MEDICAL STUDENTS' BIOMEDICAL AND CLINICAL KNOWLEDGE: COMBINING LONGITUDINAL DESIGN, EYE TRACKING AND COMPARISON WITH RESIDENTS' PERFORMANCE

Ilona Södervik, Henna Vilppu, Erika Österholm, Mirjamaija Mikkilä-Erdmann

Abstract

This study combines longitudinal and individual process-level analyses to investigate medical students' biomedical knowledge and how they generate a diagnosis for a patient case text. The diagnostic processes were investigated using the eye-tracking method, and students' processes were compared with those of residents. The results showed that students differed in their diagnostic performance in the beginning of the clinical phase. Of the students who had biomedical misconceptions in the preclinical phase, 69% ended up with an incorrect diagnosis, while 60% of students with accurate biomedical knowledge made a correct diagnosis. The processing of a patient case text was faster among better-achieving students and residents. Furthermore, residents' illness-script activation could be seen from their eye-movement data as a relatively longer reading time regarding the sentence that concerned the enabling condition of the case. Based on the results of the study, pedagogical suggestions are discussed.

Keywords: Biomedical knowledge; clinical knowledge; illness script; eye tracking; higher education

Highlights

- Differences in biomedical and clinical knowledge among medical students are evident
- 69% of those with biomedical misconceptions made an inaccurate diagnosis
- Those who diagnosed the patient case correctly processed the text faster
- Students and residents made different use of enabling conditions for the case
- The eye-tracking method may enable the detection of the illness-script activation

1. Introduction

In recent decades, several studies have aimed to understand how medical students acquire a high level of competence on their way to achieving medical expertise (see e.g. Kuipers & Kassirer, 1984; Schmidt & Boshuizen, 1993; Schmidt & Rikers, 2007). There is a general consensus that basic science or biomedical knowledge provides a foundation for clinical knowledge (Kaufman, Keselman, & Patel, 2008; Woods, 2007), and recent studies have shown that students with better biomedical knowledge succeed better in clinical reasoning tasks (see e.g. Ahopelto, Mikkilä-Erdmann, Olkinuora & Kääpä, 2011; Nivala, Lehtinen, Helle′, Kronqvist, Paranko′ & Säljö, 2013). However, less is known about how the level of biomedical knowledge relates to diagnostic accuracy in the early clinical phase of medical studies. In this study, the development of medical students' biomedical knowledge concerning one of the most essential but complex systems — the central cardiovascular system (CCVS) — is followed up twice during their preclinical phase and compared to students' success in a clinical reasoning task at the beginning of the clinical phase.

Previous studies have provided interesting insights into novices' and experts' diagnosis of written patient cases (see e.g. Boshuizen, van de Wiel & Schmidt, 2012), but with a few exceptions (e.g. Vilppu, Mikkilä-Erdmann, Södervik & Österholm-Matikainen, 2017), such research has not focussed on the processes by which participants use the case description text while coming to a diagnosis. In this study, students' processing of a patient case text concerning a pulmonary embolus — in which understanding of the content is expected to suggest biomedical knowledge of the CCVS —is studied via eye movements, interviews and written tasks to study whether there are differences among students in their processes. Eye tracking offers a suitable method to study diagnostic processing, as eye movements indicate cognitive processing during the task (to read more about the widely accepted eye-mind hypothesis, see Just & Carpenter, 1980). To evaluate students' diagnostic processes, their processing was compared with that of medical residents, who had completed their six-year basic medical degree studies and already begun their specialisation training.

Different theories have been developed concerning how medical students construct a coherent entity of biomedical and clinical knowledge during their studies, thereby gradually learning to solve complex clinical problems (see e.g. Boshuizen & Schmidt, 1992; de Bruin, Schmidt & Rikers, 2005; Feltovich & Barrows, 1984; Kuipers & Kassirer, 1984; Patel, Evans & Groen, 1989; Schmidt & Boshuizen, 1993). Patel et al. (1989) argued that biomedical knowledge and clinical knowledge can be considered worlds apart, meaning that clinical knowledge and biomedical concepts construct separate knowledge structures; hence, biomedical knowledge is involved in clinical reasoning only in the sense that it provides coherence when solving exceptionally complex patient cases. However, at present, there is more empirical evidence supporting another perspective, where biomedical and clinical knowledge need to be integrated (i.e. encapsulated) for medical expertise to be achieved (Boshuizen et al., 2012). According to encapsulation theory, successful medical education and later expertise development cause biomedical knowledge to become constructed into knowledge structures that comprise concepts under which many lower-level details and interrelations of information are organised (see Schmidt & Rikers 2007; Schmidt & Boshuizen, 1993; Van de Wiel, Boshuizen & Schmidt, 2000).

In the preclinical phase, which takes approximately the first 2.5 years of medical studies, students' main aim is to build an extensive understanding of basic scientific knowledge, such as the anatomy and physiology of the human body. However, previous research has shown that the learning of biomedical contents poses challenges for medical students and often suggests the abandoning of certain misconceptions (see e.g. Chi, 2005; Mikkilä-Erdmann, 2002). In addition, from the very beginning of medical studies, students begin to acquire some practical experience, first via observing physicians' work in healthcare centres and later via real patient encounters. As students begin to use their biomedical knowledge in the clinical context, their biomedical knowledge starts to integrate with clinical experience and they begin to reorganise *illness scripts* (see e.g. Charlin, Boshuizen, Custers & Feltovich, 2007; Feltovich & Barrows, 1984; de Bruin et al., 2005; Schmidt & Rikers, 2007). An illness script (Feltovich & Barrows, 1984) is an integrated model of medical abnormalities, which specifies illness in terms of enabling conditions serving as background factors to influence the probability that an individual has contracted a disease (e.g. travelling to a malaria-endemic area) and possibly contribute to the fault, that is, the pathophysiological malfunctioning constituting the biomedical core of a disease (e.g. an enlarged spleen). This fault may give rise to certain consequences that are typically complaints, signs and symptoms (e.g. a high fever every other day) (Boshuizen et al., 2012; Custers, Boshuizen & Schmidt, 1998).

