university of groningen / faculty of mathematics and natural sciences / artificial intelligence

Clinical reasoning in ambulance nurses

The influence of experience on confirmation bias

Han Havinga

August 2023

Master Thesis

Artificial Intelligence

University of Groningen, The Netherlands

Internal supervisor:

Dr. Fokie Cnossen (UHD Artificial Intelligence, University of Groningen)

External supervisor:

Bert Dercksen (Medisch Manager Ambulancezorg, UMCG)

Abstract

Clinical reasoning is an important skill in the work of ambulance nurses. Improving clinical reasoning can lead to better healthcare outcomes. Clinical decision support systems can aid ambulance nurses in their clinical reasoning. To develop better clinical decision support systems it is necessary to understand how clinical reasoning works and how it is impaired. One of the detriments to clinical reasoning is confirmation bias.
This thesis aims to determine whether confirmation bias occurs more often in novice ambulance nurses than in experienced ambulance nurses. To this end a simulated patient encounter was constructed of two medical cases. The analysis of these patient encounters shows that novice ambulance nurses were not more likely to suffer from confirmation bias than experienced ambulance nurses.
These findings indicate that confirmation bias is not a major detriment to the clinical reasoning of novice ambulance nurses, and that it would be more efficient to develop clinical decision support systems that are focused on other obstacles in clinical reasoning.

Table of contents

1 Introduction

In the Netherlands the threshold for calling the emergency number has lowered over the years. This causes patients to call the emergency number in non life threatening situations more often. This leads to an increase in the diversity of cases that ambulance nurses have to handle, for which their standard protocols are no longer sufficient. When ambulance nurses cannot use protocols for support, the reliance on clinical reasoning increases. With an increase in clinical reasoning and a higher variety in cases, this also increases the number of diagnostic mistakes, which can lead to worse health outcomes.

Ambulance nurses in the Netherlands are in a unique position within the healthcare system. In contrast to hospital nurses they work on their own without a medical doctor present. Ambulance nurses, in contract to medical doctors, are not trained extensively to diagnose patients. However, in cases where it is ambiguous what is wrong with the patient, ambulance nurses need to use clinical reasoning to develop a working diagnosis to decide whether the patient should be transferred to the hospital or not. The process of clinical reasoning is further impaired by the limited tools that ambulance nurses have access to and the time pressure they are usually under.

Clinical decision support systems may provide a way to help ambulance nurses in their clinical reasoning. These systems use a database of information about diagnosis to support medical personnel in their clinical reasoning. However, such systems are most efficient when they are built to support real life situations, instead of being based on our general theoretical understanding of clinical reasoning.

To be able to build efficient support systems we need to understand how clinical reasoning works in ambulance nurses. One common occurrence in clinical reasoning is cognitive errors. The main cognitive error that occurs is confirmation bias. Such biases tend to be most detrimental to novice clinicians, as experts learn from their experience to handle such biases more efficiently.

We want to find out whether confirmation bias occurs more often in novice ambulance nurses compared to experienced ambulance nurses. If so, we could build a decision support system that could focus specifically on preventing the negative effect of confirmation bias on clinical reasoning. This would then result in better medical outcomes.

In chapter 2 of this thesis the theoretical framework will be described. Chapter 3 will present the methodology. Chapter 4 shows the results and chapter 5 will contain the discussion.

2 Theoretical Framework

2.1 Task analysis

Ambulance nurses are a part of the pre-hospital emergency medical services (EMS) in the Netherlands. The way that EMS services are organised internationally differs by country, and can be divided in roughly two groups: the Anglo-American system and the Franco-German system (Al-Shaqsi, 2010).
In the Anglo-American system most paramedics are only allowed to perform basic life support interventions and are trained to quickly take the patient to the hospital in a safe manner as their primary task. Sometimes paramedics are accompanied by a physician in this system, who is then able to perform more advanced interventions on scene.
In the Franco-German system ambulance nurses are trained to stay on scene and stabilise the patient, with less emphasis on getting all patients to the hospital. These clinicians are educated to perform more advanced medical interventions than their counterparts in the Anglo-American system. Fewer patients are taken to the hospital in the Franco-German system, as the patient may instead be referred to their GP or even be left at home after being treated by the ambulance nurses.
The latter system is employed in the Netherlands, in which the ambulance is always staffed with at least one ambulance nurse and a driver.

2.1.1 Ambulance care in the Netherlands: From patient call to hospital

When a member of the public calls the emergency number 1-1-2 in the Netherlands they are connected to the emergency medical dispatch centre. Here a dispatch nurse triages the call and decides whether to send an ambulance and with which priority level. Subsequently the ambulance professionals are given information from dispatch about where they need to go and what the medical situation of the patient is.

On arrival at the scene the ambulance nurses are tasked with determining what medical issues are present with the patient, what treatments need to be given on scene and whether the patient needs to be handed over to other care facilities.

Andersson (2022) described that ambulance nurses primarily get their information from the patient themselves, by asking the patient about the situation and by checking their vital signs and other measurements. If the patient needs treatment that the ambulance nurses cannot provide or if they need further testing, they will be handed over to their local care facility (e.g. their GP) or transferred to the hospital. If the treatment from the ambulance nurses is sufficient and there is no further care needed then the patient is left at home.

If the patient needs to be handed over to another care provider, the ambulance nurses will give their assessment of the patient and their working diagnosis during the handover.

2.1.2 Patient interaction on scene

It is during the interaction with the patient on scene that the ambulance nurses need to make several medically important decisions. They have to decide what treatment they will administer, what diagnosis they find likely and what further care the patient needs.

To do so the ambulance nurses need to investigate the complaints and symptoms of the patient by listening to the patient's story and history, performing physical examinations and conducting tests and images. After all information is gathered the ambulance nurses will have a picture of the patient's conditions and can form a diagnosis.
Zwaan (2012) points out that in reality this process is far more complex. It can occur that the patient's complaints do not concur with the test results or that they have a variety of complaints and symptoms that don't point to a single clear diagnosis. It can be difficult to determine which information is medically relevant and which is not. In addition ambulance nurses also need to decide what medical guidelines are relevant and how they should be applied. Andersson (2022) observed that ambulance nurses often have to accept that they don't have a complete picture of the situation.

Regardless of the quality of the information that has been gathered, ambulance nurses need to decide what further actions need to be taken. If they make a wrong decision, they could leave a patient at home who in actuality needs hospital care, or they could transfer a patient to the hospital who did not need any further treatment or testing. It is fundamental to the quality of patient care that ambulance nurses are able to make a correct estimation of the patient condition and diagnosis. The process of gathering this information and coming to a diagnosis of the patient is called clinical reasoning.

In the rest of this chapter we will look at several relevant topics. First we will look at some cognitive aspects of pre-hospital care; clinical reasoning, decision making, confirmation bias and expertise. Following this, the influence of clinical decision support systems and dispatch information will be described. Lastly, the experimental set-up of previous research and the current study will be discussed.

2.2 Clinical reasoning

Trowbridge (2015) defines clinical reasoning as "the cognitive and non cognitive process by which a healthcare professional consciously and unconsciously interacts with the patient and the environment to collect and interpret patient data, weigh the benefits and risks of actions, and understand patient preferences to determine a working diagnostic and therapeutic management plan whose purpose is to improve a patient's well-being."

Clinical reasoning encompasses a lot of different sub-tasks and cognitive abilities. In this research the focus is specifically on what can be called diagnostic reasoning, which is the process of determining the diagnosis of a patient. Diagnostic reasoning has been explained under different paradigms, one of which is decision making. Under the decision making paradigm, diagnostic reasoning is explained as updating a hypothesis with imperfect information, namely the clinical evidence that is gathered during a patient encounter (Hunink, 2001).

2.2.1 Naturalistic Decision Making

Historically decision making was viewed through normative decision making models, which assumed that humans base their decisions on rational considerations. Models such as Bayes' theorem were used to predict what decisions would be optimal. However, while normative models may work in a lab setting, estimating the parameters of a completely rational model is not feasible in real-life complex situations (Folk et al., 2012).
Within cognitive psychology this normative view has evolved to an understanding of decision making where it is recognized that humans are limited in their cognitive resources and that they are not completely rational decision makers. Instead human beings make use of heuristics to come to decisions in a quick and efficient manner (Kahneman et al., 1982). This leads to the development of the model of Naturalistic Decision Making (NDM) (Klein, 2008), which intends to describe how people actually make decisions in real-life situations. These models take a different viewpoint, in which the focus is not on humans making suboptimal decisions, but rather on the question how humans are able to make decisions in suboptimal circumstances, such as time limits, uncertainty and other unstable conditions (Orasanu & Connolly, 1993). Given that ambulance nurses often have to work under straining circumstances, naturalistic decision making models are far more applicable than normative models.

Hammond's (1987) cognitive continuum theory (spectrum of intuitive and analytical processes), Rasmussen's (1983) cognitive control (otherwise known as the skill-based, rule-based, and knowledge-based model) and Klein's (1993) recognition-primed decision making model are all NDM models. All these theories came to the same conclusions, in Klein's (2008, p. 457) words: "People were not generating and comparing option sets. People were using prior experience to rapidly categorize situations. People were relying on some kind of synthesis of their experience –call it a schema or a prototype or a category – to make these judgments. (...) The static notion of decisions as gambles, which portrays people as passively awaiting the outcomes of their bets, did not fit leaders who were actively trying to shape events."

In comparison to normative models, NDM models include the prior perception and recognition of situations and the consequent generation of possible actions, instead of only focusing on which choice is made amongst presented options.

2.2.2 Recognition-primed decision making

Recognition-primed decision making (RPDM) is a model by Klein (1997; 1993) that's also part of the NDM approach. RPDM focuses on the notion that pattern recognition forms the basis of complex situation assessment. Specifically, it pays attention to the notion that experts have the ability to focus on a subset of best options and disregard bad choices in an instant, while novices need a lot of time to deliberate the situation. Chase and Simon (1973) observed this phenomenon in master chess champions.
Klein (1993) developed this model on the basis of in-depth interviews with firefighter commanders about their work. To illustrate the model, here follows an example of the way that an experienced firefighter fights a fire:
When a firefighter arrives on the scene of a fire, they immediately start collecting cues from the environment. These cues then activate long-term memory representation of earlier

experiences. The pattern matching that occurs between the memories and the current situation leads to the firefighter being able to recognize what type of fire it is. If the current situation is not a familiar one it may take more cue collection before pattern matching can occur. Once the fire type has been recognized, the possible actions that are known to be successful for fighting this type of fire are assessed. If there is not one obvious action that should be taken, the possible actions are estimated on their efficiency by running mental simulations with them. The first option that provides a satisfying outcome in these simulations is the one that is selected. See figure 2.1 for a visual presentation of this process.

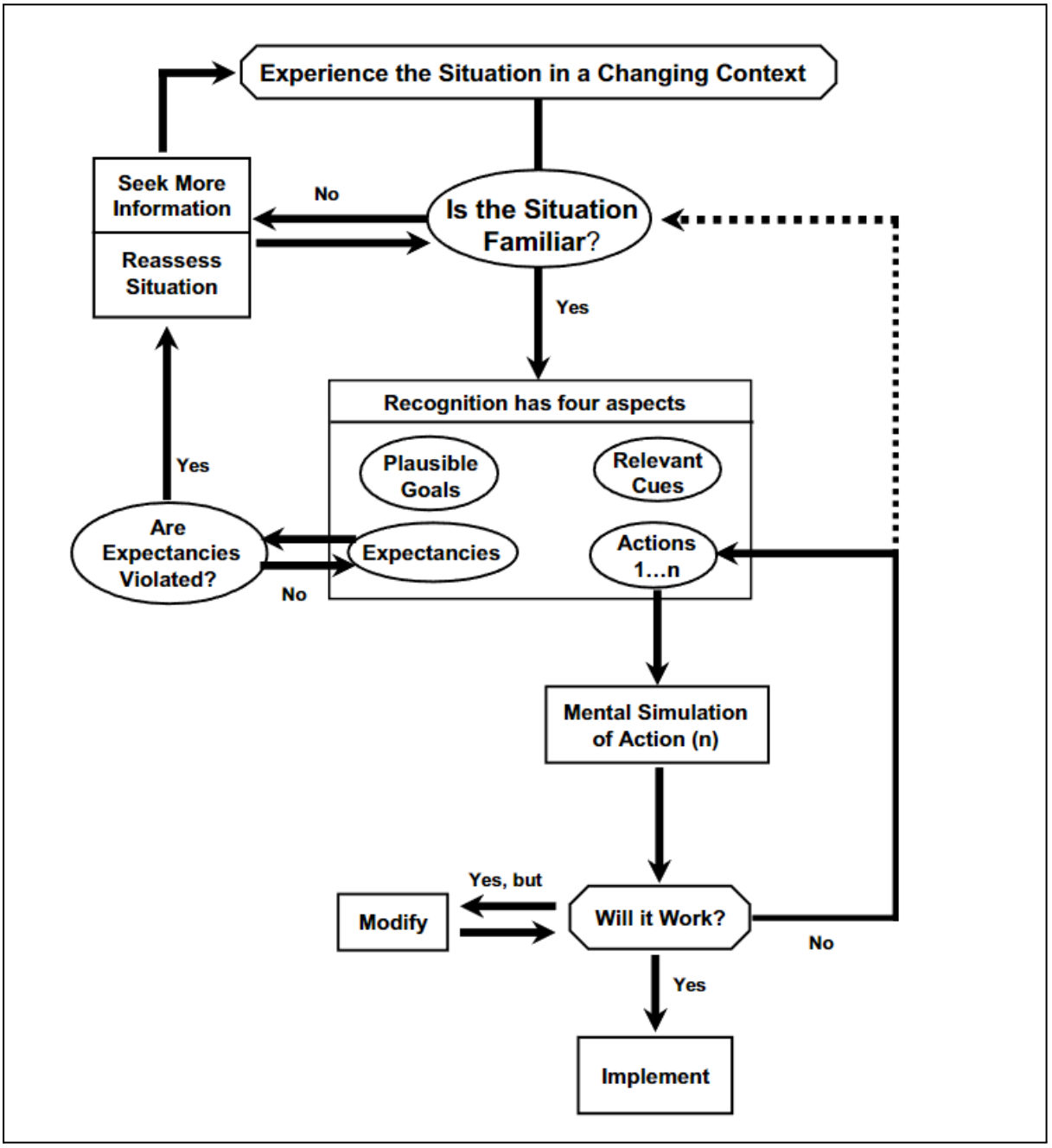

Figure 2.1: ***Model of recognition-primed decision making.*** Image from Klein et al. (1993)

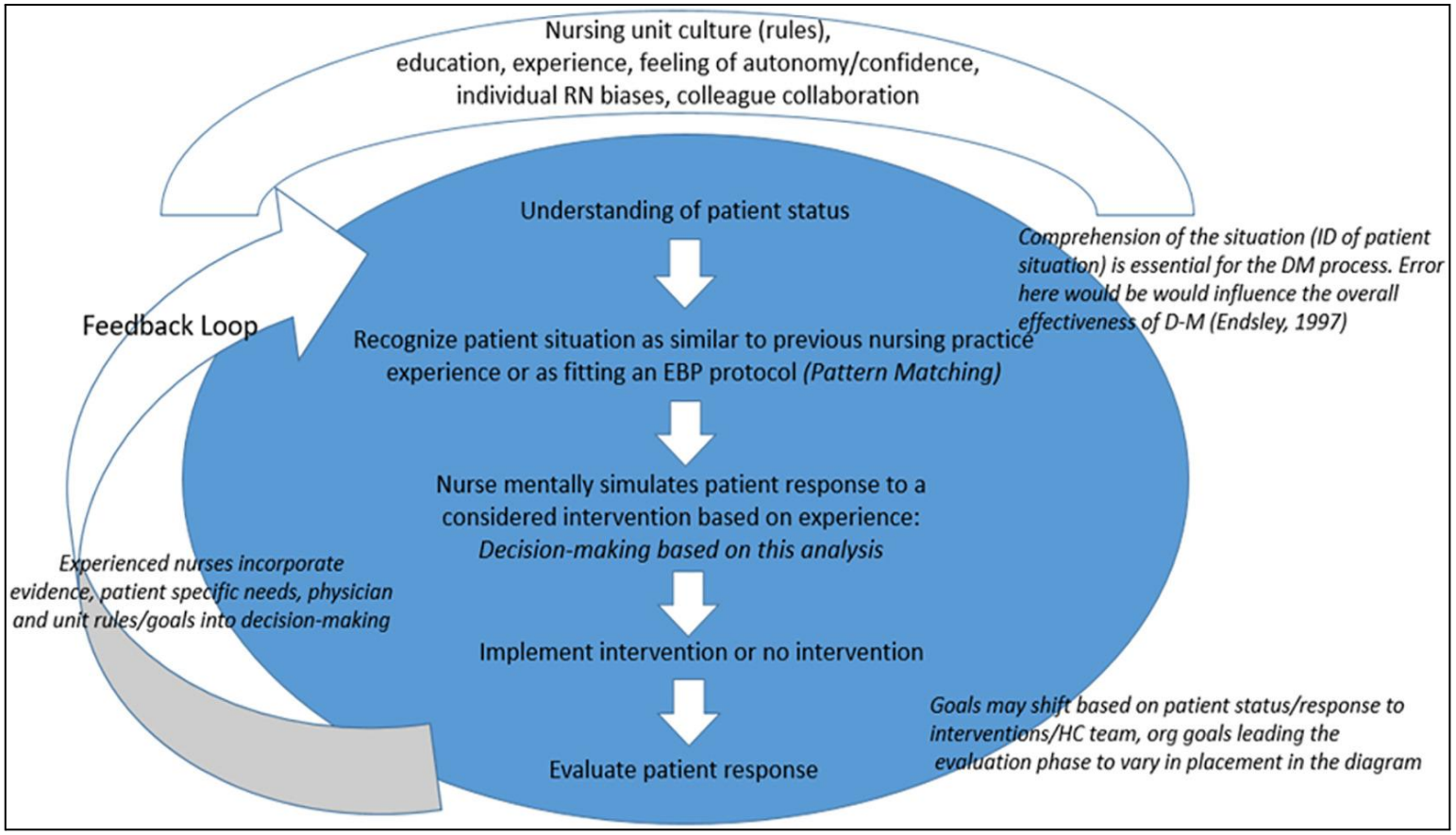

Figure 2.2: ***Diagram of the Practice-Primed Decision Model.*** Image from Nibbelink et al. (2019)

The way that experts become so efficient at decision making is that they have experienced many different situations within their domain, which allows their long-term memory representations to come to resemble the regularities in their domain. Kushniruk (1998) observes that the speed with which physicians generate hypotheses in their diagnostic reasoning shows that the semi-automatic recognition-primed decision making is at work.

Nibbelink (2019) derived a theory based on the RPDM model in the context of acute care nursing. They found that RPDM is a solid theoretical basis because it concerns decision making in complex situations under time constraints and other unstable conditions, which aligns with the tasks of acute care nurses. The acute care nursing model they derived from the RPDM model is called the Practice-Primed Decision Model (PPDM).
In this model an acute care nurse comes to understand the patient status by matching the status of the current patient with those of previous patients they have encountered. They then use this understanding of the patient status to run a mental simulation of how medical interventions would affect this patient. After an intervention has been administered, the nurse would then re-evaluate how the patient status has changed and decide whether further interventions are needed. This cycle then repeats until the patient status has improved. See figure 2.2 for a visual representation.
Nibbelink subsequently suggests that more research on the differences between decision making in experienced and novice nurses could help expand upon this model. The PPDM model has a large focus on interventions, similar to pre-hospital care. However, there is no focus on transferring patients to other care facilitators or establishing a specific diagnosis.

Even so, the PPDM model provides a basis for how RPDM can be understood in the context of emergency medical care.

2.3 Confirmation bias

2.3.1 Why confirmation bias?

It is inherent to the job of an ambulance nurse that they receive information before they arrive to see the patient. The clinical reasoning process already starts when the nurse receives the dispatch information, which they can use to start generating differential diagnosis or plan ahead actions to undertake with the patient. Pelaccia (2014) showed that physicians tended to create their differential diagnosis before they saw their patient. Because the dispatch information is given prior to any other information this may lead the nurses to be biased towards this information.

One of the most common biases is confirmation bias. Confirmation bias is the tendency to confirm the idea you have instead of trying to disconfirm it. Every person has the tendency for confirmation bias, and it is prevalent in the medical field.

2.3.2 Definition

Confirmation bias is often defined by two different mechanisms, the manner in which information is sought and the manner in which information is interpreted (Croskerry, 2002; Klayman, 1995).

2.3.2.1 Evidence search

In confirmation bias people tend to seek information that confirms their current hypothesis, instead of disconfirming it, even if this information would be more diagnostic (Einhorn & Hogarth, 1978).
This tendency is called positive testing. Wason (1960) introduced the “rule identification task” to show the phenomenon. The task goes as follows:

Imagine that the sequence of three numbers (e.g., 2-4-6) follows a rule. Your task is to diagnose that rule by writing down another sequence of 3 numbers. Your instructor will tell you whether or not your sequence follows the correct rule.

Propose that the participant in this experiment believes the rule to be that “numbers must go up by two”. They are then more inclined to test a sequence that they believe to be accurate (such as 1-3-5) than a sequence they believe to be false (such as 1-2-3). The underlying rule in the experiment was “any sequence of ascending numbers”, which would be impossible to discover if only sequences to confirm the participants assumption were tested.

