Chapter 6 End-User Evaluation

Following the refinements to the user interface, and inclusion of additional elements to encourage reflection, as described in the previous chapter, an evaluation was undertaken involving the intended end-users of this simulation environment. The focus of this evaluation was to explore:

1. Whether the overall flow of a single consultation is adequate to represent an authentic medical consultation.
2. Are the history-taking, examination, investigation, and management interfaces sufficiently authentic and usable, such that they support the learning process?
3. Does the use of multiple consultations enhance the authenticity of the learning experience?
4. Is SIMPRAC able to support learner reflection?
5. Does SIMPRAC encourage learner reflection?
6. In what ways can support for reflection by the learner be improved?
7. Are users engaged or distracted by the simulation, and what are their opinions on this learning environment.

This chapter includes an outline of the methods used in the evaluation, and is followed by the evaluation results. The results come in two parts. First, are the qualitative observations made during the think-aloud sessions. Second, are the quantitative results from the questionnaire that was administered to all users, as well as the activity data coming from the activity logs and action records.

6.1 Users

This evaluation cycle involved ten medical students, five general practitioners, and two consultant medical officers (see Table 7). The medical student and general practitioner groups represent the potential end-users of the software. While a minimum of five individuals are required to undertake an effective think-aloud evaluation (Nielsen, 1993), a larger pool of student participants was chosen, to gain additional qualitative insight into how student learners respond to the layer of reflection added to SIMPRAC. It was also believed that it would be informative to compare the usage of the software by these two groups. The consultant medical officers were specialists in Chemical Pathology and both had extensive experience in the management of hyperlipidaemia. As the number of individuals involved in specialist management of hyperlipidaemia is relatively small, it was not feasible to have five specialist medical practitioners use the software. This group was not intended as an end-user of the software, but an evaluation of their activity would inform appropriate scoring of activity items, as well as providing a series of users against which future students or general practitioners could compare performance.

All participants were volunteers, and approval for the study involving medical students was obtained from the University of Sydney Human Research Ethics Committee and the Graduate Medical Program Research Committee. The medical students were recruited by sending an invitation to participate, via email, to all students enrolled in the final two years of the four year University of Sydney postgraduate medical program. All participants were asked to read an information sheet (see Appendix D, page 246) before consenting to participate. Written informed consent was obtained from each of the medical students (see Appendix E, page 248). The general practice and specialists

medical practitioners were recruited by a direct approach, with a request to participate in an evaluation of the educational software.

6.2 Methods

Each evaluation involved observation using the simplified think-aloud (Nielsen, 1993) followed by administration of the evaluation questionnaire (See Appendix C, page 241). At the beginning of each evaluation, each user was informed that the purpose of the session was to evaluate the software, and not themselves. Each user was also given a brief demonstration on how to use the application. This was achieved by the author logging into the demonstration case, and giving a five minute overview of all the patient interaction elements. The end of consultation review screens were not demonstrated at this time.

The think-aloud sessions were undertaken as described in more detail on page 89 of Chapter 5. Each participant was asked to think-aloud, stating what they were doing and thinking as they used the interface. Since it was emphasised that we were evaluating the software, not the users, they were encouraged to critique their experiences as they worked through the case. In addition, all use of the software was time-stamped and logged by SIMPRAC for later analysis.

After each participant had completed their period of using the simulation, they answered a questionnaire that was comprised of three parts (see Appendix C). Part A was designed to collect demographic information, including sex, year of medical course, and previous experience with computer-based simulation environments. Part B was a series of semantic differential responses to each aspect of the interface and about the reflection, and was scored using a five point Likert scale (Jackson and Furnham, 2000).

Questions were asked in both the positive and negative sense. Finally, part C was a series of open questions that asked the user what they liked most and least about the simulation, as well as asking for general comments.

Table 7: Users and software configuration used during each evaluation. Software configuration 1 included Microsoft Windows 2000, Java Software Development Kit 1.4.1_01, Microsoft Internet Explorer 6.0. Software configuration 2 included Microsoft Windows XP, Java Software Development Kit 1.4.2, and Microsoft Internet Explorer 6.0.

User	Status	Gender	Year of Medical Course	Software Configuration
ST01	Student	F	3	1
ST02	Student	M	3	1
ST03	Student	F	3	1
ST04	Student	F	4	1
ST05	Student	M	4	1
ST06	Student	F	4	1
ST07	Student	F	4	1
ST08	Student	F	3	2
ST09	Student	F	3	2
ST10	Student	M	4	2
GP2[#]	General Practitioner	F	N/A	1
GP3	General Practitioner	M	N/A	2
GP4	General Practitioner	F	N/A	2
GP5	General Practitioner	M	N/A	2
GP6	General Practitioner	F	N/A	2
EX1	Specialist	M	N/A	2
EX2	Specialist	M	N/A	2

The numbering starts at GP2 because GP1 was a test user during the development process.
N/A: Not Applicable.

Comparisons between Student and General Practitioner group data was assessed using a two tailed Students t-test assuming unequal variances. Linear associations between properties were assessed by calculating Pearson's correlation coefficient. Associations were considered to be significant if the p value was less than 0.05. Non-linear associations were assessed for selected data sets, as indicated later, by first curve fitting the data, then recalculating the correlation coefficients on the predicted values. Statistical analyses were performed with SPSS version 12.0 statistical software (SPSS

Incorporated, Chicago). Curve fitting was undertaken using DataFit 8.0 (Oakdale Engineering, Reading).

6.3 Results

The users completed between one and four consultations. Detailed data logs were not available for the first four medical students due to the evolving nature of the application. As a result of the deficiencies in the data logging, the time these four students spent reviewing their activity and assigning relevance scores could only be estimated (see Table 8 to Table 10 on page 135 for summary data). It was calculated as the time from the final submission of the management options, at the end of the consultation, to the time the user submitted their relevance scores. This calculation slightly over-estimates the actual duration of the review process, as it does not allow for the few seconds it takes for the review component to initialize. Based on the log data for the subsequent students, it took the review component an average of 13 seconds to display at the end of the first consultation, but only 7 seconds at subsequent consultations.

The ten medical students were evenly split over the last two years of the four year University of Sydney postgraduate medical program. The fact that there were equal numbers from both years occurred by chance.

The consultations took from 47 to 107 minutes to complete. On average, the students (74 ± 20 minutes) (Mean ± SD) used the application for longer than general practitioners (65 ± 16 minutes), although this was not statistically significant due to the wide variation, and limited number of participants. In four of these cases, the evaluation was terminated before the fourth consultation because the users had other commitments, and the evaluation had taken longer than the 60 minutes they had expected. In the case

of ST10, the evaluation was terminated, as a result of the user becoming frustrated with the inability of the simulation to recognise their questions and provide appropriate responses.

6.3.1 Think-aloud observations

Very little direct feedback was provided by the users on how the application could be improved. Instead, the users became focused primarily on "solving" the clinical case. A short summary of each think-aloud session is given below for each user. These summaries highlight some of the issues that were discovered during the evaluation, and these are discussed in detail in Chapter 7.

ST01: Female in year 3 of the medical course.

Prior to commencing the simulation ST01 expressed her doubts about her ability to undertake the simulation due to lack of medical knowledge. This student used the application for almost 57 minutes before terminating the encounter at the end of the second consultation as a result of a shortage of time. During the first consultation the student asked 24 questions in total. On four occasions they used the free text method but this only retrieved a suitable match twice. The other 20 questions were asked from the category lists. This user was observed to have some difficulty elucidating the patient's vital signs. For example, when using the thermometer it had to be explained to the user that they should select the head and neck view of the patient and click over the patient's mouth. The user also wanted to know the respiratory rate but this option was not available. ST01 was unfamiliar with the management of hypertriglyceridaemia in the presence of Glycogen Storage Disease, so an explanation was given to the student. During the review process this student used the bar chart representation of their actions but did not use the pie or line chart representations.

ST02: Male in year 3 of the medical course.

ST02 used the application for 59 minutes and, as with ST01, terminated the session early, in this case, just before the end of the second consultation, due to time constraints. During the examination of the patient, this student wanted to know the liver span given that the palpation tool had indicated that the liver was enlarged. This information was not available. While reviewing the patient's test results, ST02 indicated that they would like abnormal results to be flagged appropriately, as is usually done with pathology reports. For example, abnormally high results to be marked with an H, bold, or highlighted in a different colour. During the review phase, this user only reviewed the bar charts.