In real settings, physicians must address extremely complex and multifaceted patient cases that are influenced by the patients' background information, personal sensations, symptoms and findings. Therefore, the physician's challenge is to differentiate substance (e.g. the relevant symptoms) from competing noise (e.g. irrelevant symptoms) in each case. Illness scripts are thought to help physicians to find patterns of diseases, filter out irrelevant information, rule out several diseases and construct working diagnoses (Monajemi, Schmidt & Rikers, 2012). Previous studies have shown that because medical students are often unable to recognise these patterns, they may fail to make the correct diagnosis (Monajemi et al., 2012). Further, a characteristic of more experienced physicians is that they seem to make better use of the enabling conditions of a specific case compared to novices (Schmidt & Rikers, 2007). This means that an experienced doctor recognises predisposing factors better than novices; thus, for medical experts, enabling factors could be efficient promoters of illness-script activation.

1.1 Research questions

The research questions in this study are as follows:

1. *Are there differences among the participants in their diagnostic accuracy and do they differ at the stage of reading in which they find correct diagnosis of a patient case text?* It is suggested that residents with more clinical experience will be more accurate and efficient in making a diagnosis than students (see e.g. Charlin, et al., 2007). However, as diagnostic reasoning is a complex skill, it is also suggested that there would be variation among medical students in their diagnostic accuracy and efficacy.

1.1. *Do the processing times for a patient case text differ between participants who give a correct versus an incorrect diagnosis?* Following encapsulation theory and the results of previous eye-tracking studies, it is expected that those who diagnose the case correctly will process the case more quickly (e.g. Charlin et al., 2007; Mann, Williams, Ward & Janelle, 2007; Schmidt & Rikers, 2007).

1.2. *Do the participants giving a correct versus an incorrect diagnosis differ in their use and processing of different-level sentences of the patient case text?* Based on previous research, it is hypothesised that those who diagnose the case correctly will be more effective in directing their attention to the task-relevant areas of the text and hence, expected to mention more relevant issues in the written answers compared to the less successful participants. Moreover, based on illness-script theory, it is suggested that residents may make more use of sentences that relate to the enabling conditions of the patient case text than medical students (see Schmidt & Rikers, 2007).

2. *Are the level of biomedical knowledge and the occurrence of misconceptions about the CCVS, and medical entrance examination results, related to success in the clinical reasoning task among third-year medical students?* Previous studies have shown that the level of biomedical knowledge is related to the level of clinical reasoning (e.g. Nivala et al., 2013). Since it has been suggested that understanding the pathophysiology of pulmonary embolus requires accurate biomedical knowledge related to the CCVS, we suggest that those students who diagnose the case correctly may have fewer misconceptions and higher scores for their biomedical knowledge of the CCVS.

2. Method

2.1 Participants and design

The participants were native (*language removed)*-speaking students from one (*country removed*) medical school (see Table 1). The students were followed up concerning their biomedical knowledge about the CCVS three times during their initial study years in 2009–2010. The whole class of medical students participated in group study phases 1–2 (more detailed description and results are provided in Ahopelto et al., 2011). However, only those who were involved in the single-case study phase are focussed on in the present study.

Table 1 The study design

Follow-ups of the biomedical knowledge 2009-2010: group-level studies in lecture hall context		***Clinical knowledge 2011-2012: single-case study in eye-tracking laboratory***	
Study phase 1	*Study phase 2*	*Study phase 3*	
in semester 2; 2009	in semester 4; 2010	in semester 6; 2011	2012
N =31	*N* = 31	*N* = 33	*N* = 13
1st year students	2nd year students	3rd year students	internal residents
Measurement 1: Biomedical tasks	*Measurement 2: Biomedical tasks*	*Measurement 3: Patient case tasks*	

A total of 39 third-year students and 13 internal medicine residents from one (*nationality removed*) university hospital participated in the single-case study phase, which was conducted in spring 2011 for students and spring 2012 for residents. The eye-tracking data were poor for six students; hence, 33 students (24 women) and 13 internal medicine residents (eight women) were included in the study. The participating students represented 27% of the third-year medical student population, with the total size of the cohort being 122. The residents represented the total number of residents currently working in the internal medicine ward. The students participated in the study during their free time, whereas the residents participated during their working time at the clinic. Participation in all studies was voluntary, and informed consent was obtained. Approval for the study was obtained from the ethics review board of the (*name of university removed*).

2.2 Procedure

Study phases 1–2: In the first study year, the participants' biomedical knowledge concerning the CCVS was measured through written tasks in a lecture context before the students attended a blood circulation, respiration and fluid balance course, in which the content of the CCVS was studied. In the second year, their biomedical knowledge related to the CCVS was followed up with written tasks completed in a lecture context.

Study phase 3: The single-case eye-tracking study phase was administered in the eye-tracking laboratory, and it consisted of an orientation and a trial phase. The orientation was parallel to the actual trial, and its purpose was to familiarise the participants with the test situation. Eye movements were recorded during the reading. The session lasted for 1–2 h for the students and 0.5–1 h for the residents.

At the beginning of the eye-tracking study, the eye tracker was adjusted accordingly and calibrated, and instructions regarding the orientation and trial phases, which had identical structures, were given. The participants were told that they would be shown a short text concerning a patient case, and after each factual slide, an interspersed slide would ask them to answer paper-and-pencil questions without the text (students) or orally respond to the questions presented on a computer screen (residents). After the questions had been answered, participants were instructed to continue reading by pressing an arrow key on the keyboard. Participants were informed that they should read the text carefully, as it would not be possible to move back to the previous slide. There were no time limits, and the participants could move on at their own pace.

After the trial, a stimulated recall interview was conducted, in which the eye-tracking data were reviewed with the participants and questions were posed about parts of the text where issues of interest, such as longer fixations or rereading, emerged in the eye movement data.

2.3 Materials

The participants' medical entrance exam scores were used as a background variable. Medical entrance exams in (*nationality removed*) are highly selective (only about 15% of the applicants are accepted). The examination consists of mainly open-ended questions measuring biomedical knowledge in biology, chemistry and physics. Applicants' responses are scored by the medical staff of the faculties.