This relates to the phenomenon of pseudo-diagnosticity, which was first identified by Micheal Doherty (1979). In the case of pseudo-diagnosticity participants would select information that relates to their current hypothesis and not to any alternative hypothesis, because they believe this information to be more diagnostic. People can make the incorrect assumption that information that is consistent with their current hypothesis will be inconsistent with any

alternative hypothesis. This one sided collection of evidence then results in the participant being unable to make a correct estimation of the probability between the different hypotheses.

2.3.2.2 Evidence interpretation

People tend to underweight or fail to remember dis-confirming information (Arkes & Harkness, 1980). Lord, Ross and Lepper (1979) showed that when people on two sides of an issue are given a pro and con argument they end up further apart then when they started. Darley and Gross (1983) showed that if subjects were given prior (false) information they would interpret the same information differently and become more convinced of their hypothesis.

Another effect that comes into play here is the feature-positive effect. This effect describes that people find it easier to focus on information that is present than information that is absent. Baila (1980) found that people often fail to use the absence of important cues as diagnostic information.

In the RPDM model specifically, the process of pattern matching can readily lead to confirmation bias, because the matching is done specifically on aspects that are present in both the current situation and situations that have been experienced before. This leads to an oversight when it comes to aspects of current and previous situations that are not present.

2.3.2.3 Effects of confirmation bias

The combination of searching for positive evidence and then interpreting any found evidence in a positive manner leads to incorrectly confirming the current hypothesis.

Croskerry (2002) believes that confirmation bias preserves diagnoses that are based on weak evidence and may lead to completely missing the correct diagnosis.
The occurrence of confirmation bias is also influenced by other factors. Experts tend to have less confirmation bias. Feltovich (1984) found that expert physicians were less likely to engage in pseudo-diagnosticity than novice physicians.
The task environment also has an influence. In complex situations with high stakes and time pressure the cognitive resources are more constrained and confirmation bias is more likely to occur (Woods et al., 1994). Depending on what real-life consequences to a chosen hypothesis are present, people may be more inclined to be conservative in their evidence search. In the case of ambulance workers there is more risk in leaving a patient that needs care at home (a false negative), than sending a healthy patient to the hospital (a false positive). Therefore, they may be more likely to test for a more detrimental diagnosis. This would then also introduce a bias, which can be seen in the practice of defensive medicine. These task environments are difficult to reproduce in general lab-based experiments regarding confirmation bias, which makes the result of those experiments difficult to generalise to real-world tasks.

2.3.3 Computational models

Within RPDM it is assumed that experts develop something akin to a schema based on their experience which supports them in recognizing familiar situations. In this section the specific

way that this phenomenon occurs for confirmation bias is discussed, to the extent that research exists about it.

Confirmation bias is a heuristic that occurs commonly and unconsciously. Bilalic (2008) showed that expert chess-players, even when they stated that they were looking at alternative solutions, kept looking at the chess pieces that were relevant for their original strategy. Bilalic suggests that when people recognize a situation as familiar the schema for that situation is activated and inhibits allocation of attention to information that is not relevant to the schema. They assume that there is a similar mechanism behind the confirmation bias.

Confirmation bias occurs because it takes less effort and resources to focus only on specific information. If someone wants to diagnose what is wrong with a patient, it is more efficient for specific symptoms to only activate a small portion of the available diagnostic knowledge. However, errors occur when different diagnoses can manifest the same symptoms. When these symptoms are confused, the wrong portion of diagnostic knowledge can be activated (Johnson et al., 1988). This can lead to inefficient diagnosis practices or misdiagnoses. Mehlhorn et al. (2011) also found that when a symptom is perceived as being uncertain, there is no activation of specific information in memory. This would explain why ambiguous evidence is less likely to be weighted against the current hypothesis.

Although these studies shine a light on parts of the mechanisms behind confirmation bias, there is not yet a cohesive computational model specific to the phenomenon of confirmation bias.

2.3.4 Previous research regarding confirmation bias

Cognitive biases occur in all people and medical professionals are not an exception to this (Pines & Strong, 2019). Confirmation bias has been found to occur regularly in physicians (Mendel et al., 2011). Feltovich, Spiro and Coulson (2001) showed that medical students tended to not let counter-evidence change their mind about their current hypothesis in regards to complex cardiological diagnosis.Leprohon and Patel (1995) found that sometimes aspects of patient information were mis-remembered by nurses who work at the dispatch centre when they did not fit the nurse’s hypothesis.

Confirmation bias can be detrimental to clinical reasoning and diagnostic accuracy (Mendel et al., 2011; Prakash et al., 2017). Berge and Mamede (2013) conclude from a review of several experimental studies that confirmation bias causes diagnostic errors. However, confirmation bias among ambulance nurses has not been studied specifically.

Because confirmation bias can negatively affect results, several studies have tried to reduce confirmation bias by using de-biasing techniques. However, these techniques have not been found effective in decreasing cognitive errors (Parmley, 2006; Sibbald et al., 2019; Zwaan & Singh, 2020).

2.4 Experts

2.4.1 Definition

To be able to say something about expert and novice ambulance nurses, we need to define what expertise is. Experts are seen as individuals who have a high skill level within their domain. However, the way in which you measure this skill level is under discussion. There is no uniform accepted definition of expertise.

Ericsson (2011) found that the amount of deliberate practice is the best indicator for superior performance, to which Ritter (2014) adds that experiential, social and cognitive factors also contribute to expertise.

The RPDM model (Klein, 2008) states that expertise is developed through receiving feedback on your decisions. In medical fields this feedback is often less available in comparison to other fields. For example, ambulance nurses in general don't receive any updates on a patient that they have transferred to in-hospital care, which means that they don't receive feedback on whether the transfer of care was the right decision and whether their working diagnosis was the correct one. Even so, ambulance nurses are not completely without feedback. When they are actively treating a patient they can see whether their actions are improving or deteriorating the state of the patient. This provides them some feedback on their working diagnosis.
However, because there is no feedback outside of the immediate treatment, it is more difficult to say to what degree ambulance nurses develop their expertise, in the sense that Klein defines it.

In several experiments in the medical field researchers have taken to selecting participants based on peer nomination or other recognized accomplishments. Elstein (1990) mentions that this practice has been taken over by looking at years of experience instead, because it is hard for clinicians to objectively identify the expertise of colleague's, since the results of their work are not publicly available (in comparison to other fields).

Some researchers choose to divide experts and novices by a specific amount of years of experience, for example Ward et al (2010), or a specific age, see Eva and Cunnington (2006). How long it takes to become an expert is a topic of discussion (Larkin et al., 1980). Therefore, choosing an arbitrary age or years of experience at which one is deemed an expert is not an objective measure, which can skew result interpretations.

When only a small group of participants is used for a study, these participants may be handpicked in regards to several measures of expertise. For example, in Pelaccia's (2015) study participants are selected based on 6 different requirements to consider them experts, ranging from education to peer nomination. In Hoffman's (2009) study this was done on the basis of 7 requirements.

However, when it comes to studies with larger populations of participants, it becomes harder to establish a group of experts and novices in such a manner. Dividing novices and experts in this way also ignores the notion that expertise is a spectrum in which expertise develops over time and through effort, instead of a binary divide.

There are no established criteria to define an expert ambulance nurse. Therefore this study will use the measure of years of experience in a non-categorical manner instead, while acknowledging that experience is not a complete indicator of expertise.

2.4.2 Previous research regarding expertise

Experts are generally better at performing tasks than novices. One of the reasons performance is increased in medical experts may be because of the reduction in biases (Richie & Josephson, 2018).

Thiele et al (1991) argues that novice nurses have limited ability to recognize relevant information cues, because of their lack of clinical experience, which leads to more fault in their decision making. While Hobus (1987) found that expert physicians were better at using patient context to come to a correct diagnosis. And Smith et al (2013) found that experienced paramedics provide better care, through greater cue gathering and inferential thinking.

However, Kostopoulou (2015) and Krupat (2017) found no correlation with experience when it comes to diagnostic accuracy in physicians.

Krems (1994) shows that within both neurologists and internists there is no difference in how many correct final diagnoses there are between experts and novices, but does show that experts have less confirmation bias than novices.

Previous research shows that expert psychiatrists have reduced confirmation bias (Mendel et al., 2011) and that medical students with more experience are better at adjusting their hypothesis when encountering disconfirming information (Arocha & Patel, 1995). However, a study among psychologists found there to be no difference in confirmation bias (Parmley, 2006) between novices and experts.

As can be seen from this short summary, the influence of experience differs between the different medical domains and does not consistently result in a decrease in confirmation bias. Within ambulance nurses the influence of experience on the presence of confirmation bias has not been researched specifically. Because the result of previous research differs between medical domains, we cannot simply assume that physicians or paramedics (or any other medical professional) will show the same correlation between experience and confirmation bias as ambulance nurses.

2.5 Clinical decision support systems

Debiasing strategies are not very effective when it comes to confirmation bias. Another approach to reduce confirmation bias comes in the form of automated systems that allow for access to information and cognitive support during the time when decisions are made.

A clinical decision support system (CDSS) is a computerised system that integrates clinical and environmental information to support the decision making process of clinicians. CDSSs do this by accumulation, validation and transformation of data into actionable information. There is a broad range of different types of CDSSs, ranging from alerts when a patient's

vitals change to systems that can suggest diagnoses and treatments. The technological implementation can also vary greatly. CDSSs can be built on expert systems, simple metrics or AI algorithms.

CDSSs are most successfully used in the medical field in interpreting diagnostic tests. Such tasks may include comprehending medical images (Doi, 2007) or automated biopsy (Tsukada et al., 2000). These applications are however not as useful during prehospital emergency care.

Hajioff (1998) found that clinicians are most interested in a CDSS that can support them in the selection of diagnoses and treatments. However, such systems are often rated very poorly when it comes to usability in the field (Wickens & Dixon, 2007). The reason for this is that selecting a diagnosis is a more complex task than analysing a diagnostic test. Medical inference demands the integration of a lot of complex and nuanced information, whereas determining the outcome of a diagnostic test requires an integration of information that is more reliable and accurate. Furthermore, the reasons that a system selects a certain diagnosis are hard to understand for the user. Not understanding how the system works and the system performing with a relatively lower accuracy to other automations causes there to be a low trust in diagnosing systems and subsequently a reluctance in using them.

Bashiri (2019) reviewed 14 different CDSSs that are used in emergency prehospital care in the American paramedic system. They concluded that the difference in geographic location, knowledge level and skill of the specialist leads to differences in the requirements for decision support systems. They also found that it is key that CDSSs are part of the natural workflow and are able to offer practical advice at the time and place of decision making. Ishak et al. (2010) concluded that decision support systems should be suitable for the user so that it can be easily understood how the system works and all decisions can be replicated. Furthermore, Yates et al (2003) found that decision support systems developed under normative models of decision making did not improve decision quality and did not get adopted in the field. Instead decision models need to take into account how people actually make decisions to be able to integrate into the existing workflow. To facilitate this Kushniruk (1998) suggests that it is important to understand the medical cognition to accurately capture the way in which clinicians work.

Muhiyaddin (2020) found that one of the drawbacks of using CDSSs is the interruption in the patient-clinician communication and the increase in time that clinicians have to spend on documentation in the CDSS system (such as inputting observations into the system). In a prehospital setting, in which patient encounters are often time-critical, these drawbacks are substantial.

Even though it is difficult to overcome these obstacles, CDSSs have proven to be able to increase diagnostic accuracy and care quality. Both Muhiyaddin (2020) and Bashiri (2019) found that in most cases CDSS had a positive impact on patient care. In Bashiri's study the CDSSs that were reviewed were often designed to improve a specific aspect of prehospital care. For example, some of the CDSSs mentioned were focused on improving care in older patients, cardiac patients or severely burned patients.

In conclusion, CDSS are most useful when they are specific and reliable. They must be developed to fit into the natural workflow of the clinicians to reduce any interruptions in the

communication with the patient and to be found useful by the clinicians. Furthermore, they are often developed to improve a specific aspect of a clinician's work.

In the case of prehospital emergency care in The Netherlands this means any CDSS that is developed must take into consideration the clinical reasoning of the ambulance nurses, the time critical nature of the work and the importance of not interrupting the patient interactions. It is also important to understand which aspects of ambulance care are in need of support.

As mentioned earlier, confirmation bias is possibly increased in ambulance nurses by nature of their task. The fact that ambulance nurses receive dispatch information before they arrive on scene is unique to prehospital emergency care. This provides an opportunity to develop an CDSS which could integrate with the dispatch information and may provide a de-biasing effect. However, to be able to build a CDSS that can be used efficiently it must first be understood how ambulance nurses handle dispatch information and confirmation bias in their current work.

2.6 Dispatch information

Andersson (2019) found that clinical reasoning starts before ambulance nurses arrive on scene, and that the start of this clinical reasoning is based on the information they received from dispatch. Gunnarsson and Stomberg (2009) also found that ambulance nurses regard the information from other operators to be one of the factors that influence their decision making in emergency care situations. Although the nurses experienced dispatch information as an influence on their decision making, the research does not tell us how dispatch information affects clinical reasoning. There is no further research into the influence of dispatch information on ambulance nurses. However, there is research into the influence of clinical context on clinical reasoning and diagnostic accuracy.

Pelaccia et al. (2015) looked at the influence of prior information on the clinical reasoning of emergency physicians. They found that the clinical context had a major influence on what cues physicians collected in the first moments that they met the patient. Because the study was based on 15 physicians, who were all interviewed regarding one event in their clinic, it is hard to generalise the results. Pelaccia et al. suggested that ideally a group of physicians should be observed solving the same case to determine how consistent the influence of prior information is.

Sibbald et al. (2011) looked at the influence of patient history, a form of clinical context, on diagnostic accuracy in residents performing a physical patient examination. They found that the patient's history had a positive impact on diagnostic accuracy only when the resident already guessed the correct diagnosis before starting the physical examination. In dispatch information the symptoms are often accompanied with a probable diagnosis, in the literature this is referred to as diagnostic suggestion. Kostopoulou (2015) found the same improvement in diagnostic accuracy as Sibbald when physicians were presented with a list of possible diagnoses at the beginning of a medical vignette. But what Sibbald neglected to take into account is that clinical context is not necessarily reliable.

Bond (2018) and Durning (2012) both show that incorrect diagnostic suggestions have a negative impact on diagnostic accuracy.

Dispatch information has a risk of providing incorrect information about the patient scene. One study mentions that in 30% of cases the dispatch information did not concur with the eventual assessment of ambulance nurses on scene (Lindström et al., 2011). The reason that dispatch is at a higher risk of providing incorrect information is that they often only have access to second- or third-hand information (Karlsten & Elowsson, 2004).

Based on these studies we can assume that dispatch information could lead to both positive and negative impact on diagnostic accuracy, depending on the accuracy of the provided information.

Sibbald (2011) also suggests that clinical context can lead to confirmation bias. If dispatch information leads to more confirmation bias and the dispatch information is incorrect, this can lead to a wrong diagnosis and a diminished quality of patient care. The goal of providing dispatch information to ambulance professionals is to have a positive influence on clinical outcomes.There has not been specific research to determine whether incorrect dispatch information negatively affects the diagnostic accuracy of ambulance nurses. It is important to determine whether this is the case, and if so, in what way CDSS could be implemented to reduce this effect.

2.7 Previous studies

Clinical reasoning is inherently not directly observable and it can be hard to articulate for participants because it is in part a subconscious process. Because of these reasons there is not one straightforward way in which clinical reasoning is studied. In this section two main approaches in which clinical reasoning and decision making is studied will be discussed, namely: field experiments and simulated patient encounters.

2.7.1 Field experiments

One approach to studying clinical reasoning is to observe clinicians as they do their work in the field. This often includes observing interactions with real patients in a participant's regular work environment. Examples of this research approach are the studies by Hoffman (2009) and Pelaccia (2015).

Hoffman (2009) observed 8 ICU nurses to study what cues they collect during patient interactions. In their experiment the participants were asked to think aloud during the encounter and were also interviewed about the patient encounter afterwards.
In Pelaccia’s (2015) study 15 emergency physicians were observed before and during the first few seconds of a patient encounter. Participants were filmed during the patient interaction and interviewed about the videos afterwards to examine their clinical reasoning.

The purpose of using a field experiment to study clinical reasoning is that the environment in which clinical reasoning is performed is as realistic as possible and to be able to collect richer data. As is the nature of field studies, this also means that it is difficult to control other environmental variables in real life patient encounters. This, in combination with the fact that all patient encounters were unique, makes it hard to compare the different encounters to one another.
Another drawback of field experiments is that it is labour-intensive to gather the data. Both of

these studies had a small sized group of participants and in the case of Hoffman all the participants worked at the same ICU unit. This makes it harder to generalise the results of such studies to a wider population.
Pelaccia (2015) mentions in their paper that one of the drawbacks of interviewing the participants afterwards is that there is no guarantee that the reasoning that participants report in the interview is the same as the reasoning that was used during the patient encounter. In the interview afterwards participants already have all the information and may be retroactively filling in the blank when talking about their actions in the video recording of the patient encounter.

Another difficulty in field studies is that, especially in emergency services, operational demands always take priority. This means that the observer or any measuring equipment cannot interfere with the quality of care that is delivered to the patient. This limits the degree in which experimental studies can be performed in the field. Often simulations of patient encounters are used to some degree for this reason.

2.7.2 Simulated clinical encounters

In recent years two studies regarding confirmation bias among medical professionals have used simulations. Mendel (2011) studied confirmation bias in psychiatrists and medical students. In Mendel's study one medical vignette of a patient was presented to participants and participants could choose one of two presented diagnoses. The vignette described patient symptoms in a way to bias participants towards one diagnosis. Participants were given 12 pieces of information, which were explicitly titled in support of one of the diagnoses, that they were allowed to select. This selection of information was then measured to determine the presence of confirmatory search patterns.
Parmley (2006) studied confirmation bias among psychologists. In their experiment participants received two medical vignettes, followed by a second part containing additional information about each vignette one week later. The second part of the vignette contained information that was either consistent or inconsistent with the diagnosis indicated by the initial information. After each part the participants were asked to give a diagnosis based on the information.

The benefits of these simulations is that a far larger group of participants can be studied (150 and 102 participants for Mendel and Parmely respectively). Furthermore, the participants all receive the same case, with any variations being controlled. For example, Parmely had a control condition in which the information stayed consistent throughout both parts of the vignette. This allows the influence of change in information to be determined.

However, the set-up that Mendel and Parmely employ is more abstract in comparison to real-life clinical cases. In Mendel's case, participants only have the choice between two diagnoses and the evidence they are provided is explicitly labelled in favour of a diagnosis. In real clinical cases the information is more ambiguous. This set-up does allow testing for confirmatory search patterns, but makes it harder to determine how participants weigh this information. Telling a participant beforehand whether the information will support a diagnosis may suppress the effect that confirmation bias can cause, where people disregard information that does not align with their hypothesis.

In the case of Parmley's experiment, the participants were not allowed to perform information search at all. But in contrast to Mendel's study, participants were not limited in what diagnosis they could select.
In both studies the participants are only asked for their working diagnosis twice per vignette. This is in contrast to field studies in which clinicians continuously update their diagnosis while interacting with the patient.

The four studies discussed here are mostly focused on either determining how data is gathered by clinicians or how their diagnoses change, but don't take both of these mechanisms into account. Knowing what actions people take in situations prone to confirmation bias and how the information gathered from those actions is weighted is important to understand how to support professionals in their decision making.
In the current study we want to measure both mechanisms of confirmation bias and do so in a realistic set-up that allows a large group of ambulance nurses to participate. To determine the influence of dispatch information it is necessary to perform an experimental manipulation, as it would be difficult to quantify the quality of dispatch information for every unique patient encounter. Therefore a simulated patient encounter will be used instead of a field experiment, in which participants are presented with the same case.

2.8 Current study

It is important to understand how clinical reasoning works in ambulance nurses to be able to build tools that support ambulance nurses in their task. One of the task aspects that may influence clinical reasoning is the dispatch information nurses receive. This information may lead to both benefits and drawbacks when it comes to task performance. Those drawbacks may be caused by the increase of confirmation bias.

If experts are better at incorporating clinical context and have reduced bias, then we can assume that the drawbacks of receiving dispatch information should be decreased for ambulance nurses with experience. Is the decision making and information search of experts different from novices to allow dispatch information to be incorporated more effectively? And do experts experience less confirmation bias that allows them to incorporate dispatch information more efficiently? Knowing the answer to this question will help in developing tools to support (novice) ambulance nurses in integrating dispatch information to their benefit and reducing diagnostic errors.

Research question: Do years of experience influence the amount of confirmation bias that ambulance nurses experience in clinical reasoning?

Additionally, it is important to confirm whether dispatch information increases diagnostic accuracy when it is accurate and decreases diagnostic accuracy when it is inaccurate.

The initial hypothesis was that the confirmation bias would be decreased in experts. More specifically, experts would show less positive testing and under-weighing of evidence than novices.
This would allow experts to disregard an incorrectly suggested diagnosis quicker, decreasing the negative impact on diagnostic accuracy.

It was further hypothesised that, when an accurate diagnosis was suggested, experts would be able to effectively use this information and increase their diagnostic accuracy to the same or a higher degree than novices. This would confirm that experts use the dispatch information in their reasoning, instead of disregarding it all together.

2.8.1 Experiment

The goal was to determine whether working experience had an influence on the amount of confirmation bias. For this purpose a simulated patient encounter was used, which allows a large participant pool to engage in the same clinical case.

To test whether working experience had an effect on confirmation bias in the clinical reasoning of ambulance nurses, we designed an experiment with a 3 x 2 mixed design with 3 levels of dispatch information and 2 simulated clinical cases.

