[−1.815; 2.733]		
$c_{malign;SMAPD}$					−0.205	
					[−5.599; 5.268]	
$c_{benign;SMAPD}$						2.408
						[−3.996; 9.224]
SD: ID	0.36	0.36	0.37	0.35	0.36	0.33
SD: Exam	0.43	0.44	0.43	0.43	0.43	0.39
R^2	0.017	0.017	0.016	0.017	0.017	0.017
Num. obs.	1100	1100	1100	1100	1100	1100
LOOIC	326.6	326.9	326.7	326.6	326.4	326.5
BF	6.13	8.01	8.19	3.03	7.29	11.6

Bold Null hypothesis value outside the confidence interval.

Table 14: GLMMs predicting deviation points using measures of similarity in network structures (only most recent case collection exams).

	Model 1	Model 2
Intercept	−0.794	−0.049
	[−1.657; 0.016]	[−0.943; 0.917]
Category 2	0.105	0.108
	[−0.161; 0.406]	[−0.166; 0.388]
Category 4a	**−195.066**	**−194.588**
	[−490.302; −1.555]	[−493.289; −2.146]
Category 4b	−0.423	−0.429
	[−0.913; 0.055]	[−0.892; 0.060]
Category 5	**−0.641**	**−0.639**
	[−1.048; −0.235]	[−1.052; −0.232]
PF.sim	0.357	
	[−2.278; 3.264]	
cor_{poly}		−2.130
		[−4.965; 0.388]
SD: ID	0.43	0.32
SD: Exam	0.26	0.23
R^2	0.036	0.037
Num. obs.	1100	1100
LOOIC	1925.7	1925.3
BF	3.88	18.42

Bold Null hypothesis value outside the confidence interval.

Table 15: GLMMs predicting deviation points using network measures (only most recent case collection exams).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**−1.420**	−0.627	0.679	**−0.904**	0.305	−0.055
	[−2.306; −0.538]	[−3.127; 1.818]	[−0.378; 1.744]	[−1.630; −0.245]	[−0.661; 1.184]	[−1.324; 1.278]
Category 2	0.101	0.109	0.106	0.107	0.108	0.113
	[−0.177; 0.383]	[−0.175; 0.382]	[−0.178; 0.370]	[−0.164; 0.395]	[−0.166; 0.393]	[−0.157; 0.390]
Category 4a	**−197.600**	**−200.975**	**−200.046**	**−201.182**	**−203.769**	**−202.323**
	[−516.742; −0.784]	[−517.071; −2.003]	[−510.763; −1.679]	[−528.257; −1.813]	[−535.466; −1.622]	[−540.918; −1.138]
Category 4b	−0.431	−0.424	−0.432	−0.428	−0.429	−0.427
	[−0.917; 0.049]	[−0.890; 0.057]	[−0.908; 0.046]	[−0.885; 0.070]	[−0.897; 0.052]	[−0.901; 0.058]
Category 5	**−0.648**	**−0.640**	**−0.653**	**−0.642**	**−0.652**	**−0.644**
	[−1.067; −0.262]	[−1.048; −0.232]	[−1.059; −0.248]	[−1.050; −0.241]	[−1.057; −0.237]	[−1.040; −0.240]
ASPL	1.149					
	[−0.120; 2.359]					
degree		−0.007				
		[−0.206; 0.197]				
CC			**−0.841**			
			[−1.443; −0.225]			
Q				3.874		
				[−6.306; 14.089]		
c_{malign}					**−0.500**	
					[−0.899; −0.091]	
c_{benign}						−0.420
						[−1.237; 0.347]
SD: ID	0.29	0.43	0.20	0.40	0.23	0.37
SD: Exam	0.27	0.25	0.23	0.27	0.28	0.25
R^2	0.036	0.037	0.036	0.036	0.036	0.036
Num. obs.	1100	1100	1100	1100	1100	1100
LOOIC	1924.9	1925.6	1922.6	1925.6	1923.7	1925.7
BF	11.06	0.25	29.09	15.19	11.76	2.01

Bold Null hypothesis value outside the confidence interval.

Table 16: GLMMs predicting deviation points using similarity in network measures (only most recent case collection exams).

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Intercept	**−1.101**	−0.756	−1.064	−0.690	**−1.330**	−1.031
	[−2.191; −0.045]	[−1.533; 0.037]	[−2.196; 0.085]	[−1.767; 0.247]	[−2.488; −0.313]	[−2.417; 0.468]
Category 2	0.105	0.108	0.108	0.109	0.106	0.107
	[−0.172; 0.378]	[−0.163; 0.391]	[−0.157; 0.388]	[−0.166; 0.398]	[−0.173; 0.382]	[−0.182; 0.378]
Category 4a	**−200.845**	**−190.374**	**−199.853**	**−198.273**	**−191.156**	**−192.866**
	[−526.888; −1.653]	[−476.683; −2.028]	[−527.743; −1.638]	[−510.145; −0.775]	[−488.059; −1.333]	[−492.569; −1.618]
Category 4b	−0.429	−0.425	−0.425	−0.424	−0.431	−0.424
	[−0.908; 0.054]	[−0.905; 0.043]	[−0.902; 0.056]	[−0.872; 0.048]	[−0.913; 0.045]	[−0.904; 0.045]
Category 5	**−0.643**	**−0.639**	**−0.638**	**−0.638**	**−0.644**	**−0.643**
	[−1.041; −0.231]	[−1.043; −0.235]	[−1.025; −0.225]	[−1.033; −0.212]	[−1.050; −0.247]	[−1.042; −0.232]
$ASPL_{SMAPD}$	1.195					
	[−1.852; 4.172]					
$degree_{SMAPD}$		0.324				
		[−3.913; 4.344]				
CC_{SMAPD}			1.473			
			[−2.847; 5.686]			
Q_{SMAPD}				−0.040		
				[−1.634; 1.503]		
$c_{malign;SMAPD}$					1.869	
					[−1.075; 4.836]	
$c_{benign;SMAPD}$						1.061
						[−3.275; 5.599]
SD: ID	0.39	0.43	0.40	0.43	0.35	0.41
SD: Exam	0.25	0.27	0.27	0.27	0.27	0.26
R^2	0.036	0.036	0.037	0.036	0.037	0.037
Num. obs.	1100	1100	1100	1100	1100	1100
LOOIC	1925.4	1925.7	1925.8	1925.9	1926.0	1925.7
BF	5.29	5.08	8.26	1.93	10.85	6.86

Bold Null hypothesis value outside the confidence interval.