Measurement 1: First-year medical students were asked to complete the following written assignment: '(a) Draw the structures of the central cardiovascular system (the heart with the largest vessels). Name the structures. (b) Explain how the blood flows in the structures you drew'. There were also two 'select one from four' multiple-choice questions complementing the drawing task. The tasks were designed based on previous studies, in which typical conceptual problems in understanding the content of the CCVS have been identified (e.g. Chi, 2005; Michael et al., 2002; Ahopelto et al., 2011).

Measurement 2: In the second year of study, a similar type of written assignment was given: 'Explain the path of a red blood cell from the left ventricle back to the same place'. After accomplishing this task, the participants completed a figure task concerning the pulmonary circuit, where they were given a diagram of the pulmonary circuit and instructed to add arrows that symbolised the number of vessels (e.g. from the right ventricle to the lungs via two pulmonary arteries [one per lung], and from the lungs to the left atrium via four pulmonary veins [two per lung]). The students were also asked to mark the arrows with either a broken line to indicate oxygen-poor blood or an unbroken line to indicate oxygen-rich blood.

Measurement 3: In the third study phase, two written patient case texts were used for the orientation and trial phases of the study. The texts were read on a computer screen in the eye-tracking laboratory. The orientation and trial cases were structured identically; the topic of the first patient case was cardiac failure, while the second dealt with pulmonary embolus. The topics were chosen because they would both be familiar to the students; further, knowledge of the pathophysiology underlying these conditions suggests an understanding of the biomedical content related to the CCVS.

The idea of the patient case texts was to simulate the phases of a patient encounter in a healthcare centre. Therefore, the texts followed a patient case, including anamnesis (i.e. medical history of the patient), status and examination results from laboratory tests. The orientation patient case text was 199 words, while the trial text was 225 words in length. The texts were in (*language removed*), but an English translation of the trial patient case text is provided in the appendix.

Findings from the laboratory examinations and X-rays were presented in written form and with necessary interpretation. Hence, there were no images included in the patient case material, and it was not necessary to remember information like reference values. The patient case texts were written without using particular terminology to avoid disruption of the students' reading processes as a result of unfamiliar wording. Details concerning the case results were not required to be memorized, and the text also included some interpretation of the results, such as 'her blood pressure is 110/70 mmHg, thus normal'. The patient cases were written by the authors and evaluated and commented on by two cardiology specialists (these experts also act as medical teachers having an understanding about the knowledge level of the third-year students) to guarantee the validity of the text content. After designing the text, it was piloted with six fourth-year medical students, and the readability of the text was improved regarding the comments received from the students.

The text sentences were semantically categorised into different levels, as follows: key sentences (including relevant and essential information for the diagnosis), supplementary sentences (which included neutral information but still supported the reader in excluding incorrect options) and irrelevant sentences (containing irrelevant information). The cases were constructed such that the first slide addressed the anamnesis of the patient, signifying the patient's medical history and preliminary knowledge of the condition as given by the patient — in the framework of illness-script theory, this provided knowledge about the enabling conditions and consequences. The second slide gave information regarding the patient status, that is, the results of examining the patient in the doctor's office. Finally, the third slide provided medical examination results, such as a description of the laboratory results and X-ray findings. The second and third slides included information about faults and consequences according to illness-script theory. The relevant sentences in the orientation case included the most typical instances of the disease, whereas the trial case was intentionally designed to be more demanding; in the trial case, the relevant sentences were structured to fit to a typical but not the most prototypical manifestation of the disease (see e.g. Charlin et al., 2007).

Following each factual slide (anamnesis, status and examination results), written questions were presented to the students and oral questions were posed to the residents (the purpose of the difference was solely to save the residents time, as they were participating in the study during their busy working hours). After anamnesis, the item posed was: 'Name the most essential symptoms according to anamnesis and give a working diagnosis'. After status, the participants were asked to do the following: 'Name the most essential findings according to status and give a working diagnosis'. Finally, after the last slide, the following instruction was given: 'Make a diagnosis and name the most essential symptoms and findings that allowed you to make your diagnosis'. The purpose of these items was to measure participants' ability to identify relevant aspects from the slides and to determine which allowed the participant to make a correct diagnosis.

2.4 Setting and apparatus of the eye-tracking study phase

A Tobii T60XL Eye Tracker (Tobii Technology, Inc., Falls Church, VA) was used to record participants' eye movements during reading. Infrared cameras tracing the position of the participants' pupils were integrated into the body of the same high-resolution 24" computer monitor operating at 60 Hz, at a resolution of 1920x1200 pixels, from which the stimuli were presented. The accuracy of the eye tracker was 0.6°. Gaze remaining within a 30-pixel radius for 60 ms or more constituted a fixation. The Tobii T60XL Eye Tracker allows even large head movements, and as we wanted the reading process to be as realistic as possible, no supporting chin rests were used; thus, the research was carried out in relatively authentic conditions.

2.5 Data analysis

Study phases 1–2: The data from the biomedical questions from the first and second study years were scored, and misconceptions were quantified. Further, the number of students with misconceptions either in the first or second study year, or both, was calculated. An inter-rater reliability analysis was accomplished with 20% of the data, and the Cohen Kappa value of reliability was found to be .906. The data were statistically analysed with IBM SPSS Statistics 22 (IBM, Armonk, NY).

Study phase 3: First, the diagnoses made by the participants for the patient case in the trial were categorised as either correct or incorrect. Written answers concerning the most essential findings were digitalised and categorised in groups of relevant and irrelevant aspects by an external member of the research group using strict instructions. The oral answers were transcribed word for word, and the participants' reading processes were analysed using Tobii Studio Statistics. In addition, the numerical data were transferred from Tobii Studio to IBM SPSS Statistics 22, which was used for further analyses.

The analysis of eye-tracking data began with defining the areas of interest (AOIs), the regions in the stimulus from which the authors were interested in gathering data (Holmqvist, Nyström, Andersson, Dewhurst, Jarodzka & van de Weijer, 2011). Each slide, each key sentence and each irrelevant sentence constituted an AOI. It has been suggested that highly important sentences receive more visual attention than unimportant ones (Hyönä & Niemi, 1990), so the sentence-level analyses focused on key and irrelevant sentences, and supplementary sentences were excluded.