ST03: Female in year 3 of the medical course.

The application was used for 56 minutes by ST03 and, as with ST01, ST03 terminated the session at the end of the second consultation. At the start of the evaluation, having read the introduction, the user appeared confused about how to proceed. When asked, she indicated that she wanted to know if the test at the shopping centre was a blood test. As this information was not available from the simulation, I indicated that it was. Some difficulty was experienced initially when examining the patient but this improved rapidly after the first few of attempts. When scoring her actions this student indicated that, "not everything is immediately relevant but you would perform them anyway when seeing the patient for the first time". At the beginning of the second consultation the student wanted to ask a series of general questions, including: "Have you been ill since your last visit?", "Have you any concerns?", "Have you made any changes to your diet?", and "what is your level of exercise?" These questions were not matched by any of the options in the question database, and had to be answered by the author. This was the only user to use the medical record component to any extent.

ST04: Female in year 4 of the medical course.

ST04 completed all four consultations in 59 minutes. Instead of using the medical record component provided with the application, this student recorded the patient information and their observations on a writing pad of their own. This was the only user to use this strategy.

At the end of the history component, on being shown the hypothesis screen, ST04 asked if it was absolutely necessary to enter a diagnostic hypothesis. I indicated that it was not compulsory, so she elected to proceed to the physical examination without entering an hypothesis. This function was never used by this student and when prompted by SIMPRAC to do so, the screen was ignored. It was not clear why ST04 did not want to undertake this process. When specifically asked about this point she indicated that she just did not want to do so. This was an unexpected result, given the emphasis the postgraduate medical program has on the hypothetico-deductive approach to medical reasoning.

ST04 said she would like to explain to the patient the risk and benefits of cornstarch in the context of the patient's illness, as well as the risks of very high triglycerides. She indicated she would also explain the risk and benefits of the various therapeutic options including Gemfibrozil and fish oil. Depending on the outcome of these deliberations with the patient, she would then prescribe a course that was mutually acceptable. All three options were selected from the management interface while the student was explaining her thoughts to me. All the options were selected and saved. At this point ST04 indicated that she hadn't wanted to prescribe all three of cornstarch, fish oil, and Gemfibrozil but had wanted some feedback from the patient first.

When classifying their actions as critical, relevant, or not relevant, this student indicated that, while not all elements would be relevant in light of the information available, one would still ask these questions or perform these actions. Only the bar chart was reviewed but this was done for all stages and against both the expert and student peer group.

At the end of the evaluation session, this student said she thought this system was good, as it was often difficult to see patients due to the large number of medical students. This simulation was an alternative way of getting practice.

ST05: Male in year 4 of the medical course.

ST05 completed all four consultations in 105 minutes. This student began by asking the patient her age. He indicated that since there was significant hyperlipidaemia, syndrome X needed to be considered, and so went on to ask questions regarding symptoms of diabetes. Noting that the patient had very few symptoms, ST05 noted that this made the diagnosis of diabetes mellitus less likely. Given the history of hyperlipidaemia the student indicated he needed to search for the complications of hyperlipidaemia such as peripheral vascular disease, and so asked whether the patient had had any cold extremities. ST05 then elucidated the history of Glycogen Storage Disease. At this point, he indicated that he had no knowledge of the disease and would need further information. The information sheet was given to the user. Based on the information supplied, ST05 systematically excluded the long-term complications of Glycogen Storage Disease. After completing the history section, the student indicated that he was uncertain as to what needed to happen next. When asked, “what would you do in a normal consultation”, he moved on to physical examination of the patient. At this point the student indicated that he would like an overview of the patient’s general appearance.

As a photograph of the patient was not available within the application, this information was given verbally.

When ordering investigations, this student would like to be able to see what investigations had already been ordered during that consultation. ST05 also wanted to be able to view investigations from past consultations as a cumulative list, as is frequently done on pathology reports.

ST05 wanted to provide education to the patient on her medical condition, offer dietitian review and prescribe allopurinol for management of hyperuricaemia. None of these options were available. Following this feedback, allopurinol was added as an option to the management database.

When scoring his actions, this student indicated that he would like more levels than just critical, relevant or not relevant.

At the beginning of this consultation, ST05 indicated that medical students were, on the whole, unfamiliar with what to do when reviewing patients, as they seldom got to see returning patients, because the time spent in any one term was fairly short.

ST06: Female in year 4 of the medical course.

This student completed all four consultations in 84 minutes. ST06 found it easy to use SIMPRAC and was given the information sheet on Glycogen Storage Disease (GSD) after asking for further information on the disorder.

In the first consultation, ST06 did not request a triglyceride estimation. However, she did in the second consultation. While the triglyceride concentration was high at 4.5 mmol/L, it was much lower than it had been in the first consultation. Because the student had not ordered the test in the first consultation, there was no way for her to know that there had been an improvement, and that the management she had selected had been very effective. This was explained to the student so that she was not under the misconception that the management had been ineffective.

When reviewing the charts at the end of the third consultation, the student read that the case authors had thought that a question on muscle aches and pains (myalgia) was critical. ST06 said that she thought this was only relevant for therapy with HMG CoA Reductase Inhibitors (statins). I informed her that, although less common than for statins, muscle inflammation (myositis) and myalgia are documented as a side effect of fibrates. During the fourth consultation, the student did ask about the presence of myalgia.

ST07: Female in year 4 of the medical course.

All four consultations were complete by ST07 in 85 minutes. This user only asked questions from the category lists and never used the free text method. The reason for this was not elicited.

After taking some history and learning the patient had Type 1B GSD, ST07 indicated she needed more information on the disease. ST07 was given the GSD information sheet. Based on this information, she asked the patient if she had a family history of GSD. When getting the response that there was no family history of illness the student

indicated that it was not surprising that there was no family history, given the autosomal recessive mode of inheritance.

ST07 did not appear to have any difficulty using the application but was not able to find all the examination items she was looking for, such as respiratory rate. At the end of the review phase ST07 said, “I get frustrated with the technology and lose patience”.

ST08: Female in year 3 of the medical course.

ST08 completed all four consultations in 107 minutes. This student looked at the history-taking screen and appeared uncertain about how to take a history from the patient. However, after a short explanation, no difficulty was experienced. This student prescribed uncooked cornstarch in accordance with the information sheet. However, she also wanted to discuss more deeply with the patient issues related to compliance, and expressed some disappointment at not being able to do so. ST08 also wanted to refer the patient to a specialist, and a dietician, for further advice. Again she was disappointed at not being able to do so. During all stages of the consultation, this student was observed to be constantly verbalizing and reflecting on the information that was provided, how it could be interpreted, and what it meant for the patient’s management. This was in distinct contrast to most other users, who had to be constantly prompted for their thoughts. As with many other users, ST08 wanted a general overview of the patient’s general appearance.

At the end of the first consultation, ST08 wanted to review the patient in 14 days to review progress. During the review process at the end of the second consultation this student noted that a question related to change in weight had been considered relevant by the case author. This student, quite reasonably, asserted that this question was not

that relevant given the fact that it had only been 14 days since the patient's last presentation. This problem arose from the fact that the user is able to select the time interval after which they want to review the patient, but this does not affect the patient outcome. This could be resolved simply in at least two ways, as discussed in more detail in Chapter 7. One way would be to remove the ability of the user to select a follow up period. The other would be to indicate at the introduction to each consultation after the first, the time elapsed since the previous consultation.

In contrast to this user's constant reflection during the consultation, she spent relatively little time scoring her actions or comparing her activity to her peers or the case expert.

ST09: Female in year 3 of the medical course.

ST09 completed all four consultations in 70 minutes. This user had no difficulty using the application, but did use the inspection tool a lot during the examination stage of the first consultation. She did this in an attempt to gain an overview of the patient's appearance. When ordering investigations ST09 asked whether it was possible to view the results of investigations before adding further investigations. While not reflecting actual practice, I indicated that this was indeed possible.

This student was more flexible in the sequence with which she conducted the first consultation than other users. She started by asking some questions, examined the patient, ordered investigations, and selected some management options. However, before saving her management options, she went back to take some further history regarding overseas travel, checked the patient's weight, and added regular exercise to the list of management items. She then requested a hepatic ultrasound before saving her management selections.

ST10: Male in year 4 of the medical course.