Ambulance nurses receiving dispatch information at the start of a clinical case provides a natural “garden path” methodology. The phrase “to be led down the garden path” means to be deceived. In the methodological sense, a garden path scenario presents the participant with a strong cue for an erroneous diagnosis at the start of the case, and later on presents the correct information to be able to arrive at an accurate diagnosis. Starting the participant off with incorrect information allows for confirmation bias to be observed, specifically the mechanism of underweighting counter-evidence, as participants will encounter information that is disconfirming their earlier diagnosis.

In practice, the garden path scenario was created by manipulating the dispatch information that was given to participants at the start of the simulated clinical case. The dispatch information would contain information that was not congruent by the information given by the patient during the clinical case. This was one of three conditions and was called the non-congruent condition. Additionally to the non-congruent condition, there was also a congruent condition, in which the dispatch information contains information which was congruent with the patient information during the case. Lastly, there was also a control condition in which the dispatch information contained limited information and was neutral in any diagnostic suggestion. The actual symptoms that the patient presented with were always the same within a case.

There were two medical cases about different medical subjects to make sure that the effects that are found are not limited to the specific subject. Although the subjects are different, participants had the choice of the same patient interactions and diagnoses selection in both cases. The results of patient interactions differed between cases, in alignment with the issues of the hypothetical patient in the case. The cases were inspired by real ambulance call-outs and modified to add missing information.

Half of the participants were assigned to start in clinical case 1 and the other half started in clinical case 2. A participant could participate in both cases, but they could not repeat a case. Every time a participant started a new case they were assigned to one of the three conditions. A participant was never assigned twice to the same condition. In total a participant could participate in two cases and be assigned to two conditions. Given that all of the above conditions were satisfied, participants were always assigned to the case and

condition that had the lowest number of participants assigned to it at the moment, so the cases and conditions are divided equally over all participants.

3 Methods

3.1 Participant recruitment

Ethical approval was obtained from the Research Ethics Committee of the University of Groningen (CETO-87209483). The study was advertised by eight Dutch ambulance organisations, who send emails to their ambulance nurses with a link to the website of the experiment. Participants were self-selected and there was no compensation for taking part in the experiment. All participants gave written informed consent before starting the experiment. The experiment ran for two months, from 3 July to 24 September 2022.

3.2 Task

3.2.1 Cases

We used two old cases from pre-hospital emergency medicine, a neurology case in which the patient suffers a stroke (CVA) and a cardiology case in which the patient suffers from acute myocardial infarction. These two cases were based on real ambulance cases and modified only to add missing information.

These two cases differed in how commonly they occur in ambulance nurses' work. Cardiology cases are one of the most common, whereas neurology cases are more rare (Ambulancezorg Nederland, 2021, p. 57). In these particular cases there was also a difference in how typical the symptoms are, the cardiology case had a more typical presentation than the neurology case. An ECG gives a clear indication that a myocardial infarction was happening in the cardiology case, whereas in the neurology case there was no such straightforward indication that a CVA has taken place.

Both these cases were set up in a way that it was likely to consider different diagnoses, especially at the start of the case. All patients were male to exclude pregnancy related cases and working diagnoses.

Case 1
Case 1 is a neurology case. The patient is a 39 year old male with a background of epilepsy. The patient visits his neighbour while exhibiting slurred speech, loss of feeling on the right side and an ataxic gait. The neighbour calls the emergency number. On arrival of the ambulance the patient only shows a numbing feeling in the right arm and a headache and no issues with walking or speaking otherwise. The patient also does not display any symptoms in accordance with the FAST test. The patient has not experienced any head trauma. Painkillers are not effective in reducing the headache.
Final diagnosis: The patient has suffered a cerebral vascular accident (stroke) of which the symptoms have reduced by the time that the ambulance arrived.

Case 2
Case 2 is a cardiology case. The patient is a 57 year old male suffering from an elevated heart rate and heavy breathing. On arrival of the ambulance the patient is pale and sweaty and experiencing chest pain, but fully conscious. Tachycardia is confirmed. Painkillers and

Nitroglycerin are effective in reducing chest pain.
Patient has a history of heart issues, for which an MRI-scan was taken at the hospital previously. The results were not available yet. The patient has experienced these symptoms eight times in the last two weeks, but has not called the emergency number before. This time the symptoms did not go away and got worse, upon which the emergency number was called.
Final diagnosis: Patient is suffering from an acute myocardial infarction.

3.2.2 Dispatch information

To present different dispatch messages in the three conditions, we created a unique dispatch message for each combination of case and condition. In total 6 different dispatch messages were created, see table 3.1 and 3.2. These messages all contained basic information about the case. The congruent and non-congruent condition also contained additional information to suggest a diagnosis to the participants.

The goal was to make the non-congruent dispatch message prime a different diagnosis than the congruent diagnosis. To achieve this it was necessary to add information to the dispatch messages that will be disconfirmed during the case, as only providing information that is proven true during the case would not deviate the non-congruent dispatch message enough from the congruent condition. Only adding information that can be disconfirmed to the non-congruent dispatch message may have created an effect where the non-congruent dispatch messages are experienced by the participants as less reliable than the congruent dispatch messages, purely on the fact that a part of the information in the message has been disconfirmed. Therefore both the congruent and non-congruent dispatch messages consisted of both information that can be confirmed and disconfirmed. This replicated the real world, in which dispatch information is often unreliable because the caller is not a medical professional and cannot necessarily access the situation accurately.

The base information was provided in all conditions, this functioned as the control and gave a basic reason for the emergency response. The dispatch messages in the control condition contained a minimal amount of information to prevent priming any diagnoses. The dispatch message was not left empty, as this is unusual during real ambulance call-outs and would provide inaccurate control for the other conditions. By supplying a minimal amount of information this may provide a control without inducing bias towards any particular diagnosis.

Table 3.1: ***Dispatch messages of case 1 (neurology case)***

Control condition	”20:09, een dinsdag in oktober. 39 jaar oud, Man, Bij kennis, Ademt. Verminderd bewustzijn (niet helemaal wakker, reageert niet normaal). Hij ademt normaal. Patiënt is bij buurvrouw thuis. Verklaring melder: Buurman is opeens niet goed aanspreekbaar.”
Congruent condition (CVA/TIA)	“20:09, een dinsdag in oktober. 39 jaar oud, Man, Bij kennis, Ademt. Verminderd bewustzijn (niet helemaal wakker, reageert niet normaal). Hij ademt normaal. Patiënt is bij buurvrouw thuis. Verklaring melder: Buurman voelt links niets meer, dronkemans gang. Hij heeft plotseling problemen met spreken.

	De resultaten van het Hulpmiddel Beroerte Diagnose geven overduidelijk bewijs voor een beroerte. Deze klachten zijn begonnen binnen de toegestane behandelingsperiode: sinds vanmiddag. Geen voorgeschiedenis met beroerte, maar heeft eerder een TIA gehad.”
Non-congruent condition (Epilepsy)	“20:09, een dinsdag in oktober. 39 jaar oud, Man, Bij kennis, Ademt. Verminderd bewustzijn (niet helemaal wakker, reageert niet normaal). Hij ademt normaal. Patiënt is bij buurvrouw thuis. Verklaring melder: Buurman is plotseling niet goed aanspreekbaar. Is bekend met epilepsie. Melder denkt dat het een epileptische aanval is”

Table Legend:
Base information
Additional information that is disconfirmed in case
Additional information that is confirmed in case

Table 2: ***Dispatch messages of case 2 (cardiology case)***

Control condition	”17:40, een vrijdag in eind augustus. 57 jaar oud, Man, Patiënt is thuis. Bij kennis, hevig benauwd, snelle hartslag.”
Congruent condition (Myocardial infarction)	“17:40, een vrijdag in eind augustus. 57 jaar oud, Man, Patiënt is thuis. Bij kennis, hevig benauwd, snelle hartslag. Pijn op de borst, pijn in de linkerarm, bekend bij cardioloog.”
Non-congruent condition (Pulmonary issues)	“17:40, een vrijdag in eind augustus. 57 jaar oud, Man, Patiënt is thuis. Bij kennis, hevig benauwd, moeite met spreken tussen ademhalingen, snelle hartslag. Gebruikt medicatie voor longproblemen.”

Table Legend:
Base information
Additional information that is disconfirmed in case
Additional information that is confirmed in case

The first screen participants were shown when they started a new case was a text-field with the dispatch information for the case. The dispatch information contained information about the day and time and information about the patient. In the right bottom corner a button with the text ‘Done’ was presented, if it was pressed the participant would be shown the next screen.

3.2.3 Differential diagnoses

To be able to measure what participants’ their differential diagnoses were throughout the cases it was necessary to present participants with a way to record their differential diagnosis.

In other experiments (such as Mendel et al. (2011)) the participants only have the choice between two diagnoses, which is restrictive and unrealistic in comparison to real medical cases. In the design of this experiment the participants had a choice out of 59 possible diagnoses. To replicate the real world circumstances it would be necessary to allow participants to freely fill in text instead of selecting a predetermined list of options. However, allowing free text input would make it hard to categorise diagnoses afterwards and compare them between (and within) participants. The choice to give the participants 59 diagnoses to choose from attempts to strike a balance between realism and reliability of the results.

The next screen showed the dispatch information at the top of the page, and below this another section in which the differential diagnosis (DDx) could be selected, see figure 3.1.

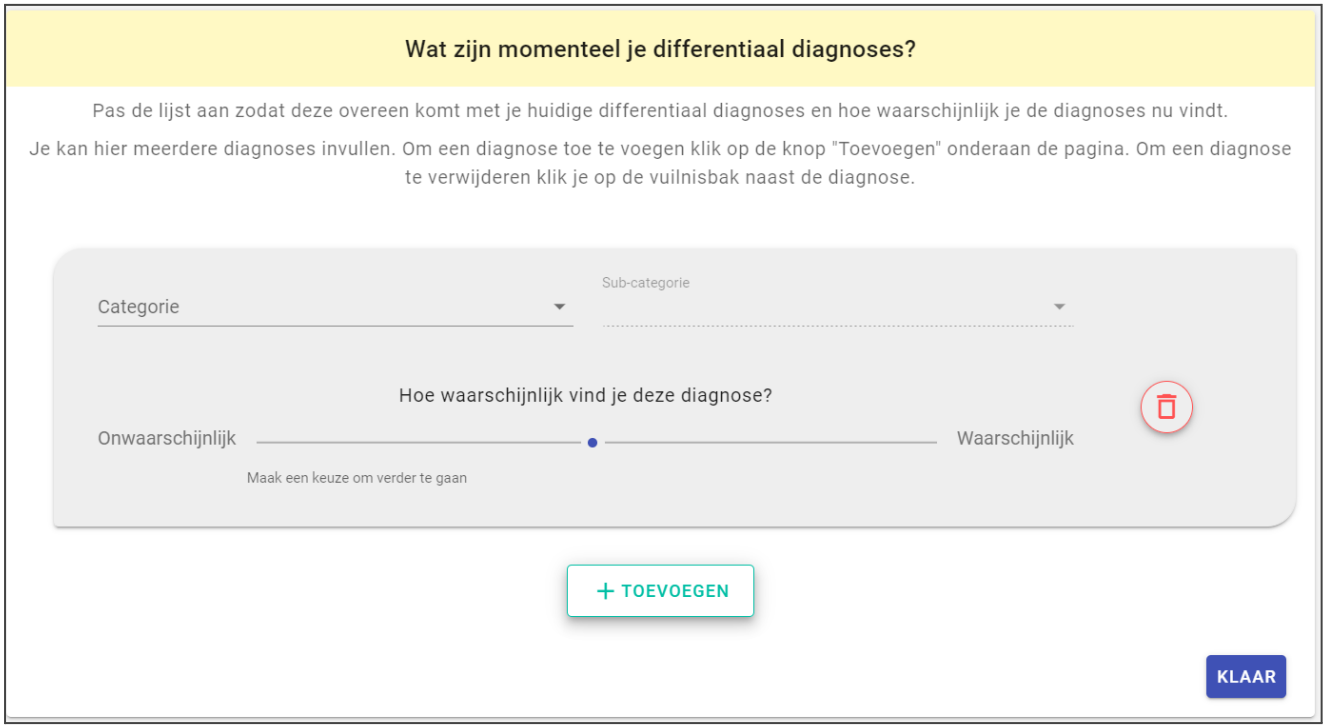

Figure 3.1: ***Differential diagnosis (DDx) section***

At the top of the DDx section the participant was presented with the question ”What are your differential diagnoses at this moment?”. Beneath this question a short text was presented with more information:
“Update the list so it is the same as your current differential diagnosis and how likely you find these diagnoses. You can add several diagnoses to the list. To add a diagnosis, click the “Add” button at the bottom of the page. To delete a diagnosis click the trash can button.”

Below this text a diagnosis field was presented, which contained two dropdown menus, a slider bar and a button with a trash can symbol. The second dropdown menu and the slider were disabled (greyed out and not clickable). The first dropdown menu was labelled “Category” and showed a list of 7 organ systems, see table 3.3. When an option in this menu was selected, the second dropdown menu was enabled. The second dropdown menu was labelled “Sub-category” and contained a list of specific diagnosis options, relevant to the

organ system that was selected in the first dropdown menu. For the diagnosis options of all organ systems, see table A.1 in the appendix.

Table 3.3: ***List of organ systems***

Pulmonary
Cardiovascular
Psychiatry
Gastroenterology
Urology
Neurology
Miscellaneous

Note. These options are presented in alphabetical order in Dutch.

When an option from the second dropdown menu was selected, the slider bar became enabled. Above the slider bar the question ”How likely do you find this diagnosis?” was presented. The left side of the slider was labelled “unlikely” and the right side was labelled “likely”. Below the slider a short text was presented: “Make a choice to continue”, which would disappear when the slider was moved.
The trashcan button allowed the participant to remove the diagnosis field, which was also stated in the tooltip of the button. Below the diagnosis field a button labelled ”Add” was shown. When pressed, another empty diagnosis field was added to the screen, which was also stated in the tooltip of this button. The participant could add a maximum of 10 diagnosis fields to the screen in this manner. If 10 diagnosis fields were on the screen the “Add” button would become disabled.

Several considerations were made to the visual presentation of the diagnoses selection to make sure that no unintended anchoring bias was introduced. First of all, the dropdown menus showed the list of options all at once, removing the need to scroll through the list and reducing unintended emphasis on the diagnosis at the top of the list (see Figure 3.2). Secondly, the first time a participant was asked to fill out their DDx, one diagnosis field was shown rather than several diagnosis fields. If 10 empty diagnosis fields were shown, this would suggest to the participants that they have to fill out 10 diagnoses exactly.
After the participants had filled out the DDx for the first time, the list was automatically saved and displayed again when the participant chose to update it. This way participants did not have to fill out the whole list again, which would have taken a lot of time and may even lead to them abandoning the experiment.

In the right bottom of the screen the participant could press the button labelled “Done”, which would show them the next screen.

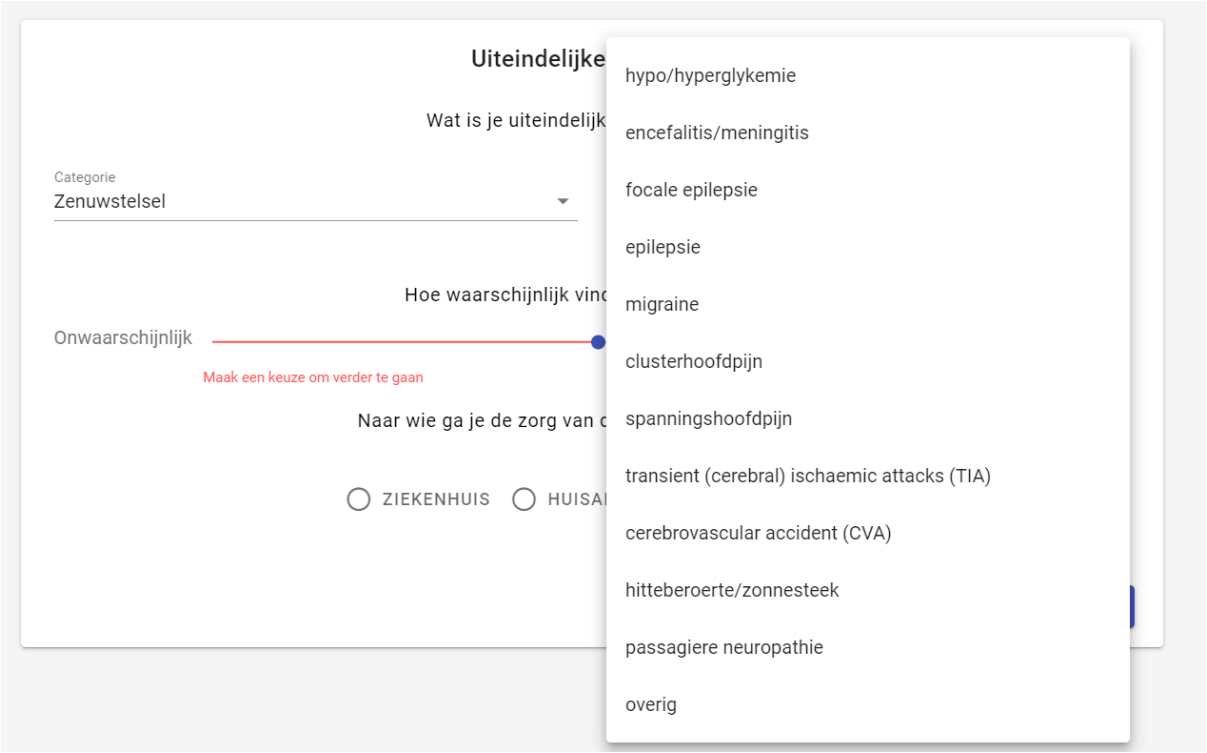

Figure 3.2: ***Open second dropdown menu to select specific diagnosis.*** In this example the first dropdown menu has “Neurology” selected. All the options that are shown in the open second dropdown menu belong to the “Neurology” category.

3.2.4 Actions

The next screen showed the main screen of the task, see figure 3.3.

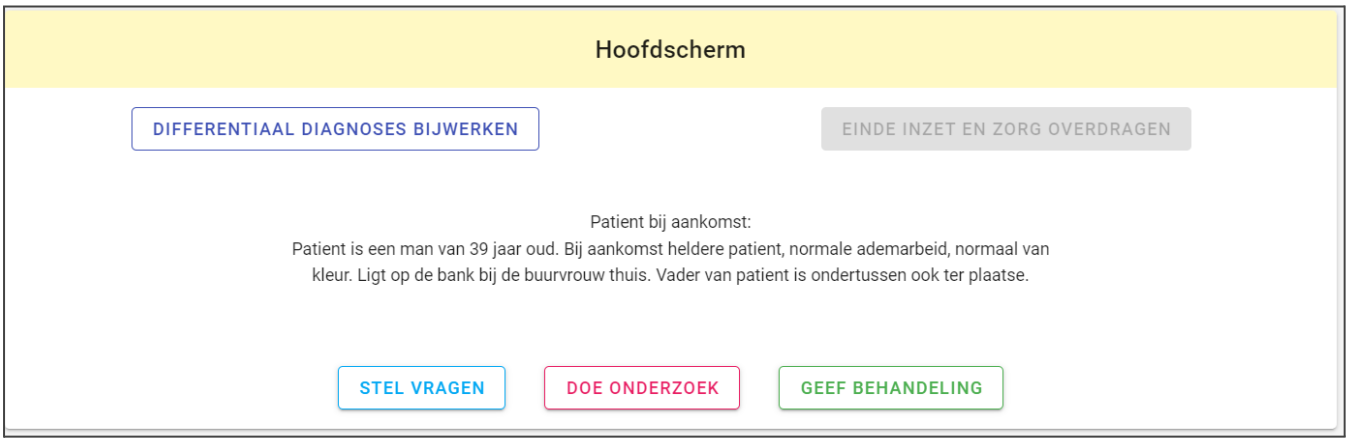

Figure 3.3: ***Main screen***

On the top of the screen two buttons were shown. One was labelled “Edit differential diagnosis” and the other was labelled “End case and transfer care”, the latter button was disabled. In the middle of the screen a text was shown which gives a visual description of the patient on arrival at the scene. At the bottom of the screen three buttons were shown, which present three categories of actions, labelled “Ask questions”, “Perform examination” and “Administer treatment”.

The “Edit differential diagnosis” button led to a screen where the DDx section was presented, without the dispatch information. All diagnosis fields that had been filled out by the participant previously were displayed. The participants could edit, add or remove diagnosis fields on this screen. They could press the “Done” button to navigate back to the main screen.
The three buttons on the bottom of the main screen each led to a screen with a collection of buttons that were labelled with different topics, see figure 3.4 for an example and table 3.4 for an overview of the different topics each of the three buttons would lead to.

Table 3.4: ***Topics belonging to actions categories.***

Actions categories	**Ask questions**	**Perform examination**	**Administer treatment**
	Medical history	Look	Medication A-E
	Complaints	Listen	Medication F-N
	Context	Feel	Medication O-Z
		Measure	Other interactions
		Head-to-Toe	

Figure 3.4: ***Example of topics for the category “Perform examination”.*** See table 3.4 for English translation of topics.

Each topic button would lead to a new screen where a collection of buttons labelled with different actions would be presented, see table A.2 in the appendix for an overview of all actions and figure 3.5 for an example.

Figure 3.5: ***Example of actions for the topic "Measure" in "Perform examination".*** See table A.2 in the appendix for translation of actions.

The goal was to present a collection of actions to the participant that they would be able to select during a case simulation, to determine which actions were selected in which order.