The metric of dwell time, which means the sum of all visits in an AOI, was chosen. An individual visit is defined as the time interval between the first fixation on the active AOI and the end of the last fixation on the same active AOI where there have been no fixations outside the AOI. The whole slide was considered the best metric of the processing time for a certain slide, including the reading time and possible thinking time, before moving to the question slide. Reading times for the sentences at different levels on each slide were calculated together; thus, in addition to sentence-by-sentence analysis, it was possible to compare the reading times for all key or irrelevant sentences per slide. When comparing the dwell times of sentences at different levels, the total dwell duration of a certain AOI was first divided by the number of characters, as the sentences had different lengths.

Stimulated recall interviews were transcribed word for word and analysed; moreover, the most common answers were classified and quantified. Certain representative answers are presented as examples of the qualitative analyses.

Descriptive statistics and non-parametric statistical tests were conducted in analysing the results, specifically Shapiro–Wilk normality tests, Mann–Whitney U, Kruskal–Wallis and Wilcoxon signed-ranks tests. Effect sizes together with 95% confidential intervals were calculated, and limiting values for the tests were reported according to Field (2009).

3. Results

3.1. Diagnostic accuracy and differences among the participants in the stage of reading at which they find the correct diagnosis?

A total of 15 of 33 students (45%) made a correct diagnosis of pulmonary embolus after reading the patient case. Looking only at participants who solved the case, 6/15 (33%) of these students made a correct working diagnosis after the first (anamnesis) slide, 12/15 (80%) did so after the second (status) slide and the rest came to the correct diagnosis after the last slide. In contrast, more than half of the students (n = 18; 55%) failed to diagnose the case correctly, meaning that after reading the whole patient case text, they gave an incorrect diagnosis or no diagnosis at all. In accordance with our expectations, all residents (n = 13; 100%) made a correct diagnosis after reading the patient case. The residents were efficient in establishing a diagnosis, considering that all of them already had a correct working diagnosis after reading the first (anamnesis) slide.

3.1.1. Did the processing times for a patient case text differ between participants who gave a correct versus an incorrect diagnosis?

When the processing times between the student groups were studied, the comparison of total dwell times of students who provided a correct versus an incorrect final diagnosis revealed that the processing time for anamnesis was almost the same in the two groups (z = –.615, p = .556, r = .11, CI_r [.00, .44]), and it differed moderately, but not statistically significantly (z = –.434, p = .682, r = .08, CI_r [.00, .41]) for the second (status) slide. However, students who made an incorrect diagnosis took significantly longer to read the last slide of the text (the laboratory examination results), with a medium effect size, than those students who gave a correct answer (z = –2.31, p = .020, r = .40, CI_r [.07, .65]; see Figure 1). The residents read each slide in a statistically significantly shorter time, with large effect sizes, than the students did (anamnesis: z = –4.24, p < .001, r = .63, CI_r [.42, 78]; status: z = –4.72, p < .001, r = .70, CI_r [.52, .82]; laboratory examination results: z = –5.01, p < .001, r = .74, CI_r [.57, .85]). Unlike the students, the residents' processing times per slide decreased during reading, such that the last slide took only approximately half of the time compared to the dwell time for anamnesis.

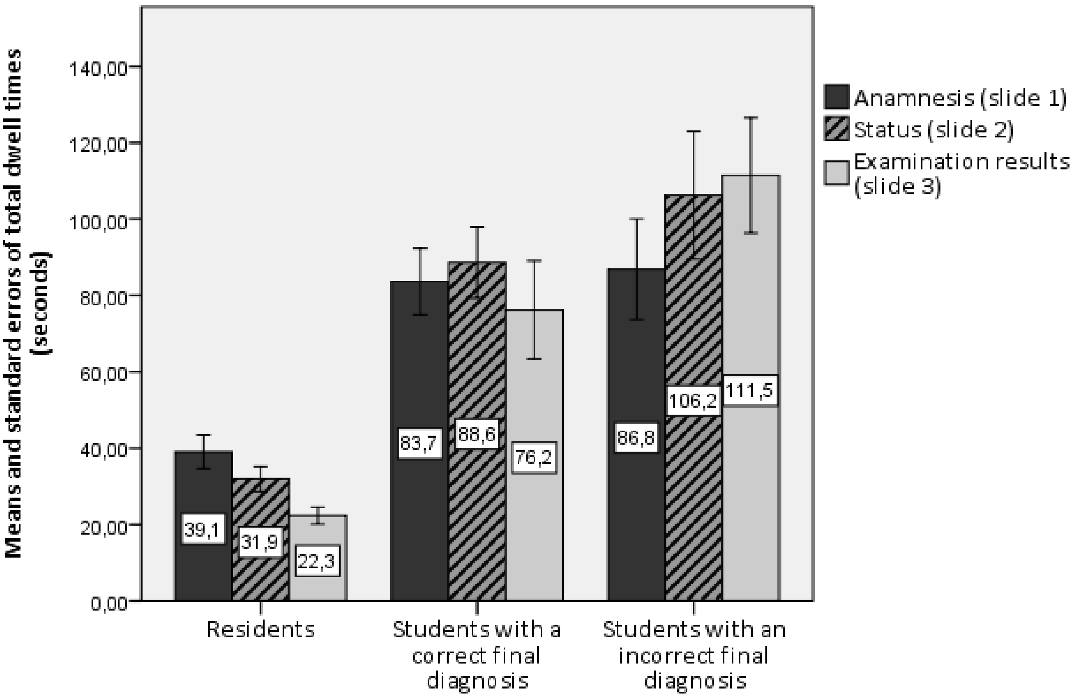

Figure 1 Means and standard errors of dwell times per slide of residents and students with a correct/incorrect final diagnosis

3.2.2. Did the participants giving a correct versus an incorrect diagnosis differ in using different-level sentences of the patient case text?