ST10 completed three consultations over 61 minutes. In contrast to all other users, this individual had had previous clinical exposure to patients with Type 1B GSD and was aware of the problems these patients experience with recurrent infections. After using the free text method of asking questions with limited success six times, ST10 said, “it might be easier to see what it will allow me to ask”. Thereafter this user searched the question category lists for appropriate questions. ST10 was disappointed to find no questions regarding strokes or transient ischaemic attacks. These questions had not been added to the corpus of available questions. This student also indicated that he wanted to determine if the patient was worried by her condition, but was unable to assess this from the available questions.

In contrast to the other users involved in the evaluation, this user was more focused on the infective complications of this disorder, rather than the biochemical derangements such as hypertriglyceridaemia or hyperuricaemia.

When selecting his management options, this user indicated that he would have preferred more specific options. For example, the ability to prescribe a diet low in simple carbohydrates but higher in complex carbohydrates.

While assigning his relevance scores to each item at the end of the first consultation ST10 said, “I did lots of little bits and pieces, didn’t I”. Then, when reviewing his activity, ST10 indicated that he would have liked to have examined the patient’s lungs, but was “thrown” by the interface. The student also noted that cornstarch was regarded as a relevant management option and asked why this was the case. I indicated that it was the treatment of choice for the metabolic derangements seen in Type 1 GSD. ST10

indicated that he had not been aware of this information. At this point the information sheet on GSD was offered to the student. In contrast to the other users who were given this sheet, this user did not read the information at any stage.

Following the question about cornstarch, this student asked the patient about her use of uncooked cornstarch.

ST10 indicated that he would also like a cumulative historical view of the pathology results.

During the third consultation, ST10 indicated that he was getting frustrated because it was not possible to ask open ended, less specific, questions.

During the review process, ST10 indicated that the question, “what treatment have you used” was an attempt to determine the patient’s compliance. As it was not possible to ask directly about this issue, this was the closest question to what he wanted. Therefore the question should have been scored as relevant by the case author, rather than not relevant.

Following the review process, ST10 indicated he did not want to continue with another consultation as he was losing interest and getting frustrated at not finding what he wanted to ask.

GP2: Female.

[Note: The general practitioners were numbered from 2 as the username GP1 had already been used as a username during the development of SIMPRAC.]

Three consultations were completed in 66 minutes. This general practitioner took a comprehensive history and was disappointed when the virtual patient was unable to answer her more specific questions. For example, the request, “can you please describe your diet?” is available within the corpus of questions. In response to the answer from the patient GP2 said, “as a diabetic you would do well”. GP2 then went on to ask: “do you have yellow spread on your bread?”, “do you use butter or margarine?”, and “what do you have for afternoon tea?” None of these questions were contained in the application’s corpus of questions.

Having asked the patient for her age, GP2 indicated that was normally provided within the practice, having been determined by the receptionist when the date of birth was recorded, and it was not necessary for the GP to ask.

Just before GP2 began examining the patient, I suggested she might like to ask about past medical history. On learning that the patient had a Glycogen Storage Disease and after reading the information sheet, the general practitioner indicated that the characteristic facial appearance, and short stature of patients with this condition, would have cued her to ask for more information on the past medical history, if seeing such a patient in their rooms for the first time.

Having completed the history, examination, and investigations, GP2 indicated that at that point in time she would like to discuss additional issues such as: does the patient understand the genetics of the disorder, what contraceptive is the patient using, and does the patient want to get pregnant.

When selecting her management options, GP2 said she would like to provide extensive dietary advice and refer the patient to a dietician. GP2 also wanted to refer the patient to a specialist with expertise in the management of this disorder. Having been informed that for the evaluation, referral was not possible, GP2 then selected from the available options.

GP2 indicated that she found the review component very valuable. Especially being able to get immediate feedback on what elements may have been missed. GP2 did, however, disagree with some of the scores given to some elements. For example, in the first consultation, GP2 thought questions about birth history were irrelevant in an adult. Also, some of the examination elements would need to be done for completeness. As an example, one would not just palpate the right upper quadrant for the liver but would examine the whole of the abdomen. While the former had been regarded as relevant, the remainder of the examination had been scored as not relevant by the author. In the third consultation, GP2 stated that she didn't think the use of allopurinol or fish-oil were relevant management options.

During the second consultation, GP2 wanted to discuss strategies for improving compliance with the cornstarch by making it more palatable. This was not possible within the current structure of the simulation.

GP2 ended the evaluation session at the completion of consultation three, as GP2 had limited time, and clinical commitments elsewhere. GP2 also said she thought this system had potential for general practitioner's to audit themselves and their practice of medicine. This general practitioner did not think this particular medical case was a good

case for general practitioners, and that it was typical of hospital specialist “silo” thinking.

GP3: Male.

All four consultations were completed in 77 minutes. GP3 indicated that he would like to have a list of the questions that he had already asked. When ordering investigations for the virtual patient, GP3 indicated that it should be possible to use common abbreviations and alternative spellings when selecting these items. For example, LFT for liver function tests, or urate versus uric acid respectively. When reviewing and scoring his activity, GP2 indicated that an image of the patient would make the case more realistic, rather than having to use the inspection tool all over the body. When comparing his activity and scores to the suggestions made by the case author, GP3 thought that the questions, “why have you come today?” and “when did it start” were redundant and therefore not critical.

During consultation two, GP3 wanted to ask, “how do you feel?” as an open-ended question to cover a number of issues, including drug side-effects. This question was not available within the software.

When reviewing his activity at the end of consultation three, GP3 disagreed with the measurement of blood pressure as being “not relevant”. While the case author had felt that measurement of the blood pressure was not relevant to the “case at hand”, and the management of the major presenting problem, this general practitioner held the view that blood pressure measurement was a relevant part of his practice. Within the context of general practice, health monitoring and disease prevention is an important activity.

Again, as suggested by GP2, this is the difference between a wholistic general practice focus and a specialist, “silo”, approach.

It was also noted that the abdominal computerized tomography (CT) scan was reported as normal even though the patient had a large liver on palpation. The normal CT report was the default result, and a case-specific report, which should have demonstrated hepatic steatosis, had not been entered. GP3 also wanted to be able to assess the patient for osteoporosis using dual x-ray absorption densitometry but this was not available as an investigation option.

GP4: Female.

GP4 completed all four consultations in 52 minutes. Having read the introductory information about the patient, GP4 wanted to know if the blood had been taken while the patient was fasting. This information was not available within the software. I indicated that the patient had not been fasting. This general practitioner found the management options inadequate and wanted to initiate a number of lifestyle changes.

At the end of consultation one, GP4 said she preferred the term, “highly relevant”, rather than “critical”. She also indicated that the user interface was causing some difficulty, as not all options were available. This was further emphasized at the end of the second consultation, where she said she wanted to prescribe a regular exercise regime, evaluate compliance with diet and medication, and wanted to consider additional issues such as a Papanicolaou smear for cervical cancer screening, as well as the patient’s vaccination status.

GP4 thought the case was “one dimensional” as she would have expected more problems. She also stated, “I find it irritating that the relevance score is nit-picking over the terms used”. This general practitioner considered that the questions, “have you had any muscle aches and pains?” and “how have you been since your last visit?” covered the same issue, in the light of the patient being on medication that can cause myalgia.

This user hypothesized that the patient had hyperlipidaemia as part of the “metabolic syndrome”. She never asked about past medical history, and never learnt that the patient had Glycogen Storage Disease.

GP5: Male.

GP5 completed all four consultations in 47 minutes. At the beginning of the evaluation this general practitioner indicated he was not good with computers. He was noted to be relatively slow at typing compared to the other users. This user only selected the questions from the various lists and did not use the free text facility.

It was difficult to get this user to think-aloud and he progressed through the case and the evaluation relatively rapidly.

GP5 would have liked an easier method for examining for lymphadenopathy, as the status of the various lymphnode groups was not returned to the user as he used the palpation tool over relevant parts of the body. For example, when palpating the neck, the carotid pulse was reported as normal but no comment was made regarding the cervical lymph nodes.

GP6: Female.

Four consultations were completed in 85 minutes. GP6 was able to use the software without difficulty, but wanted to explore issues such as why the patient had had their blood tested in the first place. It was not possible for the GP to explore such issues in the current configuration. When the hypothesis screen was displayed for the first time, GP6 sat back and said, “oh, what have I got as a hypothesis”.

When learning that the triglyceride concentration was higher than when previously measured, she again wanted to know why. For example, was depression leading to poor compliance and worsening lipid profile?