The patient information that was shown as the actions' content was extracted from the real life cases that the simulation was based upon. Any information that was missing, as not every action was performed in the original cases, was composed in collaboration with a medical supervisor.

The final collection of actions contained 87 unique actions, divided over three categories; "Ask questions", "Perform examination" and "Administer treatment".

An alternative to this design would be to present all actions at once (in the form of a list, for example), but there are two reasons why this was not the preferred design. First of all, the way the simulation was set up allows information to be seen by the user sequentially. Not only was this necessary to measure when the content of a specific action is seen by the user, but providing sequential information also increases the amount of confirmation bias (Jonas et al., 2001). Whether providing the information in sequential order was more realistic is not clear cut. An ambulance nurse is likely to observe some symptoms simultaneously (especially those that are visual) and others sequentially (those measured with instruments). The second reason to use sub-menus in the design rather than a list was that it is easier to recognize an appropriate action or category than to think of the name of a specific action (Ritter et al., 2013).

Each screen contains a button that allows the participant to navigate back to the main screen and a button that allows them to navigate back to the previous menu. Clicking an action button would show a pop-up with information about that action. This information could

be presented in text, a table or an image (see figure 3.6abc). On this pop-up the participant can click a button labelled “Done”, which would close the pop-up, or a button labelled “Edit differential diagnosis”, which would show them the screen with the DDx section, where they could edit their diagnosis fields.

In a real life ambulance case it often occurs that ambulance nurses make notes during their interaction with the patient, to be able to refer back to this later on. For example, when the heart rate is measured at first they may note that it is elevated, and later check the heart rate again and compare to the previous result for any changes. To facilitate this in the simulation a separate log or note taking functionality could have been implemented, but this would make it less clear which information the participant is observing at the current moment. Instead, any previous information about the action that has changed since the action was last selected was contained within the action’s content.
For example, when a participant selected the action to measure oxygen saturation the previous measurement would also be shown, see figure 3.6b. In the case of a patient's answers to questions, the previous statement made by the patient about this topic was also displayed.

Figure 3.6a: Example of text content in pop-up

Figure 3.6b: Example of table content in pop-up

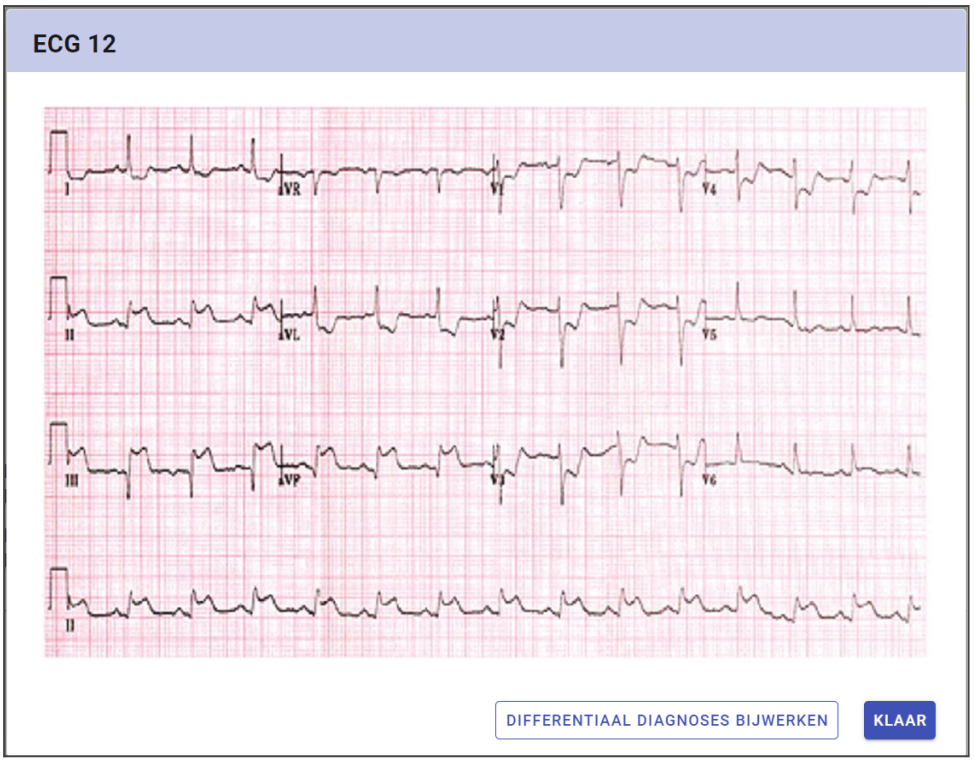

Figure 3.6c: Example of image content in pop-up

Once the participant had selected their first action in a case, the “End case and transfer care” button on the main screen was enabled from that point on.

Every time 4 consecutive actions have been selected without visiting the DDx section, the participant was shown a pop-up in which they were prompted to visit the DDx section, see figure 3.7. They could only navigate to the DDx section from this pop-up. This assured that the participants did not forget to update their DDx list.

The participant was able to select (and re-select) as many actions as they wanted. To end the case they could navigate to the main screen and select the “End case and transfer care” button. The participant would be shown a new screen titled “Final diagnosis”, see figure 3.8.

At the top of the screen the question “What is your final diagnosis?” was presented. The screen contained two dropdown menus and a slider bar, with which the participant could select their working diagnosis, in the same manner in which they would fill out a diagnosis field in the DDx section. At the bottom of the screen the question “To whom will you transfer the care of this patient?” was presented, with three options, of which one could be selected. The three options are: “Hospital”, “General practitioner” and “Patient himself”.
In the bottom right corner a button labelled “Done” was present, which navigated the participant to the next screen when pressed.

Once a participant had pressed “Done” on the final diagnosis screen they were finished with the case. They were shown a text that said that the case was finished and that they could go to the next screen to start a new case. The text also stated that if the participant did not have

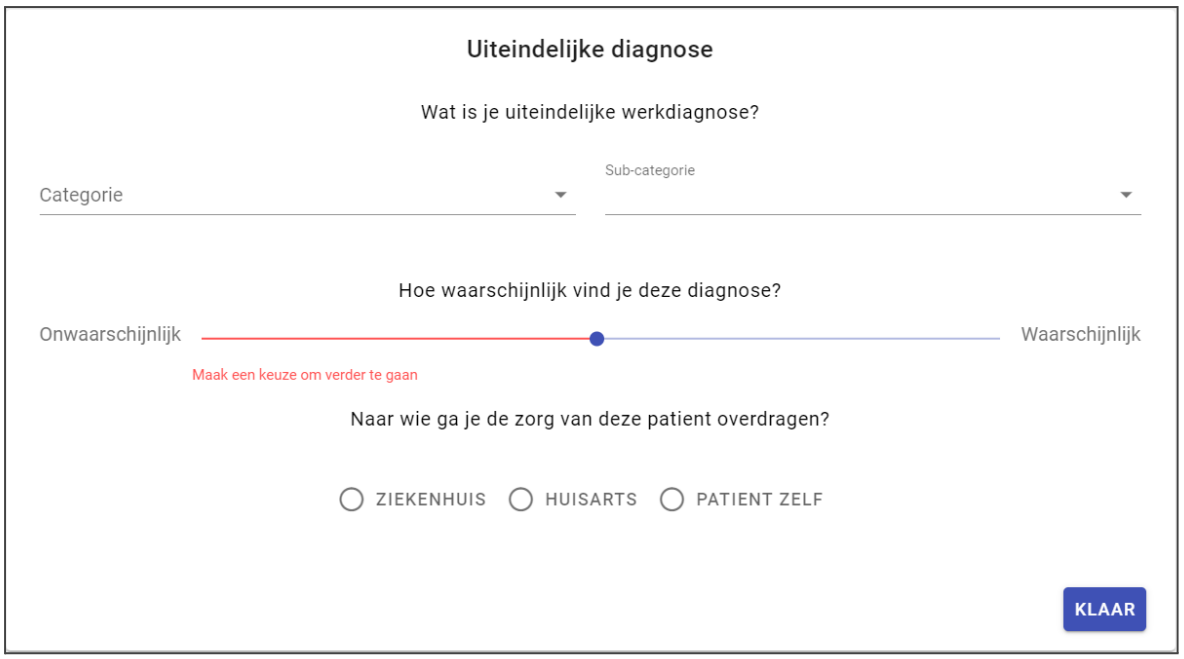

Figure 3.8: ***Final diagnosis screen***

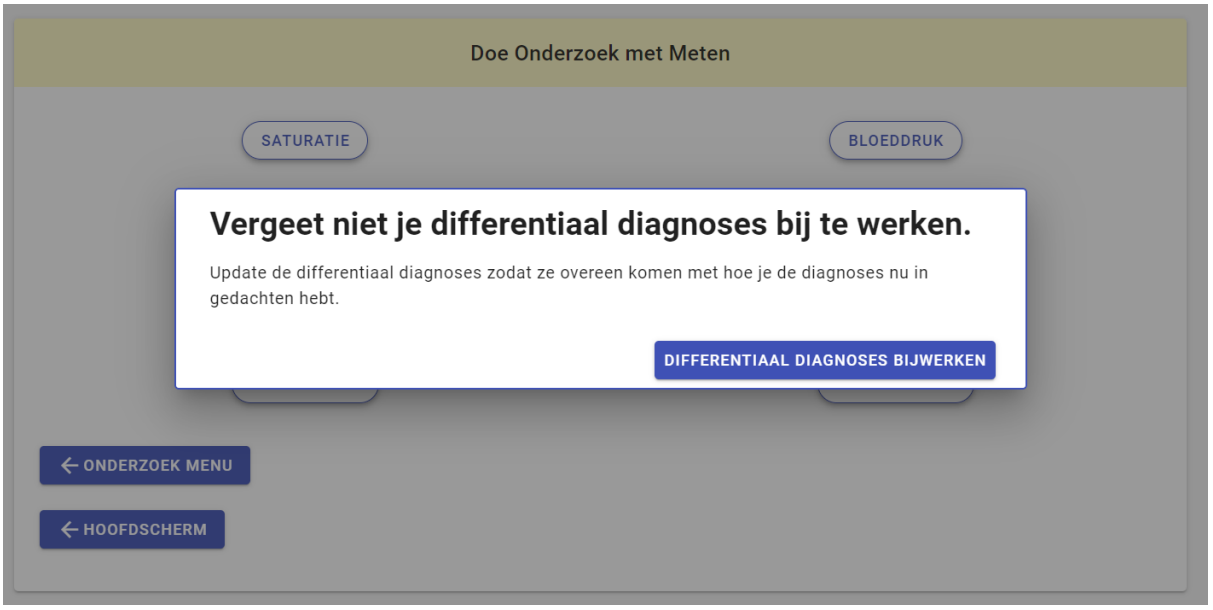

Figure 3.7: ***DDx section prompt pop-up***. The translated pop-up text: “Don’t forget to update your differential diagnosis. Update the differential diagnoses so they match the diagnoses you have in mind at the moment.”

time at the present moment they could return to the experiment at a later moment to start a new case. If the participant had finished both cases they were shown a text that said that they had finished all cases and they were thanked for their participation.

3.3 Procedure

Participants could join the experiment at any time during the period in which the experiment ran. As the experiment was designed to measure confirmation bias, it was necessary to

carefully consider under what pretence participants would join the experiment. To this end it was decided to not ask whether participants had any training in bias awareness, as was suggested by Parmley (2006), to decrease the chance of influencing the results. Instead the information in both the invitation email and on the website explained that the experiment was about clinical reasoning and that the scenarios were based on old cases. There was no mention about cognitive biases or confirmation bias in any of the given information.

To be able to keep track of individual participants' data, participants had to register an account with their email address when they first joined the experiment. They are asked to provide their information, namely: gender, age, years of work experience as an ambulance nurse, previous medical fields (if any) and total years of work experience as a nurse.

Accounts had no passwords, as participants may forget their password and become unable to resume the experiment at a later date. Participants could not access any recorded task information directly, so if anyone were to log in with the wrong email address there would be no issue regarding privacy concerns of the participant’s data.
Allowing the participants in this experiment to register with their email address, makes it possible to send reminder mails to encourage participants to finish any remaining cases. Reminder emails were sent every week to those who joined the experiment to encourage them to finish a second case. Participants received a maximum of two reminder emails. Registering with their email address also solves the issue of participants forgetting their login credentials, as they are more likely to remember their own email address than a randomly assigned subject number.

Upon registering their account participants were given a tutorial about the experiment. In this tutorial the different components of the simulation environment were shown and explained in text. The components could be interacted with on a stand-alone basis, to be able to get familiar with their mechanisms and the medical content. Following the tutorial may help close the gap between participants with different levels of familiarity with digital environments.

The tutorial also contained extra information regarding the experiment. Since this experiment was performed in 2022 the information provided to the participants explicitly states that cases in the simulation are from the year 2017 or 2018, to exclude any COVID-19 related diagnoses or actions. Participants were also asked not to discuss the cases with colleagues, so new participants would not join the experiment with prior knowledge of the content.

After the tutorial had been finished, participants were shown a screen where they could return to the tutorial or start a case.

After the participants finished a case they were able to leave feedback about the experiment.

After the experiment period concluded all participants were sent a debriefing email, in which the purpose of the study was explained.

3.4 Materials

We refer to the part of the website in which the task was performed as the simulation, as it simulates an ambulance case.

To validate the medical content of the simulation all choices and content were validated by a medical professional with experience in the field of pre-hospital emergency medicine.
The selection of working diagnoses and actions that a participant could choose from are the same for both cases.

3.4.1 Simulation software

The website on which the task could be accessed was built with custom software, see appendix B.1 for more information about this decision. The website was built on NodeJS with a VueJS frontend and an Express backend. MySQL was used as the database.

3.4.2 Actions

The collection of actions provided a comprehensive overview of all actions that would be relevant in non-trauma cases.

To gather an extensive list of actions for this simulation, several resources were used. First all actions that were performed in the real medical cases, upon which the two simulation cases are based, were extracted. Then the actions performed in standard protocols for non-trauma cases involving male patients were added (NHG, 2021; Søreide, 2008; Zemaitis et al., 2023). For treatment options all medication that is typically available in an ambulance are added as an action. This collection of actions was verified by the medical supervisor of this experiment. After the pilot several actions were added. The category “Administer treatment” was expanded with more treatment options besides administering medicine, they can be found under the topic “Other interactions”, see table A.2 in the appendix. To the medicine options “ticagrelor” and “heparin” were added. Under the category ““Perform examination” the action “auscultation heart” was added under the topic “listen”.

3.4.3 Differential diagnoses

Several steps were taken to create a collection of diagnoses that participants would be able to choose from. First of all, all differential diagnoses that were mentioned in the reports of the two medical cases upon which the simulation cases are based were included. Then a list of DDx related to these diagnoses was built with the help of the diagnostic tool Diagnosaurus (Zeiger, 2014), further confirmed by the ICD10 database (World Health Organization, 2019) and the medical supervisor. As a result of the pilot, several diagnosis options were added, see table 3.5.

Table 3.5: ***Diagnoses options added after pilot***

urine retention
panic disorder
hypo/hyperglycemie
no diagnosis

Diagnoses that are not relevant for non-trauma adult male patients were not included. Diagnoses were also excluded based on a high specificity, as they would be unlikely to be

diagnosed during an ambulance case, because they need further testing (such as cancer diagnoses).

3.4.4 Interactive patient system

Table 3.6: ***Treatment effects on patient information.***

	Underlying patient variables							
		Pain	Oxygen Saturation	Awareness	Heart rate	Blood pressure	Blood sugar	Nausea
Medication	Adrenaline				Increase	Increase	Increase	
	Atropine Sulfate				Increase			
	Esketamine	Decrease		Decrease	Increase	Increase		
	Fentanyl	Decrease		Decrease	Decrease	Decrease		Increase
	Glucose						Increase	
	Glucagon						Increase	
	Midazolam				Decrease	Decrease		
	Morphine	Decrease		Decrease	Decrease	Decrease		Increase
	Naloxone**	Increase		Increase	Increase	Increase		Decrease
	Nitroglycerin*	Decrease			Decrease	Decrease		
	Ondansetron							Decrease
	Paracetamol*	Decrease			Decrease	Decrease		
	Oxygen		Increase					

Note. The exact increase or decrease of symptoms were dependent on the case.
* Only effective in case 2
** This medication is often given to counter the effect of drugs. For the patients in the 2 simulated cases this medicine did not have any effect, except in the case where they were given morphine or fentanyl, in which case the effects of these painkillers were countered.

Both Mendel et al. (2011) and Parmley (2006) suggest that it would benefit the experimental set-up to include patient interactions, so the participant can react to the patient information. In the simulation this was implemented by letting the different treatments affect the underlying patient information. For example, if oxygen was administered, then the patient's

oxygen saturation would increase. This change in the underlying patient variables could then be observed in the vital measurements and the patient's responses to questions. See table 3.6 for a simplified overview of the treatment interaction.

The system also kept track of whether morphine, fentanyl, nitroglycerin and painkillers were administered at any point. For morphine and fentanyl it was relevant to know whether administering naloxone would have any effect on the patient. For nitroglycerin and painkillers it was to make sure that administering more would not necessarily increase their effects.

As can be seen in table 3.6, administering treatment can also induce new symptoms. For example, administering fentanyl can cause nausea in the patient.

The interaction system was not fully realistic. The amount of medicine that could be administered was a fixed dose. The amount of times that administering the same medication would have an effect was limited to two times. The influence of treatments on the underlying patient information in the simulation should be viewed as a simplified replica of how a real patient would react to treatment.

3.5 Measures

Differential Diagnosis
Every time a participant finished editing the DDx section, the list of all diagnoses and their likelihoods were recorded together with the current time. This allowed the lists to be reviewed later on in the order in which they were created.

Actions
All actions that the participant selected were recorded, as well as the order in which they were selected.

Final diagnosis
Both the final diagnosis and its likelihood was recorded together with the choice for the transfer of care.

3.6 Data Analysis

Case progression
To measure the progression through the case the amount of times that the DDx sections had been edited was counted. The DDx edits were chosen instead of the amount of time that has passed because the time it takes to go through the case is dependent on how quick a participant read and which actions they selected (some action content contained long pieces of text).

Working diagnosis
To be able to analyse the working diagnosis the differential diagnosis with the highest likelihood was selected from every DDx list and designated as the working diagnosis at that moment.

Working experience
For the working experience we used the number of years that the participants have worked as an ambulance nurse.

Experts - Novices
The amount of expertise was determined by how much working experience a participant has. Expertise was treated as a scale and not a category.

Labelling actions and differential diagnoses
The different conditions don't prime one specific diagnosis, but rather a small group of diagnoses. To determine whether participants were affected by this priming it was important to not only know which specific diagnoses they selected, but also whether they selected diagnoses from the primed group or not. This is why the diagnoses that are primed are categorised by the condition and case they are in.
The same rationale applies to the actions selection. To be able to determine whether confirmatory search was taking place it was necessary to label the actions that are positive tests to the primed diagnoses. These actions are either questions or physical examinations of which the expected result was positive given the associate diagnoses. For example, determining whether a patient suffers from incontinence was associated with the diagnosis of epilepsy, because when one assumes an epilepsy seizure has occurred they expect the patient to likely suffer from incontinence. The overview of primed diagnoses with their associated actions can be found in Table A.3 in the appendix.

Confirmatory search
We measured the amount of confirmatory search by how many actions were selected that were positive tests for a primed diagnosis in a specific condition.

Under-weighting counter-evidence
We measured the amount of under-weighing evidence by how many primed diagnoses were selected in the non-congruent condition.

3.6.1 Statistical analysis

Expertise and confirmation bias
The goal was to determine what the influence of working experience was on confirmatory search (selecting actions) and under-weighting counter-evidence (selecting working diagnoses). Confirmatory search and under-weighing evidence are tested separately. For both tests a multinomial model was used, because both the dependent and independent variables contain categorical data and the model was too complex to use a simpler test.

Expertise and confirmatory search
To determine what the influence was of the *work experience*, *progress of the case* and the *conditions* on action selection a multinomial model was used. The dependent variable in this model was the category that an action belongs to.
The data from case 1 and case 2 was tested separately, but the same model was used.

The multinomial model consisted of the interaction *Work Experience* x *Case Progress* x *Condition* and all its nested terms. The model tried to predict from which category an action was selected.
This specific model was chosen to determine whether Experts (with high *work experience*) were more likely to select less of the non-congruent primed actions in the non-congruent condition and less of the congruent primed actions in the congruent condition over the progress of the case. This would show that experts were less likely to engage in positive testing than novices.

Expertise and under-weighing evidence
To find out what the influence of *work experience, progress of the case* and the *conditions* were on the selection of a working diagnosis a multinomial model was used. To make the model easier to interpret the dependent variable was the category that the working diagnosis belongs to, instead of the specific diagnosis.
The data from case 1 and case 2 was tested separately, but the same model was used.

The multinomial model consists of the interaction *Work Experience* x *Case Progress* x *Condition* and all its nested terms. The model tried to predict which category the working diagnosis belonged to. This model was chosen to determine whether Novices (with low *work experience*) were more likely to select a non-congruent primed working diagnosis in the non-congruent condition over the progress of the case. This would show that novices were more likely to engage in over-weighting evidence that was in favour of their current working diagnosis than experts.

Effect of conditions
The goal was to determine whether the different conditions had an influence on which diagnoses were selected by the participants.
The data of case 1 and case 2 were tested separately, as different diagnoses are primed depending on the case.

The difference in the occurrence and likelihood of selected diagnosis between conditions was tested with different parts of the data. The first part of the data was only the diagnoses that were selected the first time that participants filled out their DDx. The second part of the data were all diagnoses that were selected throughout the case, apart from the final diagnosis. And the last part of the data consists only of the final diagnoses.