When comparing the relative dwell times of different-level sentences among the participant groups, the result showed that there was a significant difference, with a medium effect size, regarding solely the second slide (Slide 1: χ^2 (2) = 2.90, p = .234, r = .11, CI_r [.00, .39]; Slide 2: χ^2 (2) = 10.02, p = .007, r = .36, CI_r [.08, .59]; Slide 3: χ^2 (2) = 1.86, p = .395, r = .04, CI_r [.00, .33]). When comparing the relative dwell times of student groups with each other, it became evident that the groups did not differ statistically significantly in their reading times of sentences from different levels (anamnesis: z = - .289, p = .772, r = .05, CI_r [.00, .39]; status: z = - .615, p = .539, r = .11, CI_r [.00, .44]; laboratory results z = -

,362, p = .718, r = .06, CI_r [.00, .40]). When comparing the reading times between residents and student groups, the results showed that there was a significant difference, with a medium effect size, regarding the different-level sentences in the second slide (z = -3.11, p = .001, r = .46, CI_r [.20, .66]) , but not in the first (z = -1.69, p = .093, r = .25, CI_r [.00, .50]) or third slide (z = -1.28, p = .207, r = .19, CI_r [.00, .46]). When studying the relative dwell times of different-level sentences within the participant groups (Table 2), the results showed that there were no statistical differences between key and irrelevant sentences for the first slide, although all groups focused slightly more on key sentences. Surprisingly, all groups read irrelevant sentences in the second slide for a significantly longer time than they did the key sentences. In the third slide, all groups read the key sentences for a significantly longer time than the irrelevant sentences.

Table 2. Comparison of relative total dwell times between key sentences and irrelevant sentences

		% of total dwell time in key sentences	**SD**	***Z***	***p***	***r***	***CI_r***
Residents	Slide1	48,7	8,0	-.628	.530	.09	.00, .37
	Slide2	32,7[1]	6,7	-3.059	.002[2]	.46	.20, .66
	Slide3	59,6	9,0	-2.746	.006[2]	.41	.14, .63
Students with correct final dg	Slide1	52,4	5,2	-1.761	.078	.26	.00, .51
	Slide2	41,1[1]	9,1	-2.953	.003[2]	.44	.17, .65
	Slide3	64,0	7,3	-3.408	.001[2]	.51	.26, .70
Students with incorrect final dg	Slide1	52,6	5,8	-1.502	.133	.22	.00, .48
	Slide2	41,4[1]	7,1	-3.462	.001[2]	.52	.27, 70
	Slide3	62,6	8,4	-3.724	<.000[2]	.56	.32, .77

[1] significant difference between participant groups [2] significant difference within a participant group

A sentence-by-sentence comparison of dwell times between groups of residents, students who made a correct diagnosis and students who made an incorrect diagnosis showed statistically significant or highly significant differences between the three groups in all but one sentence (χ^2 (2) = 5.80, p = .055, r = .26, CI_r [.00, .51]). This sentence, 'The patient is recuperating from knee surgery' was the first key sentence of the patient case, and it appears that it was processed relatively longer by residents than the student groups.

Stimulated recall interviews given by the residents revealed that 6/13 (46%) of them made comments, such as ‘After reading this sentence [The patient is recuperating from knee surgery] I knew, er, I began to suspect that this patient has a thrombus or embolus’, and ‘After reading this [The patient is recuperating from knee surgery], this was a clear case for me’. Further, the other 7/13 (54%) residents referred to the surgery when giving a correct working diagnosis after the first slide. Regarding the irrelevant sentences, residents made comments such as the following: ‘Patients have a lot of findings, which in the end turn out to be confusing, for they are not related to the actual illness. You often have to try to see the forest from the trees’.

Considering that the diagnostic process took longer for most students, it was not a surprise that only a few students made comments about the first key sentence after reading the anamnesis. However, most of the students — 12 of the 15 (80%) who ended up with a correct diagnosis — described a ‘web of risks’ after reading the whole text, and 14/15 (93%) of them mentioned lower limb surgery as an exposing aspect for pulmonary embolus, but not until after reading the last slide. The following example was a typical answer from those who made a correct diagnosis: ‘Well, I think it (reasoning) started from the fact that there was this sort of web of risks, meaning diabetes, surgery, smoking […] and other kinds of signs of thrombus. And further, dyspnoea and palpitation could also relate to that (embolus). In the status, there were also things that spoke of it (embolus), there was tenderness in the calf, increased respiratory frequency, tachycardia, etc.’.

Comparing the number of relevant and irrelevant comments in the written answers of the student groups after reading each slide, the results showed that there was no difference between the correct and incorrect final diagnosis groups in the answers given after the first (anamnesis) (relevant comments: $z = -1.22$, $p = .259$, $r = .21$, CI_r [.00, .52]; irrelevant comments: $z = -.14$, $p = .901$, $r = .02$, CI_r [.00, .36]) and second (status) slides (relevant comments: $z = -1.05$, $p = .343$, $r = .18$, CI_r [.00, .49]; irrelevant comments: $z = -1.84$, $p = .108$, $r = .32$, CI_r [.00, .60]). However, when comparing students’ answers to the question presented after reading the last (laboratory examination results) slide, ‘Name the most essential symptoms and findings on which you made the final diagnosis', we found that the students who made a correct diagnosis named statistically significantly more relevant aspects of the case, with a large effect size ($z = -3.20$, $p = .001$, $r = .56$, CI_r [.27, .78]). In proportion, those who made an incorrect diagnosis mentioned significantly more irrelevant aspects, with a medium effect size ($z = -2.31$, $p = .027$, $r = .40$, CI_r [.07, .65]).

Comparing the number of relevant and irrelevant comments in the written answers of the student groups and residents, the results showed that the residents reported fewer irrelevant aspects after anamnesis and status compared to the student groups (see Table 3), (based on anamnesis: χ^2 (2) = 13.60, $p = .001$, $r = .45$, CI_r [.18, .65]; based on status: χ^2 (2) = 6.96, $p = .031$, $r = .28$, CI_r [.00, .53]). Furthermore, after reading the medical examination results, the residents reported significantly more relevant (χ^2 (2) = 20.60, $p < .001$, $r = .59$, CI_r [.36, .75]) and significantly less irrelevant (χ^2 (2) = 14.50, $p = .001$, $r = .47$, CI_r [.21, .67]) aspects related to the case than the student groups.