Similar to other users, GP6 wanted a finer level of control over the management options, including the ability to negotiate with the patient an appropriate intake of uncooked cornstarch. As an alternative to cornstarch, she would have liked to consider the use of foods with a low glycaemic index.

This user also thought that some of the history questions should have been scored differently. At the end of the second consultation, she would have like to have discussed cervical screening with a Papanicolaou smear.

EX1: Male.

Four consultations were completed in 60 minutes. This user was not currently in clinical practice but had had extensive experience managing lipid disorders, and had managed a patient with Glycogen Storage Disease. During the consultation, when he learnt the patient had GSD, he said he would need to look up the condition, as it had been some time since he had seen anyone with this disorder. This user performed a large number of

examination items, but these were targeted at the cutaneous manifestations of hyperlipidaemia.

During the management selection stage, this user wanted to order dietician review but then ordered a diet high in carbohydrate before saving the option and being shown the review screen. At this point EX1 indicated that he had wanted to add additional items and suggested that a confirmation screen was required before ending the consultation. When reviewing his actions and assigning relevance scores, EX1 suggested that the actions should be in the order that they have been undertaken.

At the end of the case, EX1 stated he felt the immediate feedback was informing his questions in subsequent consultation as some items had been forgotten. He also suggested that the 'perfect' score should be available on the graph. I indicated that this information was available from the 'expert' comparison, and textual feedback when right-clicking the bar chart. Having considered this, EX1 then suggested that rather than having to right click the graph, the detailed information should be more easily accessible. For example, by being able to call up the information with a single click. EX1 also said he would like a better definition of critical, relevant and not relevant as he had noted that he had assigned more things as critical and relevant than the author.

EX2: Male.

EX2 completed four consultations in 43 minutes. This user specializes in the management of lipid disorder but had not previously managed a patient with GSD.

During the first consultation there was evidence that this user was not sure how all elements of the interface were to be used. For example, at the investigation screen EX2

requested a number of investigations including: Triglyceride, Cholesterol, HDL cholesterol, and LDL cholesterol. The results of each test were viewed after each one was ordered. An apolipoprotein E phenotype was then requested. The user saw that this was in the second test list, i.e. tests that have been requested but are unavailable until some time in the future. At this point the user deleted the test. The same thing was done with Apolipoprotein E genotype. Glucose, liver function tests, electrolytes, urea, and creatinine were then requested. EX2 then requested lipoprotein electrophoresis. Again, it appeared on the list of tests the results for which were not currently available. Again, EX2 deleted the test. As this had happened for the third time, I indicated to EX2 that although the test was not available currently, the results would be available at the next consultation. On learning this EX2 then re-ordered apolipoprotein E phenotyping, apolipoprotein E genotyping, and lipoprotein electrophoresis.

Later, EX2 said, "One would usually assess the thyroid function, but it doesn't usually affect the triglycerides. I will do it anyway. I'm not paying for it."

Although this user learnt in the first consultation that the patient had a past history of GSD, he did not ask for additional information on this disorder, despite not having previously seen a patient with this disorder.

While assigning his relevance scores, EX2 indicated that with his examination items, he had been trying to look for eruptive xanthoma. EX2 then said, "I should have asked about skin rashes". I indicated that it was too late to ask this in the first consultation. This was the first and only question asked during the second consultation. EX2 then looked for the results of the investigations requested in the first consultation, including

apolipoprotein phenotyping, apolipoprotein genotyping and lipoprotein electrophoresis. The results of all of these were inappropriate for the case, as they only had the default "normal" results. They should have been E3/E3, E3/E3, and type IV pattern on the lipoprotein electrophoresis with dominant VLDL band in the pre beta region, respectively.

During the management stage of the third consultation, EX2 added fish-oil and then requested Basikol. This latter medication, an extract containing plant sterols, was unfamiliar to the case author, and had not been included in the database.

At the beginning of consultation four, EX2 asked the patient about adverse reactions from their medication before re-checking the lipid values. EX2 then ceased the fish-oil because, "it doesn't seem to be working", and added a statin. This combination of a statin and a fibrate is an unusual combination that would generally only be undertaken by a specialist, because the risk of myalgia and myositis is increased synergistically over either drug alone. This is not a combination that one would expect a medical student or general practitioner to use.

Summary

Overall, the users did not have any major difficulties using the application once they had learnt how to use the various interface elements. For example, ST03 had some initial difficulty using the history-taking screen. However, after asking a couple of questions she was then able to proceed without apparent difficulty. All the users appeared to be engaged by the simulation and were more interested in the diagnosis and management aspects rather than providing feedback about the software interface. This

speaks to the engaging nature of the problem and the, at least, neutral nature of the interface for most people.

The consultations flowed appropriately, although there was some variation in the way people used the system. Most took a fairly linear approach by taking a history, performing an examination, ordering investigations, and then preparing a management plan. Others, such as ST09, took a less linear approach and moved from investigations back to further history-taking and physical examination, before finalizing the management plan.

The history taking, and management interfaces caused the most problems. These arose from structural problems in the case, as well as limitations in the software implementation. The limitations in the software implementation were noted in the formative evaluation and relate to the limited ability to match free text enquires with concepts within the database. The structural problems relate, firstly, to concepts not being present in the database. For example, users were unable to find questions on the patient's consumption of butter, as this particular question was not held in the database. The same problem was found for generic management terms such as "provide education". This very general and non-specific action had not been included in the list of available management actions. Nevertheless, it would be easy to add this to the list and assign it a relevance score for each patient state. The second structural problem was that appropriate responses were not available for examinations and investigations, where it had not been anticipated that users would choose these particular items. For example, GP3 requested an abdominal CT and EX1 requested lipoprotein electrophoresis. In both cases, this had not been anticipated, and results consistent with

the patient's condition were not present in the database. Despite these problems, the examination and investigation interfaces were well received. However, several suggestions for improvement on the presentation of results were made, with the idea of making these reports look more like those seen in clinical practice.

The use of multiple consultations was generally well accepted, with ST05 noting that in their medical training, patients were rarely seen a second time. Nevertheless, others such as ST09 experienced some frustration with the multiple consultations, when the patient's status did not appear to undergo much change. The reasons for this lack of change are discussed more fully in Chapter 7. Time-related issues were also experienced with the multiple consultations. For example, ST08 wanted to review the patient 14 days after the first consultation, and was surprised by the feedback she received regarding the case author's assessment of the relative relevance of different actions. It had not been anticipated that a user would review the patient after such a short follow-up period: the patient status and the relevance assessments, were made on the assumption that there would be a considerably longer follow-up period.

The students appeared to vary considerably in the degree to which they were able to verbalise their thoughts and reflect on their actions. At one extreme was GP5, who progressed rapidly through the case and had to be constantly prompted to verbalise his thought processes. At the other extreme was ST08, who was able to verbalise her thoughts well, and appeared to be constantly reflecting on her actions and the implications they had for the diagnosis and management of the patient, as well as on the understanding of the disorder. Reflection by users appeared to take place during the case and during the review process to a variable degree. As an example, ST07 asked about a

family history of GSD, having read the GSD information sheet. On learning that there was no family history, she indicated that this was not surprising given the autosomal recessive mode of inheritance of Type 1B GSD. This suggests there was some reflection on the value of asking about family history. During the review process, a number of users disagreed with the relevance assignments given by the case author, suggesting they were reflecting on the merits of their actions. Interestingly, all the users appeared to consider that concordance of their relevance assignments with the case author's assignments indicated that the author was indicating that what they did was "right", while discordance was interpreted as a statement of being "wrong". None appeared to consider this as just a framework for reflecting on the usefulness of their questions and activities. A number of users found the feedback provided by the charts and action lists useful, with both GP2 and EX2 indicating that it helped remind them of some elements that may have been missed.

6.3.2 Quantitative results

A summary of the activity of each user during the patient interaction components can be found in Table 8 to Table 10, starting on page 135. These data are illustrated in the figures below. Figures 38, 39, 40, and 41 show respectively the total time each user used the simulation, the time they spent interacting with the patient, the time spent reviewing the importance of each of their actions, and the time devoted to reviewing the comparative charts. It should be noted that comparisons between groups and individual users are only completely valid where the users have undertaken the same consultation and stream. That is, valid comparisons can only be made if the patient is in the same management state for each user. In consultation one, all users are in stream one and may be compared to each other. In subsequent consultations, the users are often in different streams for the same consultation. For example, from Table 10, on page 138, it can be

seen that expert one (EX1) was in streams 1, 1, 4, and 4 for consultations one to four. This pattern was not undertaken by any of the other users. It is possible to compare EX1 with ST01 and ST05 in consultation two, as they were all in stream 1. Thereafter, however, their pattern of progress was different.