For all three parts of the data Dunn's test was used to determine whether there was a difference in occurrence and assigned likelihood of the selected diagnoses between the conditions.

Explorative
A part of the data analysis was exploratory in nature.

For the final diagnoses the likelihood regardless of diagnoses was compared between conditions with a Kruskal-Wallis test and Dunn's test as post-hoc. The choice of care transfer was also compared between conditions with a Kruskal-Wallis test.

The difference in the number of actions selected between the two cases was tested with a paired t-test and the difference in the number of actions selected between the conditions in each case were tested with an ANOVA test. The influence of working experience on the number of selected actions was tested with a simple linear regression. The influence that the conditions had on which specific actions were selected was tested with Dunn's test.

The difference in the number of times that diagnoses were updated between the two cases was tested with a paired t-test and the same was tested between the conditions in each case with an ANOVA. The difference in the number of DDx within a DDx list was also tested, between the cases with a paired t-test and between the conditions with an ANOVA.
The influence of work experience on the assigned likelihood of working diagnoses over the progression of the case was tested for both cases with a linear regression.

STATA (StataCorp LLC, 2021) was used for the multinomial models. All other tests were done in R (The R Foundation for Statistical Computing, 2021).

4 Results

4.1 Participants

The participants consisted of 47 ambulance nurses who finished one or more cases in the simulation (25 men and 22 women). Participants' ages ranged from 28 to 62 years (M = 46.3, SD = 10.0). Participants' experience as ambulance nurses ranged from 0 to 35 years (M = 14.0, SD = 10.6) and participants' total experience as a nurse ranged from 1 to 44 years (M = 24.8, SD = 11.4).

All nurses had previous medical experience, see table 4.1.

Table 4.1: ***Previous medical experience of participants***

Specialisation	**Number of participants**
Anaesthesiology	8
Cardio Care Unit	16
Intensive Care Unit	23
Emergency department	11
Other	19
None	0

4.1.1 Number of completed cases

In this analysis only completed cases were analysed, which were defined as cases that have a completed final diagnosis. There were 47 participants who completed at least one case. 29 of the participants completed both cases and 18 of the participants completed one case, so the total number of completed cases was 76.
55 participants were assigned to case 1 (neurology case) and 36 participants finished case 1 (65.45%). 55 participants were assigned to case 2 (cardiology case) and 40 participants finished case 2 (72.7%). There was no significant difference in task abandonment between the two cases ($X^2(1) = 0.68111$, $p = 0.4092$).
However, there was a significant difference in which order the participants saw the cases that they finished ($X^2(1) = 17.066$, $p = <0.01$), see table 4.2. Most cases that were finished were the first case that participants saw. And most participants who finished cases were first presented with the cardiology case, instead of the neurology case.

For the number of completed cases over case 1 and case 2 for all conditions see table 4.3. There was no dependence between the cases and conditions ($X^2(2) = 1.234$, $p = 0.5396$).

Table 4.2: ***Order of finished cases***

	First case	**Second case**	**Total**
Neurology case	13	23	36
Cardiology case	33	7	40
Total	46	30	76

Table 4.3: ***Number of completed cases.***

	Congruent condition	**Non-congruent condition**	**Control condition**	**Total**
Neurology case	10	15	11	36
Cardiology case	12	12	16	40
Total	22	27	27	76

4.2 Confirmation bias

The goal was to determine the effects of working experience on the amount of confirmation bias in the clinical reasoning of ambulance nurses. The hypothesis was that novice nurses suffer more from confirmation bias in their clinical reasoning and that experienced nurses are able to use dispatch information more efficiently than novice nurses.

Confirmation bias consists of two mechanisms: Confirmatory search and under-weighing counter evidence. We tested for both these mechanisms separately.

In the results the two cases were analysed separately, because these cases differ too much for the data to be analysed as one dataset. For the multinomial model, analysing the two cases as one dataset would have made it necessary to include a four-way interaction, which would make the model overly complex and hard to interpret.

For all models described in this section we will describe the relevant interactions for the research question. The complete overview of the interactions of all models can be found in the appendix, see table A.4-A.7.

4.2.1 Confirmatory search

When a participant engages in confirmatory search they will select tests that they expect will confirm their working diagnosis. To measure confirmatory search the actions that were selected by participants have been categorised. These categories are determined by whether an action was a positive test for a primed diagnosis. The overview of primed diagnoses with their associated actions can be found in Table A.3 in the appendix.

It was hypothesised that novices will engage more in confirmatory search than experts and thereby select more actions that are positive tests for primed diagnosis. To determine what the influence was of the *work experience*, *progress of the case* and the *conditions* on action selection a multinomial model was used. The dependent variable in this model was the category that an action belongs to. The models are presented by case. Both cases use the same model, as explained below.

The multinomial model consisted of the interaction *Work Experience* x *Case progression* x *Condition* and all its nested terms. The model tried to predict from which category an action was selected. The full model was selected because the three way interaction is the interaction that shows the effect in which we are interested. Models that were more complicated did not achieve a higher AIC in case 1 or case 2.

The base outcome of the model (the reference group) was the category of actions that do not belong to either the congruent or non-congruent diagnoses. For the condition variable the reference group was the control condition. All statistical significance was dependent on the specific terms of this model with this specific reference group.

4.2.1.1 Case 1 (Neurology)

In case 1 none of the main terms in the model were significant and neither were any interactions. The only significant effect was the intercept.

The question was whether experienced nurses (with high *work experience*) were more likely to select less actions from the non-congruent category in the non-congruent condition and less actions from the congruent category in the congruent condition over the progress of the case. The effect of work experience over the course of the case was not significant in any of the conditions, see table A.4 in the appendix for all the multinomial model results.

However, case 1 shows that the intercept for actions from both the congruent and non-congruent category were significantly different from the neutral category (see Table 4.4). This means that without the influence of work experience, conditions and case progression participants were more likely to select neutral actions than actions from the congruent or non-congruent category.

Table 4.4: ***Significant univariate effects in action selection of case 1***

Variables	Category	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
Intercept	non-congruent actions	-4.012381	1.064985	-3.77	0	-6.099713	-1.925048
Intercept	congruent actions	-3.111454	0.7330461	-4.24	0	-4.548198	-1.67471

4.2.1.2 Case 2 (Cardiology)

In case 2 the progression of the case had a significant effect on the amount of actions from the non-congruent category that were chosen relative to neutral actions (see table 4.5). As the case progressed participants were less likely to select non-congruent actions.

The question was whether experienced nurses (with high *work experience*) were more likely to select less actions from the non-congruent category in the non-congruent condition and less actions from the congruent category in the congruent condition over the progress of the case.The effect of working experience and the conditions were not significant and neither was the interaction between working experience, conditions and the progression of the case. See table A.5 in the appendix for all the multinomial model results.

Case 2 also showed that the intercept for actions from both the congruent and non-congruent category was significantly different from the neutral category (see table 4.5). This means that without the influence of work experience, conditions and case progression participants were more likely to select neutral actions than actions from the congruent or non-congruent category.

Table 4.5: ***Significant univariate effects in action selection of case 2***

Variables	Category	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
Case progression	non-congruent actions	-0.0390072	0.0152328	-2.56	0.01	-0.0688629	-0.0091516
Intercept	non-congruent actions	-0.9908031	0.3324028	-2.98	0.003	-1.642301	-0.3393055
Intercept	congruent actions	-1.066744	0.2829552	-3.77	0	-1.621326	-0.5121623

4.2.2 Under-weighing evidence

When a participant engages in under-weighing counter-evidence they will be less likely to change their working diagnosis when confronted with counter-evidence to their current working diagnosis. To measure under-weighing evidence the diagnoses that were selected by participants have been categorised. These categories are determined by whether a diagnosis was primed by the dispatch messages in the congruent or non-congruent condition. The overview of primed diagnoses with their associated actions can be found in Table A.3 in the appendix.

Based on the theoretical framework, experienced nurses were expected to select less non-congruent diagnoses in the non-congruent condition than novices, but experienced nurses were not expected to select less congruent diagnoses in the congruent condition than novice nurses. It was expected that this would lead to experienced nurses selecting more congruent final diagnoses than novice nurses over all conditions.

To find out what the influence of *work experience, progress of the case* and the *conditions* were on the selection of a working diagnosis a multinomial model was used. To make the

model easier to interpret the dependent variable was the category that the working diagnosis belongs to, instead of the specific diagnosis. The models are presented by case. Both cases used the same model, as explained below.

The multinomial model consists of the interaction *Work Experience* x *Case progression* x *Condition* and all its nested terms. The model tried to predict which category the working diagnosis belonged to. The full model was selected because the three way interaction is the interaction that shows the effect in which we are interested. Models that were more complicated did not achieve a higher AIC in case 1 or case 2.

The base outcome of the model (the reference group) was the category of non-primed diagnosis. For the condition the reference group was the control condition. All statistical significance is dependent on the specific terms of this model with this specific reference group.

4.2.2.1 Case 1 (Neurology)

In case 1 the main effects of work experience and condition were significant, see table 4.6. The main effect of progress over the case was not significant. See table A.6 in the appendix for all the multinomial model results.

The effect of working experience indicated that experienced nurses were more likely to select congruent diagnoses than novice nurses. Between the conditions the difference between the congruent and the control condition was significant, which indicated that participants in the congruent conditions were more likely to select congruent diagnosis rather than neutral diagnoses, relative to participants in the control condition.

The question was whether novice nurses (with low *work experience*) were more likely to select a working diagnosis from the non-congruent category in the non-congruent condition over the progress of the case than experienced nurses. This would show that novices were more likely to engage in under-weighing counter evidence that was not in favour of the working diagnosis.

The interaction between work experience, the progression of the case and the conditions was significant in both the congruent and non-congruent condition, compared to the control condition, see figure 4.1abc. This indicates that in both the congruent and non-congruent condition more experienced nurses were more likely to select a congruent diagnosis rather than a neutral diagnosis over the progression of the case. However, this interaction did not show that novices were significantly more likely to choose non-congruent diagnosis in the non-congruent condition, as the research question states.

Although work experience had a significant influence in the selection of working diagnoses over the whole case, in the final diagnosis there was no difference in work experience by those who selected a congruent (M= 16.5 , SD= 10.58), non-congruent (M= 13.33 , SD= 13.05) and neutral diagnosis (M= 14.43 , SD= 9.68); Kruskal-Wallis, $X^2(2) = 0.3727$, p=0.83.

Table 4.6: ***Significant univariate effects and three-way interactions in working diagnosis selection of case 1***

Variables	Category	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
Work experience	congruent diagnoses	0.1538931	0.0524359	2.93	0.003	0.0511207	0.2566656
Congruent condition	congruent diagnoses	3.226684	1.139954	2.83	0.005	0.9924149	5.460952
Congruent condition x Work experience x Case progress	congruent diagnoses	0.0386635	0.0123261	3.14	0.002	0.0145047	0.0628223
Non-congruent condition x Work experience x Case progress	congruent diagnoses	0.0407176	0.0117629	3.46	0.001	0.0176626	0.0637725

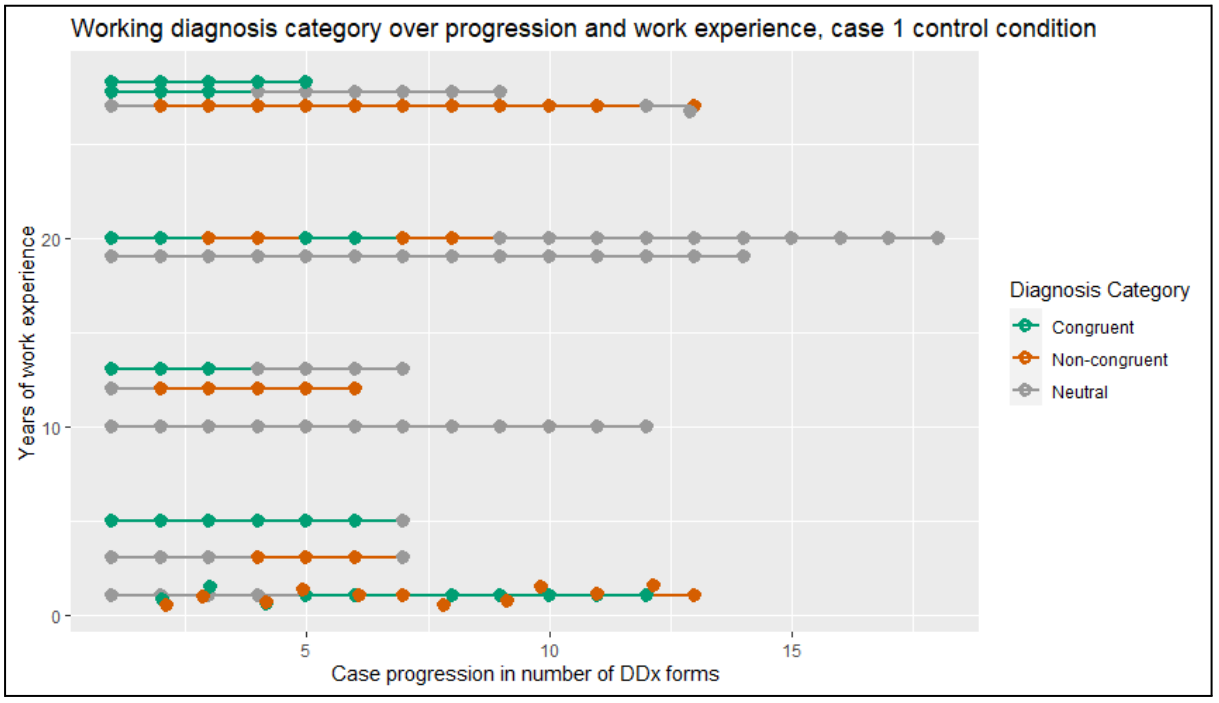

Figure 4.1a: ***Working diagnosis category over work experience and progression of case 1 in the control condition***

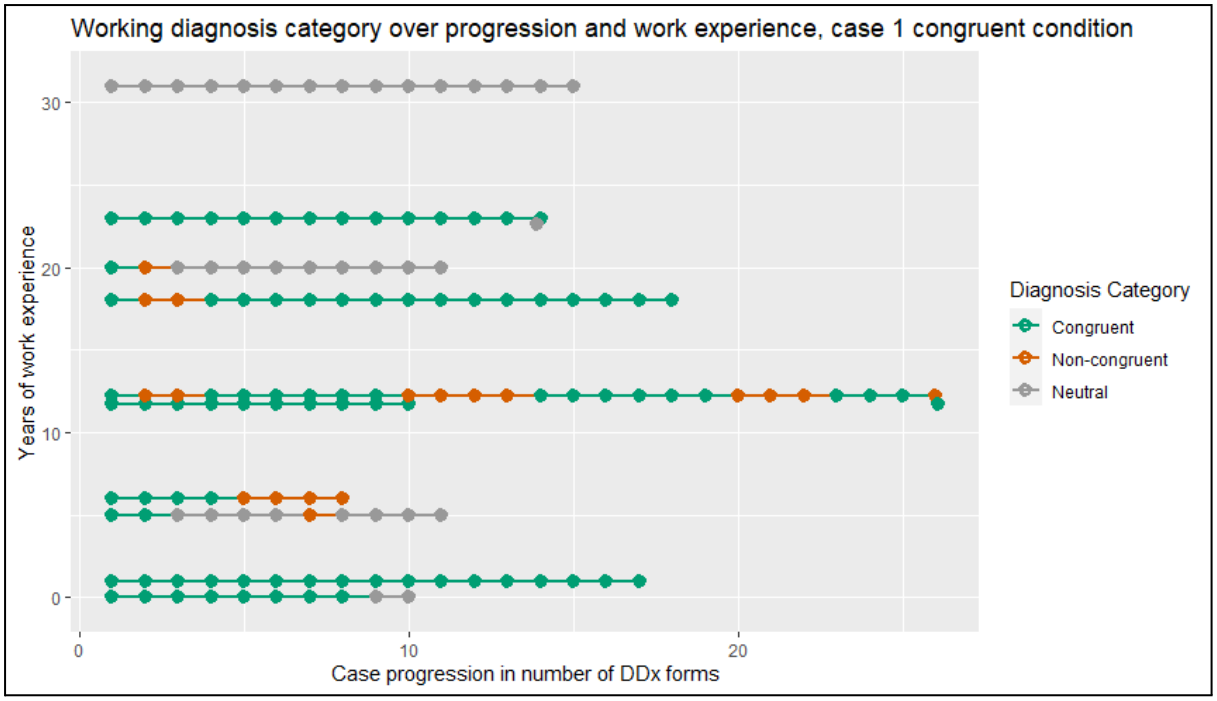

Figure 4.1b: ***Working diagnosis category over work experience and progression of case 1 in the congruent condition***

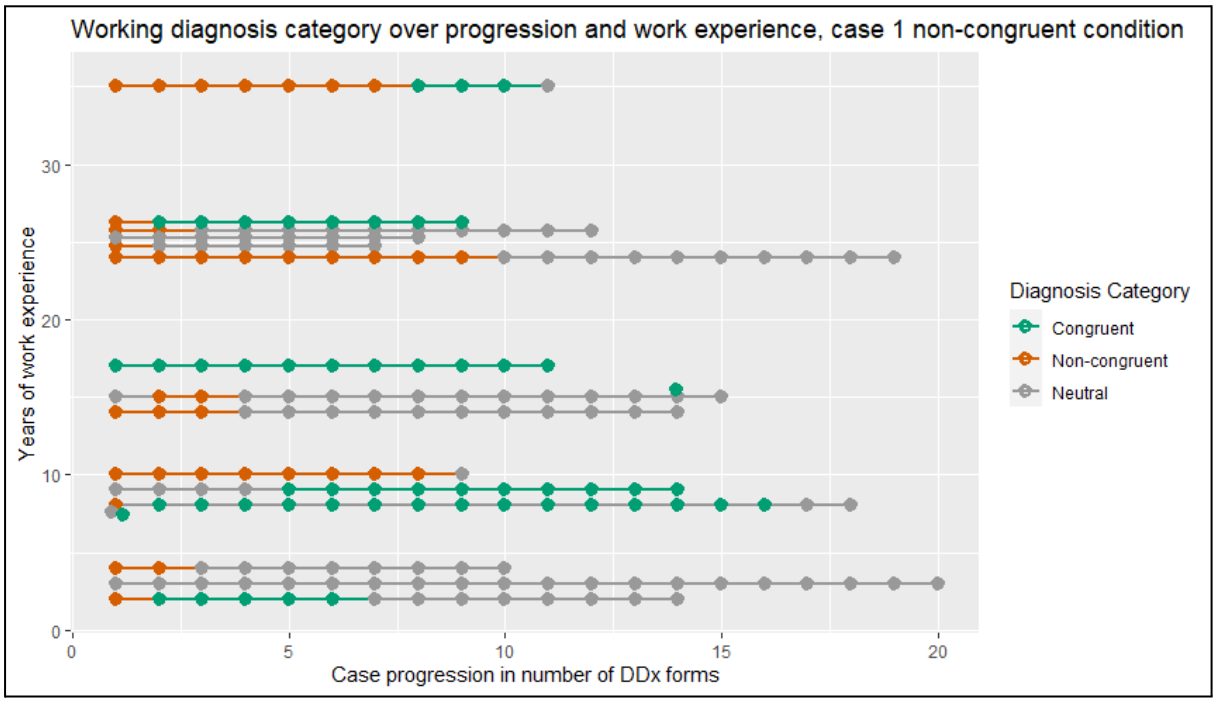

Figure 4.1c: ***Working diagnosis category over work experience and progression of case 1 in the non-congruent condition***

4.2.2.2 Case 2 (Cardiology)

In case 2 the main effect of case progression was significant, see table 4.7. The main effects of working experience and the conditions were not significant. See table A.7 in the appendix for all the multinomial model results.

The case progression had a significant effect on the number of diagnoses from the congruent category that were chosen relative to neutral actions. As the case progressed participants were more likely to select congruent actions, see figure 4.2.