Table 3 Number of relevant/irrelevant comments when asked to name the most essential symptoms and findings on which the participants made their working/final diagnosis

		Residents (n=13)		**Students with correct final diagnosis (n=15)**		**Students with incorrect final diagnosis (n=18)**	
		Relevant	Irrelevant	Relevant	Irrelevant	Relevant	Irrelevant
After SLIDE 1	Comments (n)	2,5	0,5*	2,6 [a]	1,5 [b]	2,1 [a]	1,6 [b]
After SLIDE 2	Comments (n)	2,2	0,4*	2,4 [c]	0,6 [d]	2,1 [c]	1,0 [d]
After SLIDE 3	Comments (n)	5,3*	0,3*	3,9 [e]	1,1 [g]	2,1 [f]	1,6 [h]

* Significant difference between residents and student groups
Between same letters no significant difference
Significant differences between e-f: $z = -2.31$, $p = .027$, r = .40, CI_r [.16, .62]; g-h: $z = -2.31$, $p = .027$, r = .40, CI_r [.16, .62]

3.3 Development of biomedical knowledge in relation to success in the patient case task

Medical students' biomedical knowledge about the CCVS increased during the preclinical studies from the first (M = 9.46/12.00; SD = 1.86) to the second year of study (M = 10.89/12.00; SD = 1.13). In the first study year, 12/31 (39 %) participants had one or several misconceptions, whereas in the second study year, 6/31 (19 %) had them. Thus, 16/31 (52 %) of the students had one or several biomedical misconceptions related to the CCVS in either the first or second study year or both.

The development of biomedical knowledge was compared between the students who made a correct diagnosis and students who made an incorrect diagnosis. It became evident that students who made a correct diagnosis of the patient case had received slightly, but not significantly, higher scores related to biomedical measures, in the entrance examination and in the first study year. They also had fewer (not significantly) misconceptions in both years than students who were not able to give a correct diagnosis in the third year (see Table 4). In addition, 11/16 (69 %) of those students, who had one or several biomedical misconceptions in the first and/or second year, were not able to diagnose the case correctly, whereas solely 5/16 (31 %) from those ended up with a correct diagnosis. On the contrary, from those who had no biomedical misconceptions in the preclinical phase, 9/15 (60 %) made a correct diagnosis.

Table 4 Entrance examination scores, scores of the biomedical knowledge and the amount of misconceptions related to CCVS classified by the validity of patient case diagnosis (CD = correct diagnosis, FD = false diagnosis)

	N		**Minimum**		**Maximum**		**Mean**		**Std. Deviation**		***Z***	***p***	***r***	***CI_r***
	CD	**FD**	**CD**	**FD**	**CD**	**FD**	**CD**	**FD**	**CD**	**FD**				
Entrance exam	14	16	77.00	76.00	97.00	99.00	87.36	83.69	5.97	6.36	-1.54	.12	.28	.00, .58
Biomed 1st year	14	17	7.20	3.60	12.00	12.00	9.73	9.25	1.80	1.94	-.48	.63	.08	.00, .42
Misconceptions 1st year	14	17	.00	.00	4.00	3.00	.57	.76	1.16	1.03	-.89	.37	.16	.00, .49
Biomed 2nd year	14	17	9.00	8.25	12.00	12.00	10.61	11.12	1.05	1.16	-1.54	.12	.28	.00, .58
Misconceptions 2nd year	14	17	.00	.00	2.99	2.00	.21	.29	.58	.59	-.58	.56	.10	.00, .44

4. Discussion and conclusions

The purpose of this study was to establish how the level of medical students' biomedical knowledge of the CCVS develops during the first three years of medical studies. Furthermore, we were interested in how the third-year students and residents would solve a patient case of pulmonary embolus, as well as how the processing times of a patient case text differed between the participants who gave a correct versus an incorrect final diagnosis. We also investigated how the participant groups differed in making use of different levels of sentences, such as relevant and irrelevant aspects of the patient case text. Finally, we compared the medical entrance examination results and the longitudinal data regarding the development of biomedical knowledge between the students who gave a correct versus an incorrect diagnosis after reading a patient case text.

In the third study year, less than half of the students made a correct diagnosis after reading the patient case text, but as expected, all residents already made a correct diagnosis after reading the anamnesis.

Considering that the early identification of relevant hypotheses is an important feature of expert behaviour in medicine (e.g. Charlin et al., 2007), this result was not a surprise. When investigating participants' reading processes, we found that residents read the case remarkably faster than the student groups did. Schmidt and Boshuizen (1993) have suggested that students' clinical processing is slower than that of more experienced physicians because students must activate their biomedical knowledge in a conscious fashion, a cognitive activity that takes considerable time because no ready-made structures are available. Further, the students who made a correct diagnosis read faster than students with an incorrect final diagnosis did, which is in line with previous studies in which overall reading time has been found to correlate positively with experienced text difficulty (Rayner, 1998). Almost all students who ended up with a correct diagnosis gave the correct working diagnosis after reading the second slide, which may partially explain why their reading sped up after the second slide.

Although the residents were generally faster readers than the students, sentence-by-sentence inspection revealed an interesting finding regarding the first key sentence of the case ('The patient is recuperating from knee surgery'), in which the relative processing times did not differ between the students and residents. This means that the residents read this sentence concerning enabling factors of the current diagnosis relatively longer than students did. We suggest that this first key sentence may have activated some residents', but not students', illness-script system; this finding was also supported by stimulated recall interviews and the result that all residents made a correct diagnosis after the first (anamnesis) slide. Stimulated recall interviews supported the assumption that the diagnostic processes of residents and students differed remarkably. For residents, the first key sentence acted as an effective illness-script activator either immediately or no later than after reading the first slide. In contrast, although almost all students who ended up with a correct diagnosis mentioned the knee operation as one important basis to justify their diagnosis, most mentioned it only after reading the whole patient case text. Previous studies have suggested that experts make increasing use of the knowledge of specific cases' enabling conditions (Schmidt & Rikers, 2007), meaning that enabling factors in a patient case are often efficient promoters of illness-script activation for experts. Our results show that this was also the case among residents, at least when it came to this case. Therefore, in patient encounters, the role of anamnesis — in which the enabling conditions are discussed — may be crucial in illness-script activation and the diagnostic process. This is a result that should be addressed in medical education.