A summary of the responses to the questionnaire can be found in Table 11 while a full listing of the questionnaire results can be found at Appendix I on page 266.

Table 8: Summary of medical student activity for consultation one to consultation four. Duration is the amount of time the user spent interacting with the patient. Time in parentheses is the total time using the application. Stream represents the state the patient was in for that consultation (see Figure 21 on page 60). Questions, Examinations, Investigations, and Management are the total number of these items selected by the user. "Review activity" is the amount of time the user spent reviewing their actions and deciding whether they were critical, relevant, or not relevant. "Review charts" is the amount of time the user spent reviewing the charts and feedback, and comparing themselves to either their peer average, or an expert.

User	Consultation	1	2	3	4
ST01	Duration (56.48 min)	35.18	7.55		
	Stream	1	1		
	Questions	22	1		
	Examinations	22	0		
	Investigations	5	4		
	Management	1	2		
	Review activity (min)	6.85	0.95		
	Review charts (min)	Unknown	Unknown		
ST02	Duration (59.18 min)	43.95	2.58		
	Stream	1	3		
	Questions	13	2		
	Examinations	21	0		
	Investigations	8	6		
	Management	1	Ceased		
	Review activity (min)	6.23			
	Review charts (min)	Unknown			
ST03	Duration (55.53 min)	38.05	6.37		
	Stream	1	3		
	Questions	23	1		
	Examinations	31	12		
	Investigations	7	7		
	Management	2	2		
	Review activity (min)	5.30	1.57		
	Review charts (min)	Unknown	Unknown		
ST04	Duration (58.88 min)	24.50	5.47	9.92	5.42
	Stream	1	4	4	6
	Questions	8	4	4	1
	Examinations	32	1	3	0
	Investigations	9	6	9	6
	Management	3	2	2	2
	Review activity (min)	3.13	0.77	0.67	0.30
	Review charts (min)	Unknown	Unknown	Unknown	Unknown
ST05	Duration (105.67 min)	55.43	9.92	6.02	4.53
	Stream	1	1	3	3
	Questions	47	7	3	6
	Examinations	53	0	0	0
	Investigations	8	7	8	10
	Management	2	3	3	3
	Review activity (min)	9.98	1.05	0.42	0.62
	Review charts (min)	8.58	2.65	2.03	1.07
ST06	Duration (84.42 min)	42.47	5.35	4.62	6.40
	Stream	1	4	4	6
	Questions	9	2	5	6
	Examinations	59	0	0	0
	Investigations	9	8	9	10

User	Consultation	1	2	3	4
	Management	4	4	4	4
	Review activity (min)	8.68	0.67	0.90	0.98
	Review charts (min)	4.85	1.78	2.00	2.48
ST07	Duration (85.38 min)	42.85	8.50	4.13	5.00
	Stream	1	4	4	6
	Questions	24	3	2	2
	Examinations	24	0	0	0
	Investigations	10	7	8	5
	Management	2	3	3	3
	Review activity (min)	5.95	1.00	0.95	0.55
	Review charts (min)	5.70	2.45	0.97	0.73
ST08	Duration (106.87 min)	67.10	9.22	9.10	6.58
	Stream	1	1	5	4
	Questions	20	5	3	4
	Examinations	30	0	0	2
	Investigations	10	1	5	5
	Management	1	1	2	4
	Review activity (min)	3.92	0.37	0.32	0.53
	Review charts (min)	4.35	0.93	0.37	0.32
ST09	Duration (70.53 min)	26.13	13.20	8.23	5.28
	Stream	1	3	3	3
	Questions	18	5	5	6
	Examinations	45	31	16	1
	Investigations	11	2	8	10
	Management	3	4	5	6
	Review activity (min)	4.00	1.10	1.03	0.70
	Review charts (min)	3.58	1.45	2.67	1.18
ST10	Duration (61.07 min)	27.85	6.32	8.15	
	Stream	1	3	3	
	Questions	40	8	16	
	Examinations	65	10	7	
	Investigations	10	3	7	
	Management	3	3	3	
	Review activity (min)	7.15	1.05	1.40	
	Review charts (min)	3.32	1.62	2.27	

Table 9: Summary of general practitioner activity for consultation one to consultation four.

User	Consultation	1	2	3	4
GP2	Duration (65.55 min)	35.30	5.95	3.53	
	Stream	1	1	5	4
	Questions	13	3	3	
	Examinations	11	0	1	
	Investigations	7	0	0	
	Management	2	2	3	
	Review activity (min)	3.62	0.37	0.35	
	Review charts (min)	8.87	1.27	3.33	
GP3	Duration (77.3 min)	18.97	11.12	9.83	7.48
	Stream	1	4	4	6
	Questions	12	3	1	0
	Examinations	19		3	3
	Investigations	14	6	7	1
	Management	1	1	2	2
	Review activity (min)	3.98	0.92	0.77	0.22
	Review charts (min)	10.88	2.50	1.25	6.68
GP4	Duration (52.23 min)	14.97	8.23	3.83	
	Stream	1	4	4	6
	Questions	5	4	4	
	Examinations	8	3	2	
	Investigations	5	5	1	
	Management	2	2	2	
	Review activity (min)	1.85	0.68	0.60	
	Review charts (min)	3.65	7.95	7.80	
GP5	Duration (46.75 min)	20.18	5.13	2.98	4.30
	Stream	1	4	4	6
	Questions	9	4	2	0
	Examinations	9	1	1	3
	Investigations	9	4	4	3
	Management	2	2	2	1
	Review activity (min)	2.37	0.60	0.48	0.35
	Review charts (min)	3.33	0.77	0.28	2.25
GP6	Duration (85.00 min)	45.52	11.05	3.27	2.63
	Stream	1	4	4	6
	Questions	14	3	3	3
	Examinations	11	0	0	0
	Investigations	5	3	1	0
	Management	2	2	2	1
	Review activity (min)	4.10	0.65	0.33	0.17
	Review charts (min)	6.57	4.67	2.03	0.52

Table 10: Summary of medical specialist activity for consultation one to consultation four.

User	Consultation	1	2	3	4
EX1	Duration (60.22 min)	25.33	6.23	4.83	2.45
	Stream	1	1	4	4
	Questions	13	2	5	3
	Examinations	43	0	0	0
	Investigations	13	5	5	5
	Management	1	2	3	3
	Review activity (min)	6.68	0.53	0.65	0.57
	Review charts (min)	5.78	1.82	2.00	0.77
EX2	Duration (43.47 min)	23.97	3.68	3.85	1.38
	Stream	1	1	4	4
	Questions	9	1	2	1
	Examinations	10	0	0	0
	Investigations	14	4	4	4
	Management	1	2	3	3
	Review activity (min)	3.12	0.62	0.47	0.25
	Review charts (min)	2.80	0.28	0.47	0.23