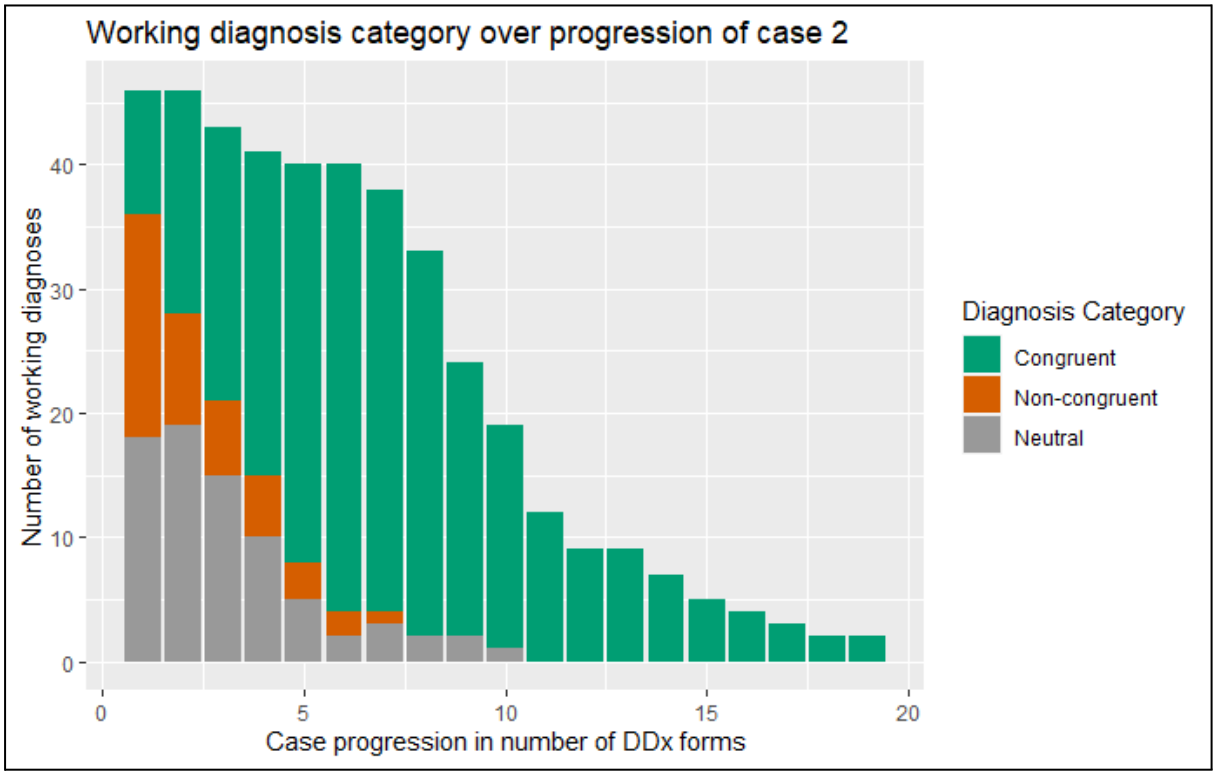

Figure 4.2: ***Working diagnosis category over progression of case 2***

The question was whether novice nurses (with low *work experience*) were more likely to select a working diagnosis from the non-congruent category in the non-congruent condition over the progress of the case than experienced nurses. This would show that novices were more likely to engage in under-weighing counter evidence that is not in favour of the working diagnosis.
The effect of work experience over the course of the case was not significant in any of the conditions.

Case 2 also showed that the intercept for diagnoses from the congruent category were significantly different from the neutral category (see table 4.7). This means that without the influence of work experience, conditions and case progression participants were more likely to select neutral diagnoses than diagnoses from the congruent category.

In the final diagnosis of case 2 there was also no difference in work experience by those who selected a congruent, non-congruent or neutral diagnosis, on account of only congruent diagnoses being selected as final diagnosis.

Table 4.7: ***Significant univariate effects in working diagnosis selection of case 2***

Variables	Category	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
Case progress	congruent diagnoses	0.4070997	0.1164237	3.5	0	0.1789135	0.635286
Intercept	congruent diagnoses	-1.922096	0.6688906	-2.87	0.004	-3.233097	-0.6110941

4.3 The effects of priming

To determine the effects of the congruency of the dispatch information on diagnosing, we used Dunn's test to test whether the number of occurrences of the diagnoses and their assigned likelihood differed between the three conditions. To understand how this changes over the case, this test was repeated separately for a dataset of only the first and final diagnosis. The results will be presented in chronological order and by case, as different diagnoses are primed depending on the case.

It was expected that in the congruent condition the congruent diagnoses would occur more often and with a higher assigned likelihood than in the other conditions. It was also expected that in the non-congruent condition the non-congruent diagnoses would occur more often and with a higher assigned likelihood than in the other conditions.

4.3.1 Influence of dispatch information on first diagnoses selection

The first DDx was the first list of possible diagnoses that was filled out directly after the dispatch information was observed by a participant. In both case 1 and case 2 there was a difference in the diagnosis selected in the first DDx between conditions.

4.3.1.1 Case 1 (Neurology)

All significant differences in the number of occurrences of diagnoses between conditions can be found in table 4.8. All significant differences in assigned likelihood of diagnoses between conditions can be found in table 4.9.

All the congruent diagnoses ("TIA" and "CVA") had significantly more occurrences or a higher assigned likelihood in the congruent condition than the other conditions.

All the non-congruent diagnoses ("Epilepsy") had more occurrences and a higher assigned likelihood in the non-congruent condition than the other conditions.

The neutral diagnosis "acute stress response" occurred significantly more often and with a higher assigned likelihood in the non-congruent condition than in the control condition.

Table 4.8: ***Significant post-hoc tests of the number of occurrences of diagnoses between conditions in case 1 in the first DDx list***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
acute stress response	Neutral	Non-congruent > Control	-2.51	0.012
epilepsy	Non-congruent	Non-congruent > Congruent	4.34	<0.01
epilepsy	Non-congruent	Non-congruent > Control	-4.06	<0.01
TIA	Congruent	Congruent > Non-congruent	-3.14	<0.01
TIA	Congruent	Control > Non-congruent	2.45	0.014

Table 4.9: ***Significant post-hoc tests of the assigned likelihood of diagnoses between conditions in case 1 in the first DDx list***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
acute stress response	Neutral	Non-congruent > Control	-2.56	0.010
CVA	Congruent	Congruent > Non-congruent	-3.19	<0.01
CVA	Congruent	Congruent > Control	-2.77	<0.01
epilepsy	Non-congruent	Non-congruent > Congruent	4.54	<0.01
epilepsy	Non-congruent	Non-congruent > Control	-4.44	<0.01
TIA	Congruent	Congruent > Non-congruent	-3.29	<0.01
TIA	Congruent	Control > Non-congruent	2.19	0.028

4.3.1.2 Case 2 (Cardiology)

All significant differences in the number of occurrences of diagnoses between conditions can be found in table 4.10. All significant differences in assigned likelihood of diagnoses between conditions can be found in table 4.11.

Two of the three congruent diagnoses (“acute myocardial infarction” and “angina pectoris”) occurred significantly more often and with a higher assigned likelihood in the congruent condition than in the control condition. There was no significant difference in the number of occurrences or the assigned likelihood of the congruent diagnosis “acute ischemic heart disease” between conditions.

One of the non-congruent diagnoses (“acute respiratory failure”) occurred significantly more often and with a higher assigned likelihood in the non-congruent condition than in the congruent condition. None of the other non-congruent diagnoses occurred significantly more

often or with a higher assigned likelihood in the non-congruent condition than the other conditions.

Table 4.10: ***Significant post-hoc tests of the number of occurrences of diagnoses between conditions in case 2 in the first DDx list***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
acute respiratory failure	Non-congruent	Non-congruent > Congruent	2.91	0.0036
acute respiratory failure	Non-congruent	Control > Congruent	3.78	<0.01
status asthmaticus	Non-congruent	Control > Congruent	2.53	0.011
acute myocardial infarction	Congruent	Congruent > Non-congruent	-3.29	<0.01
angina pectoris	Congruent	Congruent > Non-congruent	-3.39	<0.01
angina pectoris	Congruent	Congruent > Control	-3.62	<0.01
cardiac arrhythmias	Neutral	Control > Non-congruent	2.56	0.010

Table 4.11: ***Significant post-hoc tests of the assigned likelihood of diagnoses between conditions in case 2 in the first DDx list***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
acute respiratory failure	Non-congruent	Non-congruent > Congruent	2.62	<0.01
acute respiratory failure	Non-congruent	Control > Congruent	3.25	<0.01
acute myocardial infarction	Congruent	Congruent > Non-congruent	-3.58	<0.01
acute myocardial infarction	Congruent	Congruent > Control	-2.40	0.016
angina pectoris	Congruent	Congruent > Non-congruent	-3.37	<0.01
angina pectoris	Congruent	Congruent > Control	-3.61	<0.01
cardiac arrhythmias	Neutral	Congruent > Non-congruent	-2.27	0.023
cardiac arrhythmias	Neutral	Non-congruent > Control	2.78	<0.01

4.3.2 Influence of dispatch information on overall diagnoses selection

To determine the effects of the congruency of the dispatch information on diagnosing in general, Dunn's test was used to compare the selected diagnoses from all DDx lists between the three conditions. This excludes the final diagnosis and includes the first DDx list that was filled out.

4.3.2.1 Case 1 (Neurology)

All significant differences in the number of occurrences of diagnoses between conditions can be found in table 4.12. All significant differences in assigned likelihood of diagnoses between conditions can be found in table 4.13.

The neutral diagnosis “encephalitis” occurred significantly more often and with a higher assigned likelihood in the non-congruent condition than in the control condition.

Table 4.12: ***Significant post-hoc tests of the number of occurrences of diagnoses between conditions in case 1***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
encephalitis	Neutral	Non-congruent > Control	-2.43	0.015

Table 4.13: ***Significant post-hoc tests of the assigned likelihood of diagnoses between conditions in case 1***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
encephalitis	Neutral	Non-congruent > Control	-2.42	0.015

4.3.2.2 Case 2 (Cardiology)

Table 4.14: ***Significant post-hoc tests of the number of occurrences of diagnoses between conditions in case 2***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
acute respiratory failure	Non-congruent	Non-congruent > Congruent	2.89	<0.01
acute respiratory failure	Non-congruent	Control > Congruent	3.062	<0.01
status asthmaticus	Non-congruent	Control > Congruent	2.46	0.014
acute myocardial infarction	Congruent	Congruent > Non-congruent	-2.75	<0.01

Table 4.15: ***Significant post-hoc tests of the assigned likelihood of diagnoses between conditions in case 2***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
acute respiratory failure	Non-congruent	Non-congruent > Congruent	2.95	<0.01
acute respiratory failure	Non-congruent	Control > Congruent	3.44	<0.01

All significant differences in the number of occurrences of diagnoses between conditions can be found in table 4.14. All significant differences in assigned likelihood of diagnoses between conditions can be found in table 4.15.

One of the three congruent diagnoses (“acute myocardial infarction”) occurred significantly more often in the congruent condition than in the non-congruent condition. None of the other congruent diagnoses occurred significantly more often or with a higher assigned likelihood in the congruent condition than the other conditions.

The non-congruent diagnosis “acute respiratory failure” occurred significantly more often and with a higher assigned likelihood in the non-congruent condition than in the congruent condition. None of the other non-congruent diagnoses occurred significantly more often or with a higher assigned likelihood in the non-congruent condition than the other conditions.

4.3.3 Influence of dispatch information on final diagnosis selection

The final diagnosis was the last diagnosis that participants handed in. Before they could submit their final diagnosis they also had to decide where to transfer the care of the patient to. Care could be transferred to either the hospital, the GP or to the patient themselves. The final diagnosis has more medical relevance than the DDx over the rest of the case, as this diagnosis is always given to whoever will take over the care.

4.3.3.1 Case 1 (Neurology)

There were a total of 14 unique diagnoses selected as the final diagnosis of the 36 finished cases in case 1, spread out over 4 categories (see table 4.16). In both the congruent and non-congruent condition 4 participants chose a congruent final diagnosis, and in the control condition 2 participants did so. There was not a significant difference in the diagnostic accuracy over the conditions ($X^2(2) = 1.2587$, $p = 0.5329$).

Table 4.16: ***Categories of final diagnoses in case 1***

Diagnosis category	Number of diagnoses
Nervous system	18
Other	9
Mental disorders	8
Cardiovascular	1

All significant differences in the number of occurrences of final diagnoses between conditions can be found in table 4.17. All significant differences in assigned likelihood of final diagnoses between conditions can be found in table 4.18.

The non-congruent diagnosis “epilepsy” occurred significantly more often and with a higher assigned likelihood in the control condition than in the other conditions.

Table 4.17: ***Significant post-hoc tests of the number of occurrences of diagnoses between conditions in the final diagnosis of case 1***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
epilepsy	Non-congruent	Control > Congruent	2.23	0.026
epilepsy	Non-congruent	Control > Non-congruent	2.45	0.014

Table 4.18: ***Significant post-hoc tests of the assigned likelihood of diagnoses between conditions in the final diagnosis of case 1***

Diagnosis	Diagnosis category	Comparison of conditions	statistic	p
epilepsy	Non-congruent	Control > Congruent	2.22	0.026
epilepsy	Non-congruent	Control > Non-congruent	2.45	0.014

The average certainty over the conditions, regardless of which diagnosis was selected, had a significant difference (Kruskal-Wallis, $X^2(2) = 6.35750$, $p = 0.0416$). Dunn's test showed that the certainty in the control condition ($M = 75.64$, $SD = 14.79$) was higher than in the congruent condition ($M = 57.50$, $SD = 17.50$) ($p = 0.04313$) and higher than the non-congruent condition ($M = 62.53$, $SD = 15.15$) ($p = 0.04313$).

There are no significant differences in where care was transferred to between the conditions in case 1 (see table 4.19). Although this trend was not significant, in the congruent condition patients always got referred to the hospital or the GP, and the care was never transferred back to the patient themselves. While in the other conditions it did occur that patients have the care transferred back to themselves.

Table 4.19: ***Difference in care transfer decision between conditions in case 1***

Care transfer	statistic	df	p	method
GP	0.24	2	0.886	Kruskal-Wallis
Patient	1.83	2	0.401	Kruskal-Wallis
Hospital	0.43	2	0.805	Kruskal-Wallis

4.3.3.2 Case 2 (Cardiology)

All final diagnoses in case 2 were congruent diagnoses. There were a total of 2 unique diagnoses selected as the final diagnosis of the 40 finished cases in case 2, both diagnoses within the Cardiovascular category. All participants except one selected the diagnosis of "acute myocardial infarction". One participant, in the control condition, chose the diagnosis of "angina pectoris" instead. There were no significant differences within either occurrence or certainty over the conditions.

In regard to the average certainty over the three conditions, there was no significant effect; Kruskal-Wallis, $X^2(2) = 0.901$, $p = 0.637$. Regarding care transfer, all patients were sent to the hospital without exception.

4.4 Explorative analysis

4.4.1 Assigned likelihood of working diagnosis

The working diagnosis is the leading contender in a list of differential diagnoses. To be able to analyse the working diagnosis, the diagnosis with the highest certainty has been selected from every DDx list.

A simple linear regression was calculated to predict the assigned likelihood of the working diagnoses based on the number of DDx updates by the participant so far. A significant effect was found for case 1, $F(1, 479) = 17.4$, $p < 0.0001$, with the assigned likelihood decreasing by 0.6847 for every update. A significant effect was also found for case 2, $F(1, 474) = 50.51$, $p < 0.0001$, with the certainty increasing by 1.7153 for every update.
However, one participant in case 1 was an outlier in the number of times they have updated their DDx list. The average number of updates was 12.28 ($SD = 4.60$) and this participant updated 26 times. The list they updated also had a very low average assigned likelihood at the end of the case. Removing this outlier still leads to a significant equation, but with the estimation value increasing instead of decreasing, $F(1, 452) = 4.32$, $p = 0.03834$, with the assigned likelihood increasing by 0.3422 for every update.

To determine whether work experience had an influence on the assigned likelihood of the working diagnosis, a simple linear regression was used. A significant effect was found for case 1, $F(1, 479) = 49.91$, $p < 0.0001$, with the assigned likelihood decreasing by 0.56262 for every year of work experience. No significant effect was found for case 2, $F(1, 474) = 0.5229$, $p = 0.47$.
One of the participants had an abnormal pattern, where their working diagnosis throughout the whole case had an assigned likelihood of 1. This participant also had 0 years of work experience, making them an outlier regarding both certainty and work experience. Removing this participant from the dataset leads to a significant result for case 2, $F(1, 458) = 19.25$, $p < 0.0001$, with the assigned likelihood decreasing by 0.29255 for every year of work experience.

There was no significant difference in the assigned likelihood between conditions in case 1 (Kruskal-Wallis, $X^2(2) = 1.174477$, $p = 0.556$). But it was significant for case 2 (Kruskal-Wallis, $X^2(2) = 20.26999$, $p < 0.0001$). A Dunn's test shows that the non-congruent condition ($M = 80.281$, $SD = 17.62$) has a lower assigned likelihood than the congruent condition ($M = 81.784$, $SD = 30.07$, $p < 0.001$) and the control condition ($M = 88.57$, $SD = 12.898$, $p < 0.001$) .
When the previously mentioned outlier was removed from case 1, this showed a significant result (Kruskal-Wallis, $X^2(2) = 8.74496$, $p = 0.0126$) and a Dunn's test showed that the congruent condition ($M = 83.03$, $SD = 11.04$) had a higher assigned likelihood than the non-congruent condition ($M = 77.25$, $SD = 15.65$) ($p = 0.0093$) .

A simple linear regression showed that work experience did not influence the assigned likelihood in the final diagnosis for case 1, (F(1,34) = 1.813, p=0.187) or case 2 (F(1,38) = 1.297, p= 0.2619).

4.4.2 Differences between cases

4.4.2.1 Action selection

Both cases contained 87 available actions. In case 1 overall 61 unique actions were selected out of 87 available actions, in case 2 it was 69 unique actions. In case 2, 8 more actions were selected than in case 1, these 8 actions were all treatment options. All actions chosen in case 1 are also chosen in case 2. All actions of the “Ask questions” and “Perform examination” categories are chosen in case 1 and case 2 at some point.

There was a significant difference in the average number of selected actions between the cases. Case 1 (M= 43.22, SD= 15.97) had more action selections than case 2 (M= 34.23, SD= 12.77), paired t-test: t(28) = 3.1152, p= 0.0042. However, there was no significant difference between conditions in case 1 (ANOVA: F(2,33) = 2.174, p=0.13) or case 2 (ANOVA: F(2,37) = 1.619, p=0.212).
There was no significant effect of years of work experience on the number of actions selected, as shown by a simple linear regression. No significant effect was found for case 1 (F(1,34) = 0.1762, p=0.6773) or for case 2 (F(1,38) = 0.8914, p=0.3511).

Between the conditions there was only one significant difference in action selection. In case 1 the physical examination of extremities occurs more often in the congruent condition (p = 0.0440) and the non-congruent condition (p = 0.0434) than in the control condition.

4.4.2.2 DDx selection

After every four actions the participants were asked to update their DDx list with the differential diagnoses they found likely at that point in time.

There was a significant difference in the average number of times the DDx list was updated between the cases. Case 1 (M= 12.28, SD= 4.60) has more updates per person than case 2 (M= 10.08 , SD= 3.50), paired t-test: t(28) = 2.4132, p = 0.0226. Within case 1 (ANOVA: F(2,33) = 2.155, p=0.132) and case 2 (ANOVA: F(2,37) = 1.68, p=0.20) there was no significant difference in the number of updates between conditions.

There was also no significant difference in how many diagnoses the DDx list contains. Between case 1 (M= 5.18, SD= 2.55) and case 2 (M= 4.63, SD= 2.61) there was no significant difference in the average length of a DDX list; paired t-test: t(28) = 0.93413, p = 0.3582. Within case 1 (ANOVA: F(2,33) = 0.967, p=0.391) and case 2 (ANOVA: F(2,37) = 0.67, p=0.518) there was also no difference between the length of DDX lists between conditions.

A simple linear regression was calculated to predict the length of the DDX lists based on the number of updates by the participant so far. No significant effect was found for case 1, F(1,440) = 0.000239, p=0.988, but there was a significant effect for case 2, F(1,401) = 10.2,

p=0.00151, with the amount of differential diagnoses decreasing by 0.10492 for every update. Over the progression of the case, the DDX lists of the participants became shorter.

To understand whether the amount of work experience had an influence on the length of the DDx list a simple linear regression was also used. A significant equation was found for case 1, $F(1,440) = 17.41$, $p < 0.0001$, with the amount of differential diagnoses decreasing by 0.05192 for every year of work experience. There was no significant equation for case 2, $F(1,401) = 3.56$, p=0.0599.

5 Discussion

5.1 Results

5.1.1 Confirmation bias

In this study, we investigated the effect of working experience on the clinical reasoning of ambulance nurses. Our main research question was whether years of experience influence the amount of confirmation bias that ambulance nurses experience in clinical reasoning. We hypothesised that nurses who had less working experience would be more likely to suffer from confirmation bias than nurses with more working experience.

To answer this question we build an online interactive simulation of a patient encounter in a pre-hospital emergency care setting. Participants were able to select actions in the simulation to be able to ask questions to the patient, do physical examinations and administer treatments. Participants were divided over three conditions, in which the congruence of dispatch information in regard to the patient information was varied. This resulted in a congruent condition, in which dispatch information and patient information suggested the same diagnosis. A non-congruent condition, in which dispatch information suggested a different diagnosis than the patient information. And a control condition, in which the dispatch information did not suggest any diagnostic direction.

To determine whether confirmation bias occurred more often in novice ambulance nurses than in experts, the two main mechanisms that make up confirmation bias were tested for, namely: confirmatory search and under-weighing counter-evidence. Confirmatory search is the practice in which people seek out information that confirms their current hypothesis, in the case of ambulance nurses this would result in a nurse selecting actions which they expect to confirm their assumed diagnosis. Under-weighing evidence occurs when people don’t take evidence that disconfirms their hypothesis into account with the same weight as they take confirming evidence into account. This effect occurs in ambulance nurses when they are reluctant to change their working diagnosis when observing patient information that disconfirms their diagnosis.

The main result of this study was that confirmation bias was not significantly more present in novice nurses in comparison to expert nurses. This can be observed in both the mechanism of confirmatory search and under-weighing evidence.

The research question in regards to confirmatory search was whether novice nurses are more likely to select more actions that would test positive for the suggested diagnosis in the congruent and non-congruent condition than experienced nurses.
The analysis of confirmatory search showed that working experience had no significant effect, which means that novice nurses are not more likely to choose tests they expect to confirm their working diagnosis than expert nurses.

The research question in regards to under-weighing evidence was whether novice nurses are more likely to select a working diagnosis that was suggested in the dispatch information

in the non-congruent condition, and thus inconsistent with the patient information, than experienced nurses. This would show that novices are more likely to engage in under-weighing counter evidence that is not in favour of the working diagnosis.