Surprisingly, residents and better-succeeding students focused more on irrelevant than relevant sentences on the second (status) slide. Despite that reading behaviour, all residents stayed with their correct working diagnoses given after the first slide. This may derive from already graduated physicians' critical awareness, as physicians are educated to systematically test their hypotheses in a script-verification process in which the doctor attempts to determine whether the activated script or any of the activated scripts adequately fits the clinical findings until all available information is received (see e.g. Charlin, et al., 2007); staying with the first hypothesis would be considered risky and in the worst case could lead to malpractice. It is notable here that this result again highlights the need for a multimethod approach when investigating the processing of complex tasks.

Students' written answers revealed that in the last (laboratory examination results) slide, more successful students had a greater capacity to distinguish between relevant and irrelevant information than those students who did not make the correct diagnosis. This may again derive from the result that almost all students who made a correct diagnosis had already done so after reading the second slide. Thus, presumably having a strong working diagnosis in mind, these students may have been more capable of focussing on relevant aspects in the third slide. However, this result was not demonstrated in the students' reading behaviour according to the analysis of eye-movements of different-level sentences.

The results of the longitudinal study showed that the medical students' biomedical knowledge of the CCVS increased between the first and second years of their preclinical studies. The result was expected, as the preclinical phase consists mainly of studying biomedical contents, and the CCVS is one of the main topics. Finally, when comparing the biomedical knowledge scores of the students who made a correct versus an incorrect diagnosis in the patient case study phase, the scores were slightly, but not significantly, higher in the first study year among those students who made the correct diagnosis in the third year. In addition, those who had one or several biomedical misconceptions in the first and/or second year were more likely (not significantly however) to make an incorrect diagnosis in the third study year. Furthermore, it is concerning that the groups of students already differed (not significantly however) in their performance on the medical entrance examination scores, which may indicate that lower-achieving students do not catch up to stronger students in later years. According to Ericsson (2016), expertise develops in a cumulating manner, so that the gulf between stronger and weaker students may even broaden during their studies. This result is extremely concerning and requires further studies and attention in higher education.

4.1 Limitations of the study and future directions

This study has several strengths, particularly its longitudinal and multimethod approach, However, several limitations should be noted when considering generalising the results of this study. Combining a longitudinal design and the use of a time-consuming eye-tracking method in this study resulted in the sample size being rather small (student participants represented 27% of the cohort) and the sample may also be somewhat biased. First, the students who volunteered to participate in the eye-tracking study represented average or better-succeeding students, as based on our previous studies with this same cohort (Ahopelto et al., 2011), we know that the students who struggled most with their studies during the preclinical phase did not participate in the last study phase. Further, the confidence intervals of effect sizes were rather large related to several results, which needs to be taken into account when interpreting the results.

A repeating follow-up design influences participants' performance in several ways: first, the participants can learn to answer certain types of questions and second, each measurement influences the following ones. Furthermore, repeating measures might result in so called respondent fatigue, where the participants lose their interest towards the study, which again may affect their answers (e.g. Lavrakas, 2008). However, there was only one study phase during a school year, and the measures of first and second study years did not radically differ from the normal curriculum-based tests used in medical education. The measurement, which clearly differed from the normal content of medical studies, the eye-movement study, was accomplished in the third year and completed the longitudinal study. Lastly, the experimental study design of the eye-tracking study phase meant that the participants may not have read the patient case texts as they would normally do, since it was not possible to go back to the previous slide in the texts. This may have affected participants' reading patterns.

The written and oral answers were classified by an outside member of the research group according to the instructions provided. However, the protocol was not blind, as the same person also classified the diagnoses as correct or incorrect. Nevertheless, the analysis was accomplished according to a strict procedure based on the division of relevant and irrelevant aspects and correct versus incorrect diagnoses made in collaboration with medical experts. Furthermore, as residents answered the questions between the slides orally, while students did so in writing, the different implementation protocol may have affected, for example, the reasoning protocol and the length of the answers. It would be recommended to standardise the protocol in later studies.

The case topics were chosen based on their familiarity to the students from their previous studies; the study was not controlled for whether some of the students had, for example, practical experience on the topics from attending rounds in their clinical studies. Finally, this study focused on only one medical topic — although it is one of the most central subjects, namely the CCVS. Therefore, the results should be repeated in other domains.

4.2 Conclusions and pedagogical suggestions

This study is one of the first to investigate the reading of medical texts at the processing level via eye movements among actors at different levels of medical expertise. Furthermore, in this study, the development of biomedical knowledge concerning the CCVS among medical students was inspected generally and in relation to success on a clinical reasoning task. The study indicates that there are differences among medical students in their clinical reasoning skills at the early stage of the clinical phase, and interestingly, more successful and less successful students already slightly differ in terms of their performance on the medical entrance examination and their amount of misconceptions concerning the CCVS in the preclinical phase, although the differences were not statistically significant. Thus, it seems evident that we should be able to recognise potential lower achievers and find solutions to support this group right from the beginning of their medical education.

This study also highlighted the importance of illness-script activation in the diagnostic process and strengthened the previous finding implying that enabling factors may act as illness-script activators, at least among already graduated physicians (see also Boshuizen et al., 2012). Therefore, based on these results, our pedagogical suggestion is to make the role of illness scripts in the diagnostic process visible in medical education. Further, the crucial role of anamnesis as a potential illness-script activator in patient encounters should be highlighted for students.

This study focussed on medical expertise, but it also has implications for many other complex domains. On a more general level, this study contributes, for example, to the discussion among learning researchers related to the interaction of theory and practice in the development of professional expertise (e.g. Tynjälä & Gijbels, 2012). However, only a few studies have connected a longitudinal perspective related to the development of theoretical knowledge to the level of practical performance; hence, more research is needed in this area.