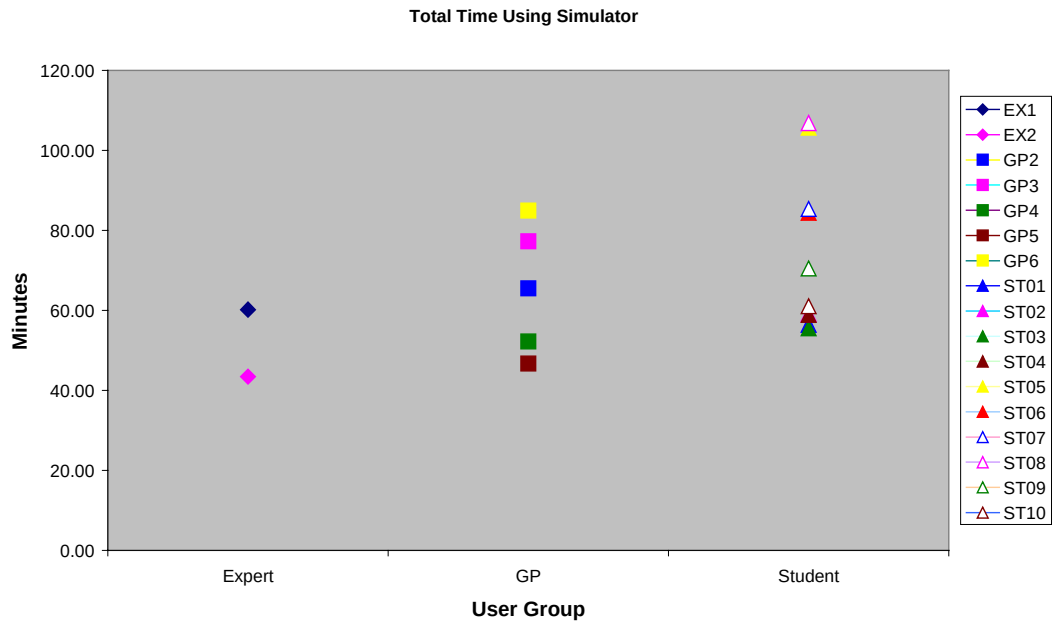

Figure 38: Total time each user spent using SIMPRAC

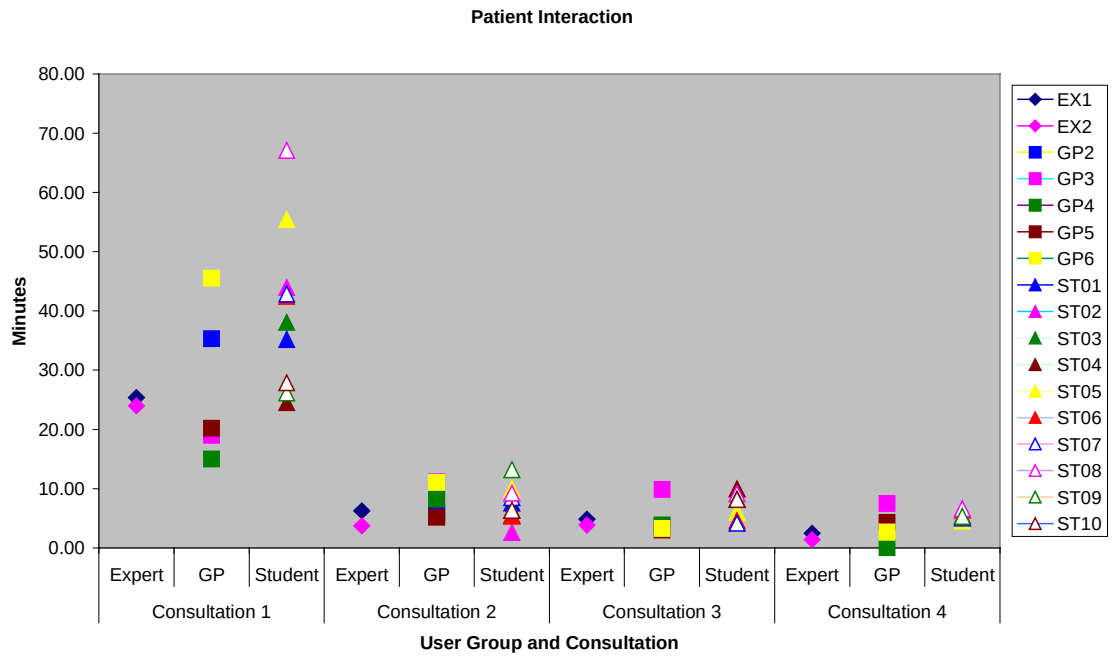

Figure 39: Total time in minutes each user spent interacting with the patient in each consultation. See Figure 9 to Figure 20 on pages 48 to 57.

From Figure 39, it appears as though the students spent more time interacting with the patient than either the experts or the general practitioners. However, as a result of the wide variance and small numbers, this difference was not statistically significant. Considerably more time was spent by each user reviewing their own activity and scoring the relevance of each item in the first consultation, than in subsequent consultations. To a large extent, this was thought to reflect the greater number of action items chosen during the first consultation, which had a predominantly diagnostic focus compared to subsequent consultations, which were more focused on assessing response to management. However, as illustrated in Figure 42, even when the duration of review is normalized for the number of actions undertaken in each consultation, there was still a tendency for the users to spend progressively less time reviewing their actions at the end of each successive consultation.

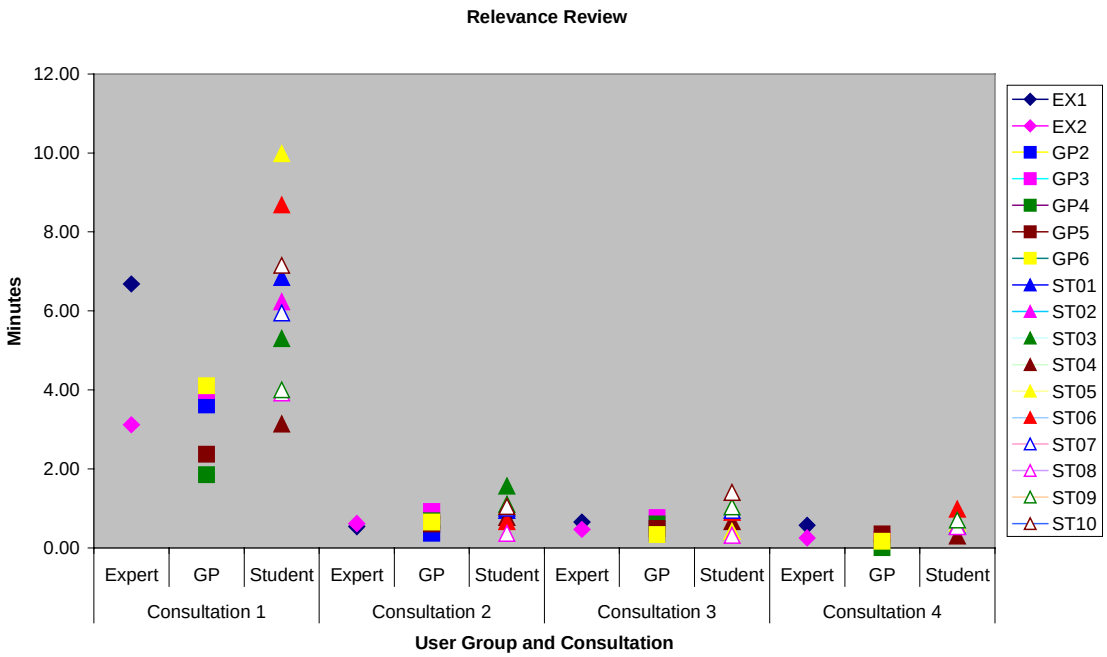

Figure 40: Total time in minutes each user spent reviewing the importance of their activity in each consultation. See Figure 25 on page 65.

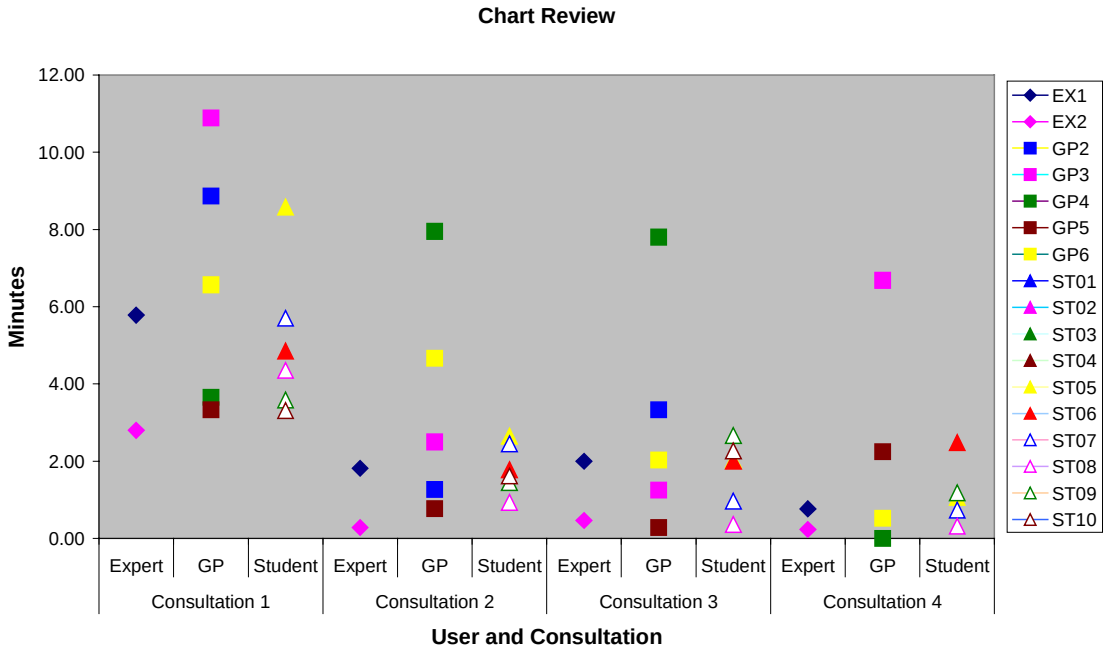

Figure 41: Total time in minutes each user spent reviewing the charts in each consultation. See Figure 26 to Figure 29 on pages 66 to 69.