The analysis of under-weighing evidence showed that working experience had a significant effect in the neurology case, but not in the cardiology case.
In the neurological case, experienced nurses' diagnoses were more in line with the actual disease then those of nurses with less experience, as experienced nurses selected more congruent diagnoses over all conditions. This is in line with previous research, which shows that experts perform better (Hobus et al., 1987; Smith et al., 2013).
However, novice nurses were not more likely to under-weigh counter evidence to the working diagnosis than expert nurses, as they did not select more diagnoses that were inconsistent with the patient information in the non-congruent condition than experienced nurses. Although experienced nurses selected more diagnoses that were consistent with the patient information, the novice nurses' decrease in selecting diagnosis consistent with the patient information was not due to under-weighing counter-evidence.

In addition to confirmation bias not being more present in novice nurses in comparison to experienced nurses, the working experience of nurses also did not influence what their final working diagnosis was at the end of a case or whether they choose to transfer the care of the patient to a hospital, GP or to the patient themselves.

5.1.2 Dispatch information

As an additional research question it was important to determine whether dispatch information increases diagnostic accuracy when it is accurate and decreases diagnostic accuracy when it is inaccurate.
The influence of the dispatch information was present at the start of the case, with more congruent diagnoses occurring in the congruent condition and more non-congruent diagnoses occurring in the non-congruent condition.
However, this effect faded over the progression of the case. The congruency of the dispatch information did not have a significant effect on the selection of the final diagnoses or where the care of the patient was transferred to. This would indicate that missing or incorrect dispatch information is not necessarily detrimental to the diagnosing of the patient.

5.1.3 Other results

5.1.3.1 Overconfidence in novices

In the explorative analysis it was found that there was a significant effect of working experience on assigned likelihood. This means that novice nurses tend to find their working diagnoses more likely than experienced nurses. This may be a sign of overconfidence, which is described as the misalignment between actual competence compared to self-rated expertise (Meyer et al., 2013). Novice nurses found their working diagnosis more likely even though they selected less congruent diagnoses than experienced nurses. Overconfidence among novices has also been found in previous research (Friedman et al., 2005).

5.1.3.2 Differences in cases

There are several major differences found in the results between the two cases. Not only did the effects of working experience and case progression differ, but so did the amount of actions that participants took to finish a case, the order in which cases were often finished and the diagnostic accuracy of the final diagnoses. Although the cases were selected to cover different medical topics, the complexity of a case was not defined specifically. It can be speculated that the cardiology case may have been simpler than the neurology case, resulting in a high diagnostic accuracy and a lower amount of actions selected. In future studies the complexity of the cases should be taken into account in choosing which medical cases to present.

5.1.4 Conclusion

The results show that confirmation bias was not more present in novice ambulance nurses than in experienced ambulance nurses. It also shows that incorrect dispatch information did not have a significant effect on the diagnostic accuracy of final diagnoses. Therefore we can conclude that the confirmation bias is not a major obstacle to clinical reasoning and is unlikely to lead to an increase in diagnostic inaccuracy. Developing a clinical decision support system focused on aiding ambulance nurses in reducing confirmation bias in their clinical reasoning may not be the most efficient way to improve clinical reasoning.

5.2 Limitations

5.2.1 Realism

In this study we simulated ambulance cases, which did not allow for all real life interactions that nurses may have during a case to be represented faithfully. In real ambulance cases the nurse is able to interact with their driver and is able to have remote interactions with the hospital. Participants in the cardiology case gave feedback that they missed the option to consult the cardio care unit remotely, to discuss ECG results and possible treatments. Being able to communicate with others can influence one's reasoning and may either emphasise or play down existing biases about the case information.

Participants in both cases also gave feedback that they wanted to have more interactions with the patient. It was often mentioned that they missed being able to see, smell or hear the patient. For example, in the neurology case several participants mentioned in their feedback that they wanted to know whether the patient smelled of alcohol. The other interaction that the participants missed was being able to ask the patient more questions. The participants specifically wanted to be able to ask more questions about the kind of pain that the patient experiences and the reason why they called for an ambulance in the first place.
Some of this information could be added to the simulation by providing more descriptions of the patient, more options to ask questions or other kinds of media, such as videos or pictures. However, the fact that it is an online simulation of a patient interaction will always limit the sort of information that can be given to a participant.

5.2.2 Assumptions

In the operationalization of the working diagnosis we have made the assumption that the working diagnosis was the diagnosis with the highest assigned likelihood of the DDx. However, we cannot ascertain that this was truly the working diagnosis of the participant at that moment. Not only because participants did not update their DDx after every action, but it may also occur that a participant forgets to update the assigned likelihood of one or more selected diagnoses. It is hard to safeguard against this uncertainty. Forcing the participant to update their DDx after every action would make it very frustrating to participate in the experiment. In the current set-up, in which the participants are prompted to update their DDx every four actions, a few participants already gave feedback that the prompt was annoying and disrupting their workflow.

Within the simulation we also had to make assumptions on what action and diagnoses options would be relevant. These decisions were further validated in a pilot study. Yet in the feedback of the experiment there were several actions and diagnoses that participants wished were available. In the neurology case participants wanted to be able to give a trauma diagnosis and a subarachnoid haemorrhage diagnosis. In the cardiology case participants missed the option to measure CO2, perform a right-sided ECG and consult the cardio care unit.

5.2.3 Remote participation

This study was performed remotely. Participants could access the website that hosted the experiment on their own devices in their own environment. In the instructions participants were asked not to communicate about the content of the experiment with their colleagues, but this cannot be guaranteed. Because this study was about bias, the risk of a participant being influenced by another participant may have had a significant impact on the results.

There is also no guarantee that all nurses who wanted to participate in the experiment were able to do so. Since participants used their own devices (rather than standard equipment in a lab environment) it is possible that the website was not usable on older devices or uncommon web-browsers. This may have influenced the participant selection and may have skewed the participant population to a lower average age, as younger people tend to be more digitally skilled.

5.3 Future research

This study shows that confirmation bias was not more prevalent in novice nurses in comparison to more experienced nurses. This would suggest that there is a different mechanism than confirmation bias at play that influences the performance of ambulance nurses. The results also suggest that incorrect or missing dispatch information may not influence novice nurses in a detrimental way, which would imply that decision support systems are not necessary to debias nurses after they have received the dispatch information.

The influence of work experience was only present in the neurology case, which was a more complex and uncommon case than the cardiology case. In further simulation studies the

complexity and commonness of cases should be varied more, to determine in which circumstances work experience has the largest influence.

The difference between the cases would also suggest that decision support systems could have a greater impact on the performance of novice nurses in more complex or uncommon cases. Further research should examine whether the complexity of a case can be predicted based on the dispatch information or information at the scene that is provided early in the case.

References

Al-Shaqsi, S. (2010). Models of International Emergency Medical Service (EMS) Systems. *Oman Medical Journal*, *25*(4), 320–323. https://doi.org/10.5001/omj.2010.92

Ambulancezorg Nederland. (2021). *SECTORKOMPAS AMBULANCEZORG TABELLENBOEK 2020*. https://www.ambulancezorg.nl/static/upload/raw/7bbd5bed-ec6e-4336-aa7d-5ff65b089745/210920+sectorkompas+ambulancezorg+tabellenboek+2020.pdf

Andersson, U., Andersson Hagiwara, M., Wireklint Sundström, B., Andersson, H., & Maurin Söderholm, H. (2022). Clinical Reasoning among Registered Nurses in Emergency Medical Services: A Case Study. *Journal of Cognitive Engineering and Decision Making*, *16*(3), 123–156. https://doi.org/10.1177/15553434221097788

Andersson, U., Maurin Söderholm, H., Wireklint Sundström, B., Andersson Hagiwara, M., & Andersson, H. (2019). Clinical reasoning in the emergency medical services: An integrative review. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *27*(1), 76. https://doi.org/10.1186/s13049-019-0646-y

Arkes, H. R., & Harkness, A. R. (1980). Effect of making a diagnosis on subsequent recognition of symptoms. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(5), 568–575. https://doi.org/10.1037/0278-7393.6.5.568

Arocha, J. F., & Patel, V. L. (1995). Novice Diagnostic Reasoning in Medicine: Accounting for Evidence. *The Journal of the Learning Sciences*, *4*(4), 355–384.

Baila, J. I. (1980). Logical Thinking and the Diagnostic Process. *Methods of Information in Medicine*, *19*(2), 88–92. https://doi.org/10.1055/s-0038-1635267

Bashiri, A., Alizadeh Savareh, B., & Ghazisaeedi, M. (2019). Promotion of prehospital emergency care through clinical decision support systems: Opportunities and challenges. *Clinical and Experimental Emergency Medicine*, *6*(4), 288–296. https://doi.org/10.15441/ceem.18.032

Berge, K. van den, & Mamede, S. (2013). Cognitive diagnostic error in internal medicine. *European Journal of Internal Medicine*, *24*(6), 525–529. https://doi.org/10.1016/j.ejim.2013.03.006

Bilalić, M., McLeod, P., & Gobet, F. (2008). Why good thoughts block better ones: The mechanism of the pernicious Einstellung (set) effect. *Cognition*, *108*(3), 652–661. https://doi.org/10.1016/j.cognition.2008.05.005

Bond, R. R., Novotny, T., Andrsova, I., Koc, L., Sisakova, M., Finlay, D., Guldenring, D., McLaughlin, J., Peace, A., McGilligan, V., Leslie, S. J., Wang, H., & Malik, M. (2018). Automation bias in medicine: The influence of automated diagnoses on interpreter accuracy and uncertainty when reading electrocardiograms. *Journal of Electrocardiology*, *51*(6, Supplement), S6–S11. https://doi.org/10.1016/j.jelectrocard.2018.08.007

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology*, *4*(1), 55–81. https://doi.org/10.1016/0010-0285(73)90004-2

Croskerry, P. (2002). Achieving Quality in Clinical Decision Making: Cognitive Strategies and Detection of Bias. *Academic Emergency Medicine*, *9*(11), 1184–1204. https://doi.org/10.1197/aemj.9.11.1184

Darley, J. M., & Gross, P. H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20–33. https://doi.org/10.1037/0022-3514.44.1.20

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica*, *43*(2), 111–121. https://doi.org/10.1016/0001-6918(79)90017-9

Doi, K. (2007). Computer-aided diagnosis in medical imaging: Historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, *31*(4), 198–211. https://doi.org/10.1016/j.compmedimag.2007.02.002

Durning, S. J., Artino, A. R., Boulet, J. R., Dorrance, K., van der Vleuten, C., & Schuwirth, L. (2012). The impact of selected contextual factors on experts' clinical reasoning performance (does context impact clinical reasoning performance in experts?).

Advances in Health Sciences Education, *17*(1), 65–79. https://doi.org/10.1007/s10459-011-9294-3

Einhorn, H. J., & Hogarth, R. M. (1978). Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, *85*(5), 395–416. https://doi.org/10.1037/0033-295X.85.5.395

Elstein, A. S., Shulman, L. S., & Sprafka, S. A. (1990). Medical Problem Solving: A Ten-Year Retrospective. *Evaluation & the Health Professions*, *13*(1), 5–36. https://doi.org/10.1177/016327879001300102

Ericsson, K. A., & Smith, J. (2011). Prospects and limits of the empirical study of expertise: An introduction. *Foundations of Cognitive Psychology*, 393–424.

Eva, K. W., & Cunnington, J. P. W. (2006). The difficulty with experience: Does practice increase susceptibility to premature closure? *Journal of Continuing Education in the Health Professions*, *26*(3), 192–198. https://doi.org/10.1002/chp.69

Feltovich, P. J., Coulson, R., & Spiro, R. (2001). *Learners' (mis)understanding of important and difficult concepts: A challenge to smart machines in education*. 349–375.

Feltovich, P. J., Johnson, P. E., Moller, J. H., & Swanson, D. B. (1984). *LCS: The Role and Development of Medical Knowledge in Diagnostic Expertise*. 44.

Folk, C. L., Remington, R., Boehm-Davis, D. A., & Boehm-Davis, D. A. (2012). *Introduction to Humans in Engineered Systems*. John Wiley & Sons, Incorporated. http://ebookcentral.proquest.com/lib/rug/detail.action?docID=875841

Friedman, C. P., Gatti, G. G., Franz, T. M., Murphy, G. C., Wolf, F. M., Heckerling, P. S., Fine, P. L., Miller, T. M., & Elstein, A. S. (2005). Do physicians know when their diagnoses are correct? *Journal of General Internal Medicine*, *20*(4), 334–339. https://doi.org/10.1111/j.1525-1497.2005.30145.x

Gunnarsson, B.-M., & Warrén Stomberg, M. (2009). Factors influencing decision making among ambulance nurses in emergency care situations. *International Emergency Nursing*, *17*(2), 83–89. https://doi.org/10.1016/j.ienj.2008.10.004

Hajioff, S. (1998). Computerized decision support systems: An overview. *Health Informatics*

Journal, *4*(1), 23–28. https://doi.org/10.1177/146045829800400104

Hammond, K. R., Hamm, R. M., Grassia, J., & Pearson, T. (1987). Direct comparison of the efficacy of intuitive and analytical cognition in expert judgment. *IEEE Transactions on Systems, Man, and Cybernetics*, *17*(5), 753–770. https://doi.org/10.1109/TSMC.1987.6499282

Hobus, P. P., Schmidt, H. G., Boshuizen, H. P., & Patel, V. L. (1987). Contextual factors in the activation of first diagnostic hypotheses: Expert-novice differences. *Medical Education*, *21*(6), 471–476. https://doi.org/10.1111/j.1365-2923.1987.tb01405.x

Hoffman, K. A., Aitken, L. M., & Duffield, C. (2009). A comparison of novice and expert nurses' cue collection during clinical decision-making: Verbal protocol analysis. *International Journal of Nursing Studies*, *46*(10), 1335–1344. https://doi.org/10.1016/j.ijnurstu.2009.04.001

Hunink, M. G. M. (2001). In Search of Tools to Aid Logical Thinking and Communicating about Medical Decision Making. *Medical Decision Making*, *21*(4), 267–277. https://doi.org/10.1177/0272989X0102100402

Johnson, P. E., Moen, J. B., & Thompson, W. B. (1988). Garden Path Errors in Diagnostic Reasoning. In L. Bolc & M. J. Coombs (Eds.), *Expert System Applications* (pp. 395–427). Springer. https://doi.org/10.1007/978-3-642-83314-4_7

Jonas, E., Schulz-Hardt, S., Frey, D., & Thelen, N. (2001). Confirmation bias in sequential information search after preliminary decisions: An expansion of dissonance theoretical research on selective exposure to information. *Journal of Personality and Social Psychology*, *80*(4), 557–571. https://doi.org/10.1037/0022-3514.80.4.557

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University Press.

Karlsten, R., & Elowsson, P. (2004). Who calls for the ambulance: Implications for decision support. A descriptive study from a Swedish dispatch centre: *European Journal of Emergency Medicine*, *11*(3), 125–129. https://doi.org/10.1097/01.mej.0000114640.63700.68

Klayman, J. (1995). *Decision Making from a Cognitive Perspective: Advances in Research and Theory*. Academic Press.

Klein, G. (1997). Developing Expertise in Decision Making. *Thinking & Reasoning*, *3*(4), 337–352. https://doi.org/10.1080/135467897394329

Klein, G. (2008). Naturalistic decision making. *Human Factors: : The Journal of the Human Factors and Ergonomics Society*, *50*, 456–460.

Klein, G. (1993). A recognition-primed decision (RPD) model of rapid decision making. In *Decision making in action: Models and methods* (pp. 138–147). Ablex Publishing.

Kostopoulou, O., Rosen, A., Round, T., Wright, E., Douiri, A., & Delaney, B. (2015). Early diagnostic suggestions improve accuracy of GPs: A randomised controlled trial using computer-simulated patients. *British Journal of General Practice*, *65*(630), e49–e54. https://doi.org/10.3399/bjgp15X683161

Krems, J. F., & Zierer, C. (1994). [Are experts immune to cognitive bias? Dependence of 'confirmation bias' on specialist knowledge]. *Zeitschrift Fur Experimentelle Und Angewandte Psychologie*, *41*(1), 98–115.

Krupat, E., Wormwood, J., Schwartzstein, R. M., & Richards, J. B. (2017). Avoiding premature closure and reaching diagnostic accuracy: Some key predictive factors. *Medical Education*, *51*(11), 1127–1137. https://doi.org/10.1111/medu.13382

Kushniruk, A. W., Patel, V. L., & Marley, A. A. J. (1998). Small worlds and medical expertise: Implications for medical cognition and knowledge engineering. *International Journal of Medical Informatics*, *49*(3), 255–271. https://doi.org/10.1016/S1386-5056(98)00044-6

Larkin, J., McDermott, J., Simon, D. P., & Simon, H. A. (1980). Expert and novice performance in solving physics problems. *Science*, *208*(4450), 1335–1342.

Leprohon, J., & Patel, V. L. (1995). Decision-making Strategies for Telephone Triage in Emergency Medical Services. *Medical Decision Making*, *15*(3), 240–253. https://doi.org/10.1177/0272989X9501500307

Lindström, V., Karlsten, R., Falk, A.-C., & Castrèn, M. (2011). Feasibility of a

computer-assisted feedback system between dispatch centre and ambulances. *European Journal of Emergency Medicine*, *18*(3), 143–147.

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Mehlhorn, K. (2011). *Memory Activation and the Availability of Explanations in Sequential Diagnostic Reasoning*.

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J. M., Maino, K., Kissling, W., & Hamann, J. (2011). Confirmation bias: Why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine*, *41*(12), 2651–2659. https://doi.org/10.1017/S0033291711000808

Meyer, A. N. D., Payne, V. L., Meeks, D. W., Rao, R., & Singh, H. (2013). Physicians' Diagnostic Accuracy, Confidence, and Resource Requests: A Vignette Study. *JAMA Internal Medicine*, *173*(21), 1952–1958. https://doi.org/10.1001/jamainternmed.2013.10081

Muhiyaddin, R., Abd-Alrazaq, A. A., Househ, M., Alam, T., & Shah, Z. (2020). The impact of clinical decision support systems (CDSS) on physicians: A scoping review. In J. Mantas, A. Hasman, & M. Househ, *The Importance of Health Informatics in Public Health during a Pandemic*. IOS Press.

NHG. (2021, oktober). *ABCDE-kaart—NHG*. https://www.nhg.org/. https://www.nhg.org/thema/spoedzorg/abcde-methode/abcde-kaart-pdf/

Nibbelink, C. W., & Reed, P. G. (2019). Deriving the Practice-Primed Decision Model from a naturalistic decision-making perspective for acute care nursing research. *Applied Nursing Research*, *46*, 20–23. https://doi.org/10.1016/j.apnr.2019.01.003

Orasanu, J., & Connolly, T. (1993). The reinvention of decision making. In *Decision making in action: Models and methods* (pp. 3–20). Ablex Publishing. https://cir.nii.ac.jp/crid/1573105975434879744

Parmley, M. (2006). The effects of the confirmation bias on diagnostic decision making. *Drexel University*, 167.

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Bertrand, C., Charlin, B., & Dory, V. (2015). Insights into emergency physicians' minds in the seconds before and into a patient encounter. *Internal and Emergency Medicine*, *10*(7), 865–873. https://doi.org/10.1007/s11739-015-1283-8

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Bertrand, C., Dory, V., & Charlin, B. (2014). How and When Do Expert Emergency Physicians Generate and Evaluate Diagnostic Hypotheses? A Qualitative Study Using Head-Mounted Video Cued-Recall Interviews. *Annals of Emergency Medicine*, *64*(6), 575–585. https://doi.org/10.1016/j.annemergmed.2014.05.003

Pines, J. M., & Strong, A. (2019). Cognitive Biases in Emergency Physicians: A Pilot Study. *The Journal of Emergency Medicine*, *57*(2), 168–172. https://doi.org/10.1016/j.jemermed.2019.03.048

Prakash, S., Bihari, S., Need, P., Sprick, C., & Schuwirth, L. (2017). Immersive high fidelity simulation of critically ill patients to study cognitive errors: A pilot study. *BMC Medical Education*, *17*(1), 36. https://doi.org/10.1186/s12909-017-0871-x

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics*, *SMC-13*(3), 257–266. https://doi.org/10.1109/TSMC.1983.6313160

Richie, M., & Josephson, S. A. (2018). Quantifying Heuristic Bias: Anchoring, Availability, and Representativeness. *Teaching and Learning in Medicine*, *30*(1), 67–75. https://doi.org/10.1080/10401334.2017.1332631

Ritter, F. E., Baxter, G. D., & Churchill, F. E. (2014). *Foundations for designing user-centered systems: What system designers need to know about people*. Springer.

Ritter, F. E., Baxter, G. D., Kim, J. W., & Srinivasmurthy, S. (2013, February 12). *Learning and Retention*. The Oxford Handbook of Cognitive Engineering. https://doi.org/10.1093/oxfordhb/9780199757183.013.0008

Sibbald, M., Panisko, D., & Cavalcanti, R. B. (2011). Role of clinical context in residents' physical examination diagnostic accuracy. *Medical Education*, *45*(4), 415–421. https://doi.org/10.1111/j.1365-2923.2010.03896.x

Sibbald, M., Sherbino, J., Ilgen, J. S., Zwaan, L., Blissett, S., Monteiro, S., & Norman, G. (2019). Debiasing versus knowledge retrieval checklists to reduce diagnostic error in ECG interpretation. *Advances in Health Sciences Education*, *24*(3), 427–440. https://doi.org/10.1007/s10459-019-09875-8

Smith, M. W., Bentley, M. A., Fernandez, A. R., Gibson, G., Schweikhart, S. B., & Woods, D. D. (2013). Performance of Experienced Versus Less Experienced Paramedics in Managing Challenging Scenarios: A Cognitive Task Analysis Study. *Annals of Emergency Medicine*, *62*(4), 367–379. https://doi.org/10.1016/j.annemergmed.2013.04.026

Søreide, K. (2008). Three decades (1978–2008) of Advanced Trauma Life Support (ATLS™) practice revised and evidence revisited. *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine*, *16*, 19. https://doi.org/10.1186/1757-7241-16-19

StataCorp LLC. (2021). *Stata Statistical Software* (Release 17) [Computer software].