Acknowledgements

The authors wish to thank the Academy of Finland for the financial support for the LeMEd-project, 128892. The first author also wishes to thank the Finnish Cultural Foundation of personal grant enabling the writing of this manuscript. Further, we thank the Faculty of Medicine in the University of

Turku, Finland, for the excellent co-operation during data gathering. We also thank our participants, who gave their time and effort for this study.

References

Ahopelto, I., Mikkilä-Erdmann, M., Olkinuora, E., & Kääpä, P. (2011). A follow-up study of medical students' biomedical understanding and clinical reasoning concerning the cardiovascular system. *Advances in Health Sciences Education*, 16, 655-668.

Boshuizen, H. P. A., & Schmidt, H. G. (1992). On the role of biomedical knowledge in clinical reasoning by experts, intermediated and novices. *Cognitive Science*, *16*, 153–184.

Boshuizen, H. P. A., van de Wiel, M. W. J., & Schmidt, H. G. (2012). What and how advanced medical students learn from reasoning through multiple cases. *Instructional Science, 40*, 755–768.

de Bruin, A. B. H., Schmidt H. G., & Rikers, R. M. J. P. (2005). The role of basic science knowledge and clinical knowledge in diagnostic reasoning: A structural equation modelling approach. *Academic Medicine, 80* (8), 765–773.

Charlin, B., Boshuizen, H. P. A., Custers, E. J., & Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical Education*, *41*, 1178–1184.

Chi, M. T. H. (2005). Commonsense Conceptions of Emergent Processes: Why some misconceptions are robust. *The Journal of the Learning Sciences, 14* (2), 161–199.

Custers, E. J. F. M., Boshuizen, H. P. A., Schmidt, H. G. (1998). The role of illness scripts in the development of medical diagnostic expertise: Results from an interview study. *Cognition and Instruction, 16* (4), 367–398.

Ericsson, K. A. (2007). An expert-performance perspective of research on medical expertise: the study of clinical performance. *Medical Education*, *41*, 1124–1130.

Ericsson, K. A. (2016). *Peak: Secrets from the new science of expertise.* Boston: Houghton Mifflin Harcourt.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211–245.

Feltovich, P. J., & Barrows, H. S. (1984). Issues of generality in medical problem solving. In: H. G. Schmidt & M.L. de Volder (Eds), *Tutorials in Problem-Based Learning*. Maastricht: Van Gorcum Assen.

Field, A. (2009). *Discovering Statistics Using SPSS*. (3rd edition). London: Sage Publications.

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye Tracking. A Comprehensive Guide to Methods and Measures*. New York, NY: Oxford University Press.

Hyönä J. & Niemi, P. (1990). Eye movements in repeated reading of a text. *Acta Psychologica, 73,* 259–280.

Just, M. A. & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87 (4), 329–354.

Kaufman, D. R., Keselman, A., & Patel, V.L. (2008). Changing conceptions in medicine and health. In S. Vosniadou (Ed.), *International Handbook of Research on Conceptual Change*). New York: Routledge.

Kuipers, B., & Kassirer, J.P. (1984). Causal reasoning in medicine: analysis of protocol. *Cognitive Science, 8*, 363–385.

Lavrakas, P. J. (2008). Encyclopaedia of survey research methods. Accessed in April 2017, DOI: http://dx.doi.org/10.4135/9781412963947.

Mann, D. T. Y., Williams, A. M. & Ward, P. & Janelle, C. M. (2007). Perceptual-Cognitive Expertise in Sport: A Meta-Analysis. *Journal of Sport & Exercise Psychology, 29*, 457–478.

Michael, J. A., Wenderoth, M. P., Modell, H. I., Cliff, W., Horwitz, B., McHale, P., "…" & Whitescarver, S. (2002). Undergraduates' understanding of cardiovascular phenomena. *Advances in Physiology Education, 26*, 72–84.

Mikkilä-Erdmann, M., Södervik, I., Vilppu, H., Kääpä, P., & Olkinuora, E. (2012). First-year medical students' conceptual understanding of and resistance to conceptual change concerning the central cardiovascular system. *Instructional Science*, 40, 745-754.

Monajemi, A., Schmidt, H. G., & Rikers, R. M. J. P. (2012). Assessing patient management plans of doctors and medical students: an illness script perspective. *Journal of Continuing Education in the Health Professions, 32*, 4–9.

Nivala, M., Lehtinen, E., Helle, L., Kronqvist, P., Paranko, J. & Säljö, R. (2013). Histological knowledge as a predictor of medical students' performance in diagnostic pathology. *Anatomical Sciences Education, 6* (6) 361-367.

Patel, V. L., Evans, D. A., & Groen, G. J. (1989). Biomedical knowledge and clinical reasoning. In D. A. Evans, & V. L. Patel (Eds.), *Cognitive science in medicine: Biomedical modelling*. Cambridge, MA: The MIT Press.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124* (3), 372–422.

Schmidt H. G, Boshuizen H. P. A. (1993). On acquiring expertise in medicine. *Educational Psychology Review, 5*, 205–221.

Schmidt, H. G, & Rikers, R. M. J. P. (2007). How expertise develops in medicine: knowledge encapsulation and illness script formation. *Medical Education, 41*, 1133–1139.

Tynjälä, P. & Gijbels, D. (2012). Changing World: Changing Pedagogy. In P. Tynjälä, M-L. Stenström & M. Saarnivaara (Eds.) *Transitions and transformations in learning and education* (pp. 205-222). Dordrecht: Springer.

Van de Wiel, M. W. J., Boshuizen, H. P. A. & Schmidt, H. G. (2000). Knowledge restructuring in expertise development: Evidence from pathophysiological representations of clinical cases by students and physicians. *European Journal of Cognitive Psychology, 12* (3), 323–355.

Vilppu, H., Mikkilä-Erdmann, M., Södervik, I., & Österholm-Matikainen, E. (2017). Exploring eye movements of experienced and novice readers of medical texts concerning the cardiovascular system in making a diagnosis. *Anatomical Sciences Education*, 10, 23e33.

Woods, N. N. (2007). Science is fundamental: The role of biomedical knowledge in clinical reasoning. *Medical Education*, 41, 1173–1177.