Figure 43 to Figure 46 chart the number of questions, examinations, investigations, and management orders selected by each individual user for each consultation. Figure 47 charts the number of questions, examinations, and investigations selected by each user in consultation one as a proportion of the total number of possible questions, examinations, and investigations. Figure 48 is the same as Figure 47, except that it also includes information on which of those actions were regarded by the case author as being critical, relevant, or not relevant. Interestingly the pattern of use was similar for all groups, whether they were students, general practitioners, or specialists. However, general practitioners tended to ask fewer questions than students and performed fewer examinations than either students or specialists. From Figure 48 it is clear that the

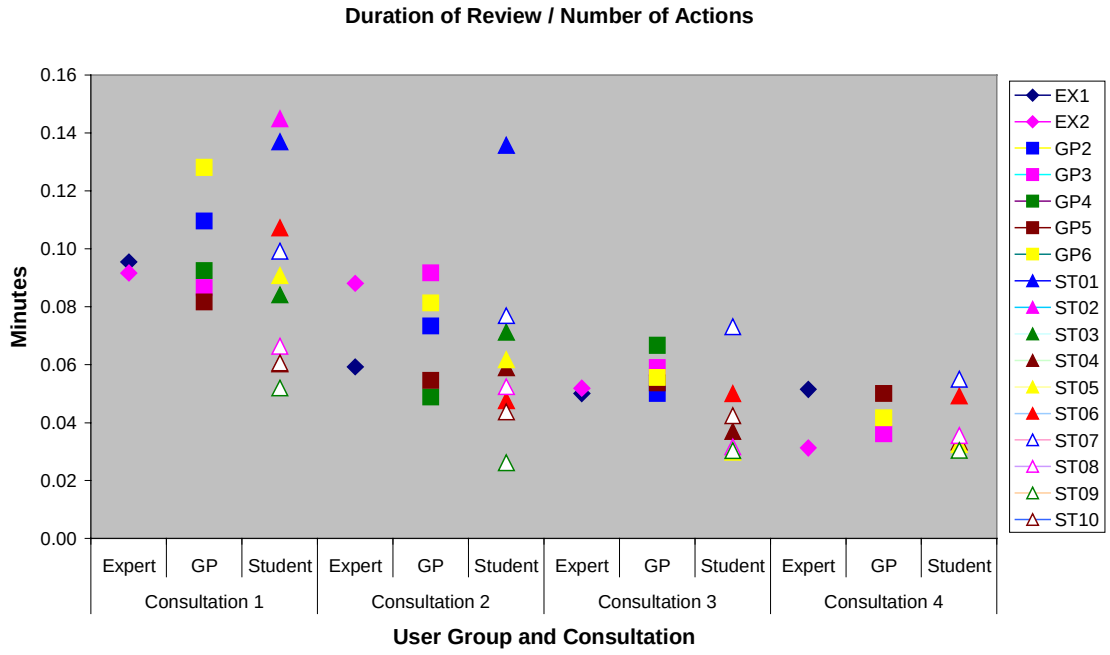

Figure 42: Time each user spent reviewing the importance of their activity divided by the total number of actions in each consultation.

higher number of questions in the first consultation was largely due to the greater number of questions asked by students that the case author regarded as not relevant. ST05 and ST10 asked many more questions in consultation one than all other users. In the case of ST05, and as stated before, having been given the information sheet on

Glycogen Storage Disease, this user asked a large number of questions to systematically exclude all the potential long-term complications of this disorder. In the case of ST10, this user was focussed on the infective complications and asked a very large number of questions that might suggest the presence of infection.

There was considerable variation in the number of examination items undertaken by the users. Observation during the think-aloud sessions, and of the log files, showed that some individuals used the inspection tool extensively. This was done to gain an overview of the patient, because a photograph of the entire patient was not available for inspection. Thus, in order to get a sense of the patient's overall appearance, a number of users used the inspection tool on multiple parts of the body. The specialists requested more investigations than either the general practitioners or the medical students in the first consultation. However, in subsequent consultations they requested similar numbers. In the management stage of the simulation, the specialist showed a clear progression in their management plan. Both specialists initially started with one intervention in the first consultation, two in the second consultation, and three in both the third and fourth consultations. Neither the general practitioners, nor the medical students demonstrated this clear stepwise escalation of management interventions.

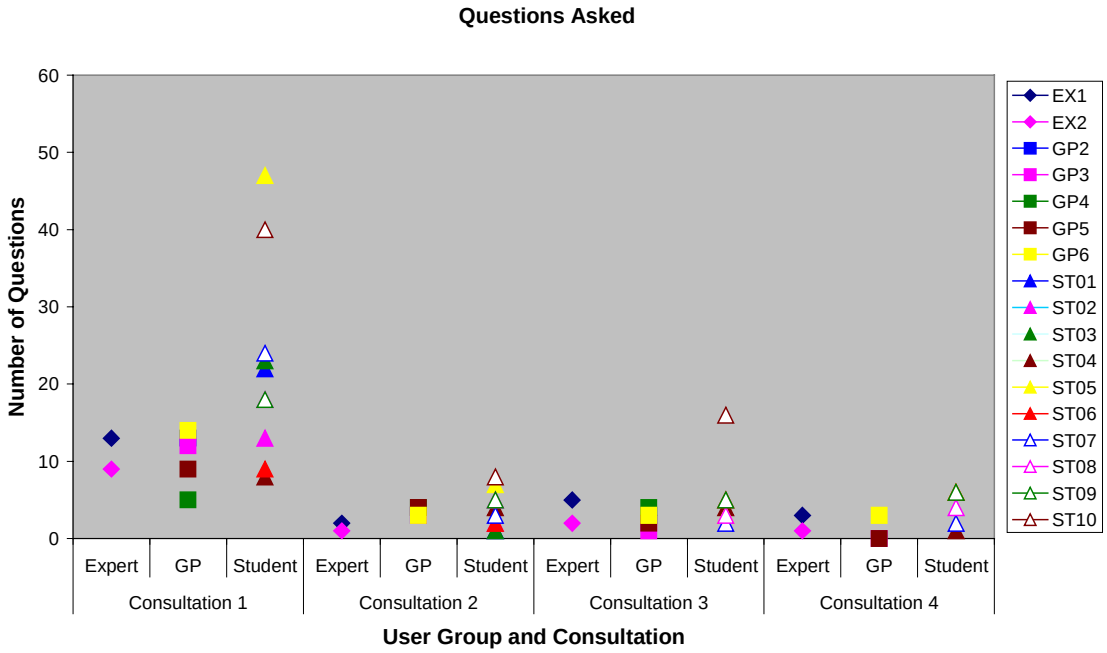

Figure 43: Number of questions asked by each user in each consultation.