The R Foundation for Statistical Computing. (2021). *R version 4.1.2* (4.1.2 'Bird Hippie') [R].

Thiele, J. E., Holloway, J., Murphy, D., Pendarvis, J., & Stucky, M. (1991). Perceived and Actual Decision Making by Novice Baccalaureate Students. *Western Journal of Nursing Research*, *13*(5), 616–626. https://doi.org/10.1177/019394599101300504

Trowbridge, R. L., Rencic, J. J., & Durning, S. J. (Eds.). (2015). *Teaching clinical reasoning*. American College of Physicians. https://cir.nii.ac.jp/crid/1130282268765075584

Tsukada, H., Satou, T., Iwashima, A., & Souma, T. (2000). Diagnostic Accuracy of CT-Guided Automated Needle Biopsy of Lung Nodules. *American Journal of Roentgenology*, *175*(1), 239–243. https://doi.org/10.2214/ajr.175.1.1750239

Wan Ishak, W. H., Ku-Mahamud, K. R., & Md Norwawi, N. (2010). *Conceptual framework for intelligent decision support system in emergency management* [Conference].

https://repo.uum.edu.my/id/eprint/3468/

Ward, P., Torof, J., Whyte, J., Eccles, D. W., & Harris, K. R. (2010). Option Generation and Decision Making in Critical-Care Nursing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *54*(4), 354–358. https://doi.org/10.1177/154193121005400418

Wason, P. C. (1960). On the Failure to Eliminate Hypotheses in a Conceptual Task. *Quarterly Journal of Experimental Psychology*, *12*(3), 129–140. https://doi.org/10.1080/17470216008416717

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. *Theoretical Issues in Ergonomics Science*, *8*(3), 201–212. https://doi.org/10.1080/14639220500370105

Woods, D. D., Cook, R. I., Johannesen, L. J., & Sarter, N. B. (1994). *Behind Human Error: Cognitive Systems, Computers and Hindsight*. Crew Systems Ergonomics Information Analysis Center.

World Health Organization. (2019). *International statistical classification of diseases and related health problems (10th ed.)*. https://icd.who.int/browse10/2019/en

Yates, J. F., Veinott, E. S., & Patalano, A. L. (2003). Hard decisions, bad decisions: On decision quality and decision aiding. In *Emerging perspectives on judgment and decision research* (pp. 13–63). Cambridge University Press.

Zeiger, R. F. (2014). *McGraw-Hill's Diagnosaurus* (4.0) [Computer software]. AccessMedicine. http://accessmedicine.mhmedical.com/diagnosaurus.aspx

Zemaitis, M. R., Planas, J. H., & Waseem, M. (2023). Trauma Secondary Survey. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK441902/

Zwaan, L. (2012). *Diagnostic reasoning and diagnostic error in medicine*. https://research.vu.nl/en/publications/diagnostic-reasoning-and-diagnostic-error-in-medicine

Zwaan, L., & Singh, H. (2020). Diagnostic error in hospitals: Finding forests not just the big trees. *BMJ Quality & Safety*. https://doi.org/10.1136/bmjqs-2020-011099

Appendix

Appendix A: Tables

Values that are statistically significant are coloured green.

Table A.1: ***Diagnosis options for each diagnosis category***

Pulmonary	Cardiovascular	Psychiatry	Gastroenterology	Urology	Neurology	Miscellaneous
Acute upper respiratory infections	Hypertension	Dementia	Malignant neoplasm of colon	Tubulo-interstitial nephritis	Hypo/hyperglycemia	Withdrawal state alcohol
Pneumonia	Angina pectoris	Depressive episode	Gastro-esophageal reflux	Calculus of kidney	Encephalitis	Poisoning
Bronchitis	Acute myocardial infarction	Panic disorder	Acute appendicitis	Urinary tract infection	Focal symptomatic epilepsy	Food poisoning
Status asthmaticus	Acute ischemic heart disease	Acute stress reaction	gastroenteritis and colitis	Torsion of testis	Epilepsy	Poisoning by narcotics
Pneumothorax	Pulmonary embolism	Stress	Acute vascular disorders of intestine	Retention of urine	Migraine	Alcohol intoxication
Acute respiratory failure	Acute pericarditis	Other	Ileus	Other	Cluster headaches	Malnutrition
Other	Pericardial effusion		Other functional intestinal disorders		Tension-type headache	General malaise

	Cardiac arrhythmia		Constipation		Transient cerebral ischemic attacks (TIA)	Electrolyte disturbances
	Heart failure		Cholangitis		Cerebral infarction (CVA)	No diagnosis
	Cardiomegaly		Acute pancreatitis		Heat-/sunstroke	Other
	Aortic dissection		Other		Transient neuropathy	
	Abdominal aortic aneurysm				Other	
	Other					

Table A.2a: ***Actions in the categories of "Ask questions"***

Medical history	**Complaints**	**Context**
Medical history patient	Complaints	General
Medical history family of patient	Start/course of complaints	Medication
Previous experience with symptoms	Additional complaints	Smoking
	Pain rating	Alcohol
	Abdominal pains	Drugs
	Chest pain	Allergies
	Shortness of breath	

	Headache	
	Appetite/bowel movement	
	Thirst/micturition	

Table A.2b: ***Actions in the categories of "Perform examination"***

Look	**Listen**	**Feel**	**Measure**	**Head-to-Toe**
Airway obstruction	Hoarseness or audible breathing	Pulse in wrist/neck/groin	Oxygen saturation	Head and face
Skin tone	Auscultation lungs	Neck stiffness	Blood pressure	Neck and spine
Breathing rate	Percussion lungs	Swellings	Heart rate	Thorax
Breathing movement	Clear speech	temperature extremities	Capillary refill	Abdomen
Bleeding	Coherent speech		ECG 12 leads	Pelvis
Pupils	Auscultation heart		ECG 3 leads	Extremities
Lateralization			Blood sugar	Back
Paralysis/weakness			Temperature	
Tongue bite				
Urinary incontinence				
Consciousness score				
Skin abnormalities				

Table A.2c: ***Actions in the categories of "Administer treatment"***

Medication A-E	**Medication F-N**	**Medication O-Z**	**Other interactions**
Acetylsalicylic acid	Fentanyl	Ondansetron	Positioning
Adenosine	Furosemide	Paracetamol	Reassurance
Adrenaline	Glucagon	Ringer's lactate solution	Explain situation
Amiodarone	Glucose	Salbutamol	
Atropine sulfate	Heparin	Ticagrelor	
Budesonide	Hydrocortisone	Tranexamic acid	
Clemastine	Midazolam	Xylometazoline	
Esketamine	Morphine	Oxygen	
	NaCl 0.9%		
	Naloxone		
	Nitrospray		

Table A.3: ***Suggested diagnoses and associated positive test actions***

Case	Neurology case 1		Cardiology case 2	
Condition	Congruent	Non-congruent	Congruent	Non-congruent
Suggested diagnoses	Transient cerebral ischemic attacks (TIA) Cerebral infarction (CVA)	Epilepsy	Acute myocardial infarction Acute ischemic heart disease Angina pectoris	All Pulmonary diagnoses except Bronchitis
Positive test actions	Lateralization	Tongue bite	ECG (3 and 12 leads)	Breathing rate
	Paralysis/weakness	Urinary incontinence	Auscultation heart	Breathing movement
	Clear speech	Head-to-Toe	Chest pain	Oxygen saturation
	Coherent speech		Heart rate	Shortness of breath
	Headache		Blood pressure	Airway obstruction
	Extremities		Pain rating	Percussion lungs
			Pulse in wrist/neck/groin	Hoarseness or audible breathing
			Capillary refill	

Note. Actions that are positive tests for both the congruent and non-congruent suggested diagnoses have been removed in this table and are considered to not be positive tests.

Table A.4: ***Multinomial model results of action selection in case 1***

Variable	Action category	Coefficient	Std. err.	z	P>|z|	[95% conf. interval]	
Congruent condition x work experience x case progression	Non-congruent	0.0024412	0.002335	1.05	0.296	-0.0021353	0.0070178
Non-congruent condition x work experience x case progression	Non-congruent	0.0001124	0.0022189	0.05	0.96	-0.0042366	0.0044615
Congruent condition x work experience	Non-congruent	-0.0954286	0.0695571	-1.37	0.17	-0.231758	0.0409007
Non-congruent condition x work experience	Non-congruent	0.0147702	0.0629408	0.23	0.814	-0.1085915	0.1381319
Congruent condition x case progression	Non-congruent	-0.0404953	0.0437244	-0.93	0.354	-0.1261935	0.0452029
Non-congruent condition x case progression	Non-congruent	-0.011848	0.0434793	-0.27	0.785	-0.0970659	0.07337
Case progression x work experience	Non-congruent	2.32E-06	0.00207	0	0.999	-0.0040548	0.0040595
Case progression	Non-congruent	0.048411	0.0411075	1.18	0.239	-0.0321582	0.1289802
Work experience	Non-congruent	0.0127119	0.0561061	0.23	0.821	-0.0972541	0.1226779
Congruent condition	Non-congruent	1.754086	1.189486	1.47	0.14	-0.5772635	4.085435
Non-congruent condition	Non-congruent	0.3946054	1.207269	0.33	0.744	-1.971599	2.76081
Intercept	Non-congruent	-4.012381	1.064985	-3.77	0	-6.099713	-1.925048
Congruent condition x work experience x case progression	Congruent	0.0018606	0.0019653	0.95	0.344	-0.0019914	0.0057125

Non-congruent condition x work experience x case progression	Congruent	0.0017381	0.0018576	0.94	0.349	-0.0019028	0.005379
Congruent condition x work experience	Congruent	-0.0726067	0.0475346	-1.53	0.127	-0.1657727	0.0205593
Non-congruent condition x work experience	Congruent	-0.0394777	0.0434945	-0.91	0.364	-0.1247252	0.0457699
Congruent condition x case progression	Congruent	-0.044173	0.0365068	-1.21	0.226	-0.1157251	0.027379
Non-congruent condition x case progression	Congruent	-0.0399672	0.0355358	-1.12	0.261	-0.1096161	0.0296818
Case progression x work experience	Congruent	-0.0015573	0.0016311	-0.95	0.34	-0.0047542	0.0016396
Case progression	Congruent	0.0321712	0.0323483	0.99	0.32	-0.0312303	0.0955726
Work experience	Congruent	0.0458816	0.0368634	1.24	0.213	-0.0263694	0.1181325
Congruent condition	Congruent	1.617085	0.8559845	1.89	0.059	-0.060614	3.294784
Non-congruent condition	Congruent	1.130945	0.8489483	1.33	0.183	-0.5329629	2.794853
Intercept	Congruent	-3.111454	0.7330461	-4.24	0	-4.548198	-1.67471

Table A.5: ***Multinomial model results of action selection in case 2***

Variable	**Action category**	**Coefficient**	**Std. err.**	**z**	**P>|z|**	**[95% conf. interval]**	
Congruent condition x work experience x case progression	Non-congruent	-0.0023101	0.002508	-0.92	0.357	-0.0072257	0.0026055
Non-congruent condition x work experience x	Non-congruent	-0.0007204	0.0017357	-0.42	0.678	-0.0041224	0.0026815

case progression							
Congruent condition x work experience	Non-congruent	0.0210882	0.0356702	0.59	0.554	-0.0488241	0.0910004
Non-congruent condition x work experience	Non-congruent	0.0111989	0.033659	0.33	0.739	-0.0547714	0.0771693
Congruent condition x case progression	Non-congruent	-0.0064628	0.0319573	-0.2	0.84	-0.0690979	0.0561723
Non-congruent condition x case progression	Non-congruent	-0.0016481	0.0354058	-0.05	0.963	-0.0710423	0.067746
Case progression x work experience	Non-congruent	0.0008909	0.0009905	0.9	0.368	-0.0010505	0.0028323
Case progression	Non-congruent	-0.0390072	0.0152328	-2.56	0.01	-0.0688629	-0.0091516
Work experience	Non-congruent	-0.0122634	0.0201549	-0.61	0.543	-0.0517664	0.0272396
Congruent condition	Non-congruent	0.1206105	0.5487427	0.22	0.826	-0.9549055	1.196127
Non-congruent condition	Non-congruent	0.0319393	0.6609061	0.05	0.961	-1.263413	1.327291
Intercept	Non-congruent	-0.9908031	0.3324028	-2.98	0.003	-1.642301	-0.3393055
Congruent condition x work experience x case progression	Congruent	-0.0023034	0.0018101	-1.27	0.203	-0.0058512	0.0012444
Non-congruent condition x work experience x case progression	Congruent	-0.0012502	0.0014098	-0.89	0.375	-0.0040135	0.001513
Congruent condition x work experience	Congruent	0.0220372	0.0307004	0.72	0.473	-0.0381345	0.082209
Non-congruent condition x work	Congruent	0.0224612	0.0322993	0.7	0.487	-0.0408443	0.0857668

experience							
Congruent condition x case progression	Congruent	-0.0070616	0.020449	-0.35	0.73	-0.047140 9	0.033017 7
Non-congruent condition x case progression	Congruent	0.0171751	0.0284357	0.6	0.546	-0.038557 9	0.072908 1
Case progression x work experience	Congruent	0.000675	0.0007238	0.93	0.351	-0.000743 6	0.002093 7
Case progression	Congruent	-0.0076557	0.0100479	-0.76	0.446	-0.027349 2	0.012037 8
Work experience	Congruent	-0.0184955	0.0181301	-1.02	0.308	-0.054029 8	0.017038 8
Congruent condition	Congruent	0.4686653	0.4438618	1.06	0.291	-0.401288	1.338618
Non-congruent condition	Congruent	-0.3528992	0.6300092	-0.56	0.575	-1.587695	0.881896 2
Intercept	Congruent	-1.066744	0.2829552	-3.77	0	-1.621326	-0.51216 23

Table A.6: ***Multinomial model results of diagnosis selection in case 1***

Variable	**Diagnosis category**	**Coefficient**	**Std. err.**	**z**	**P>|z|**	**[95% conf. interval]**	
Congruent condition x work experience x case progression	Non-congruent	0.0067081	0.0105976	0.63	0.527	-0.014062 7	0.027479
Non-congruent condition x work experience x case progression	Non-congruent	0.0186901	0.0109337	1.71	0.087	-0.002739 5	0.0401197
Congruent condition x work experience	Non-congruent	-0.119256	0.0849483	-1.4	0.16	-0.285751 7	0.047239 6
Non-congruent condition x work experience	Non-congruent	-0.0188223	0.065233	-0.29	0.773	-0.146676 5	0.109032
Congruent	Non-congruent	-0.0009363	0.1731715	-0.01	0.996	-0.340346	0.338473

condition x case progression						3	6
Non-congruent condition x case progression	Non-congruent	-0.6758348	0.2182121	-3.1	0.002	-1.103523	-0.2481469
Case progression x work experience	Non-congruent	-0.0146566	0.0069267	-2.12	0.034	-0.0282327	-0.0010805
Case progression	Non-congruent	0.1764284	0.1205582	1.46	0.143	-0.0598613	0.4127181
Work experience	Non-congruent	0.0684033	0.0481166	1.42	0.155	-0.0259036	0.1627102
Congruent condition	Non-congruent	0.8518311	1.384894	0.62	0.538	-1.862512	3.566174
Non-congruent condition	Non-congruent	1.667406	1.155382	1.44	0.149	-0.5971016	3.931914
Intercept	Non-congruent	-1.166625	0.7925108	-1.47	0.141	-2.719918	0.3866671
Congruent condition x work experience x case progression	Congruent	0.0386635	0.0123261	3.14	0.002	0.0145047	0.0628223
Non-congruent condition x work experience x case progression	Congruent	0.0407176	0.0117629	3.46	0.001	0.0176626	0.0637725
Congruent condition x work experience	Congruent	-0.2538359	0.071472	-3.55	0	-0.3939185	-0.1137533
Non-congruent condition x work experience	Congruent	-0.1586101	0.064712	-2.45	0.014	-0.2854432	-0.0317769
Congruent condition x case progression	Congruent	-0.1238957	0.1564564	-0.79	0.428	-0.4305446	0.1827531
Non-congruent condition x case progression	Congruent	-0.263429	0.1378919	-1.91	0.056	-0.5336921	0.0068342
Case progression x	Congruent	-0.0404035	0.0109385	-3.69	0	-0.0618426	-0.0189643

work experience							
Case progression	Congruent	0.1901448	0.120944	1.57	0.116	-0.0469011	0.4271906
Work experience	Congruent	0.1538931	0.0524359	2.93	0.003	0.0511207	0.2566656
Congruent condition	Congruent	3.226684	1.139954	2.83	0.005	0.9924149	5.460952
Non-congruent condition	Congruent	1.0157	0.9722216	1.04	0.296	-0.8898188	2.92122
Intercept	Congruent	-1.035474	0.7531281	-1.37	0.169	-2.511578	0.4406296

Table A.7: ***Multinomial model results of diagnosis selection in case 2***

Variable	**Diagnosis category**	**Coefficient**	**Std. err.**	**z**	**P>|z|**	**[95% conf. interval]**	
Congruent condition x work experience x case progression	Non-congruent	0.1558374	28.23954	0.01	0.996	-55.19265	55.50432
Non-congruent condition x work experience x case progression	Non-congruent	0.2128181	0.2061355	1.03	0.302	-0.1912001	0.6168362
Congruent condition x work experience	Non-congruent	-0.0819004	136.2473	0	1	-267.1217	266.9579
Non-congruent condition x work experience	Non-congruent	-0.3237215	0.2410279	-1.34	0.179	-0.7961275	0.1486845
Congruent condition x case progression	Non-congruent	1.508273	346.1999	0	0.997	-677.0311	680.0477
Non-congruent condition x case progression	Non-congruent	0.4997129	1.003756	0.5	0.619	-1.467613	2.467039
Case progression x work experience	Non-congruent	-0.1624402	0.2027196	-0.8	0.423	-0.5597633	0.2348829
Case	Non-congruent	-1.119006	0.6624282	-1.69	0.091	-2.417342	0.179329

progression							2
Work experience	Non-congruent	0.1233683	0.2121597	0.58	0.561	-0.2924571	0.5391936
Congruent condition	Non-congruent	-19.47051	1929.793	-0.01	0.992	-3801.795	3762.854
Non-congruent condition	Non-congruent	3.717534	2.965928	1.25	0.21	-2.095579	9.530647
Intercept	Non-congruent	1.481591	1.164761	1.27	0.203	-0.8012992	3.764481
Congruent condition x work experience x case progression	Congruent	-0.0299674	0.0180294	-1.66	0.096	-0.0653044	0.0053696
Non-congruent condition x work experience x case progression	Congruent	0.0276807	0.0402855	0.69	0.492	-0.0512774	0.1066388
Congruent condition x work experience	Congruent	0.0670979	0.0676873	0.99	0.322	-0.0655668	0.1997626
Non-congruent condition x work experience	Congruent	-0.2034007	0.1369779	-1.48	0.138	-0.4718724	0.0650711
Congruent condition x case progression	Congruent	0.0470753	0.2123231	0.22	0.825	-0.3690704	0.463221
Non-congruent condition x case progression	Congruent	-0.1520939	0.7546496	-0.2	0.84	-1.63118	1.326992
Case progression x work experience	Congruent	0.0227669	0.0123129	1.85	0.064	-0.001366	0.0468998
Case progression	Congruent	0.4070997	0.1164237	3.5	0	0.1789135	0.635286
Work experience	Congruent	-0.0208907	0.0478697	-0.44	0.663	-0.1147136	0.0729321
Congruent condition	Congruent	1.10996	0.9371225	1.18	0.236	-0.7267661	2.946686
Non-congruent condition	Congruent	4.418631	2.957312	1.49	0.135	-1.377595	10.21486

Intercept	Congruent	-1.922096	0.6688906	-2.87	0.004	-3.233097	-0.6110941

Appendix B: Text

Section B.1: Explanation for software decisions.

The system that was used to facilitate the experiment needed comply with several requirements, which are as follows:

- *The system needed to be online*. To be able to perform the experiment, as described in the method section, the participants needed to be able to access the simulated scenario online.
- *The system needed to be in Dutch.* This experiment focuses on Dutch ambulance nurses, who work in a Dutch environment. For this reason, all text in the experiment needed to be in Dutch, so all Dutch ambulance nurses would be able to participate and there would be no interference from possible language barriers.
- *The system needed to work on old machines/browsers.* Many hospitals and ambulance centres still work with older computers and old software. The system needs to function on computers that have browsers that don't have the latest features.
- *The system needed to measure several outputs.* Not only did the system have to measure what diagnoses were chosen and when, it also needed to keep track of what actions were selected and in what order.
- *The system needed to be interactive.* When a participant selects an action this needs to have the appropriate effect on the patient information provided.
- *The system needed to be able to be used for future research/education.* A system that is text-based and image-based, makes it easier to update than using video or voice stimulus. Especially in an education setting this allows many cases to be added and updated in the system.

These requirements were too narrow to use a ready-made solution, so for this experiment a custom system was built to run the simulation in.