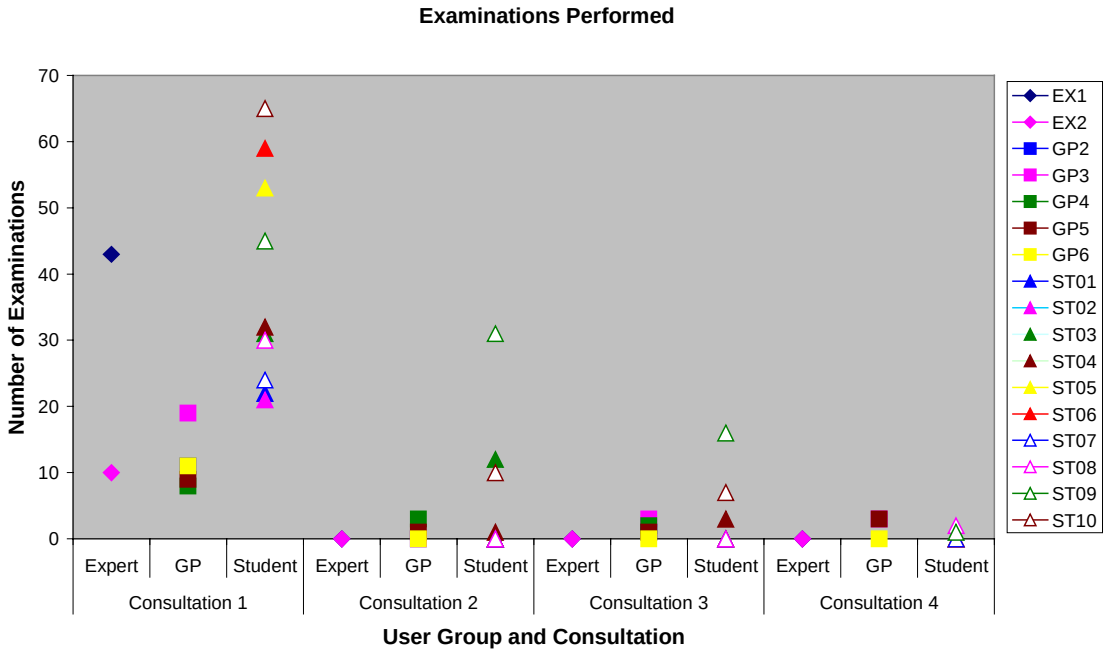

Figure 44: Number of examinations performed by each user in each consultation.

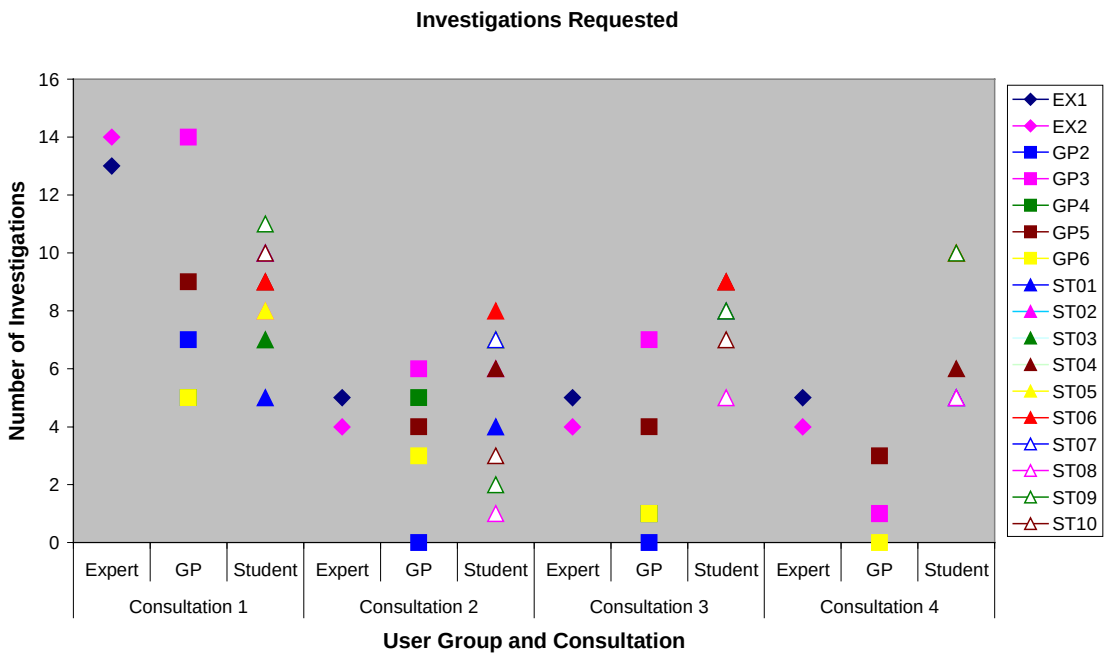

Figure 45: Number of investigations requested by each user in each consultation.

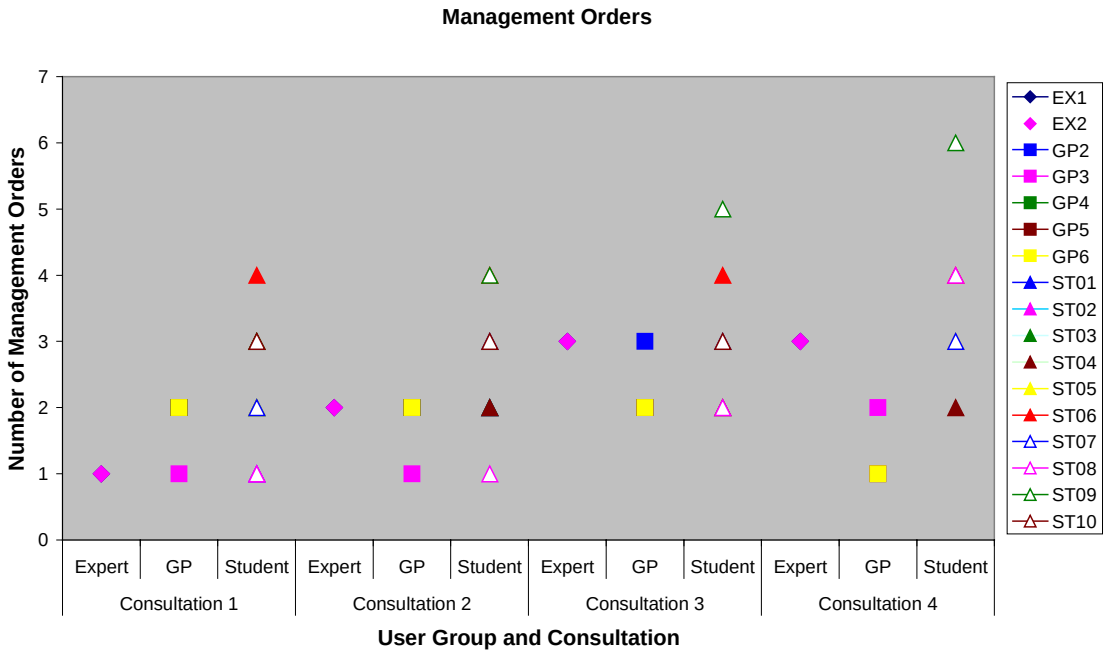

Figure 46: Number of management options chosen by each user in each consultation.

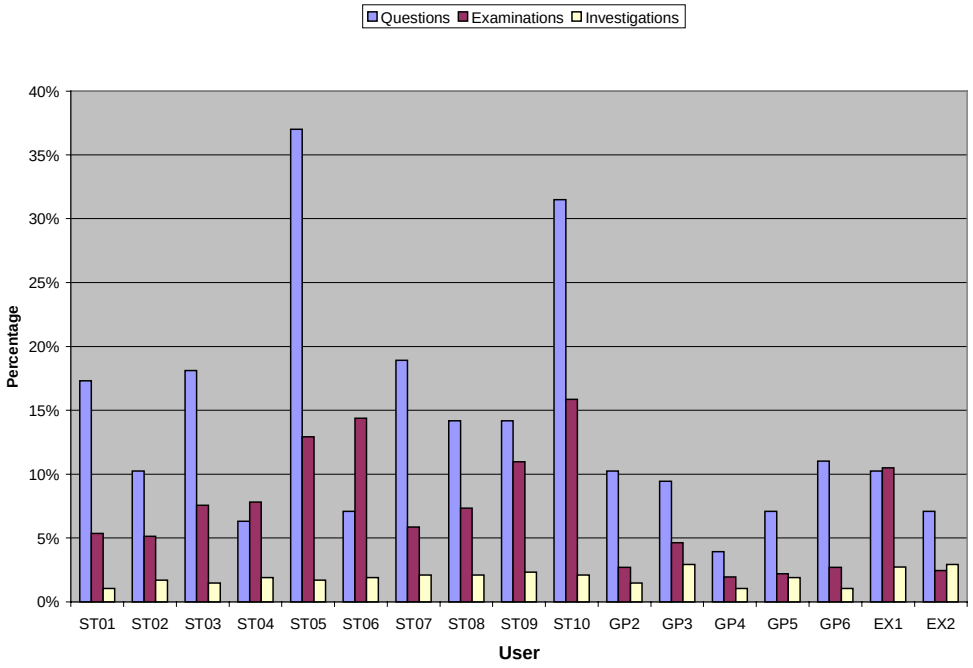

Figure 47: Number of questions asked, examinations performed, and investigations ordered in consultation one by each user, expressed as a proportion of the total number of all available questions, examinations, and investigations, respectively, held within the database.

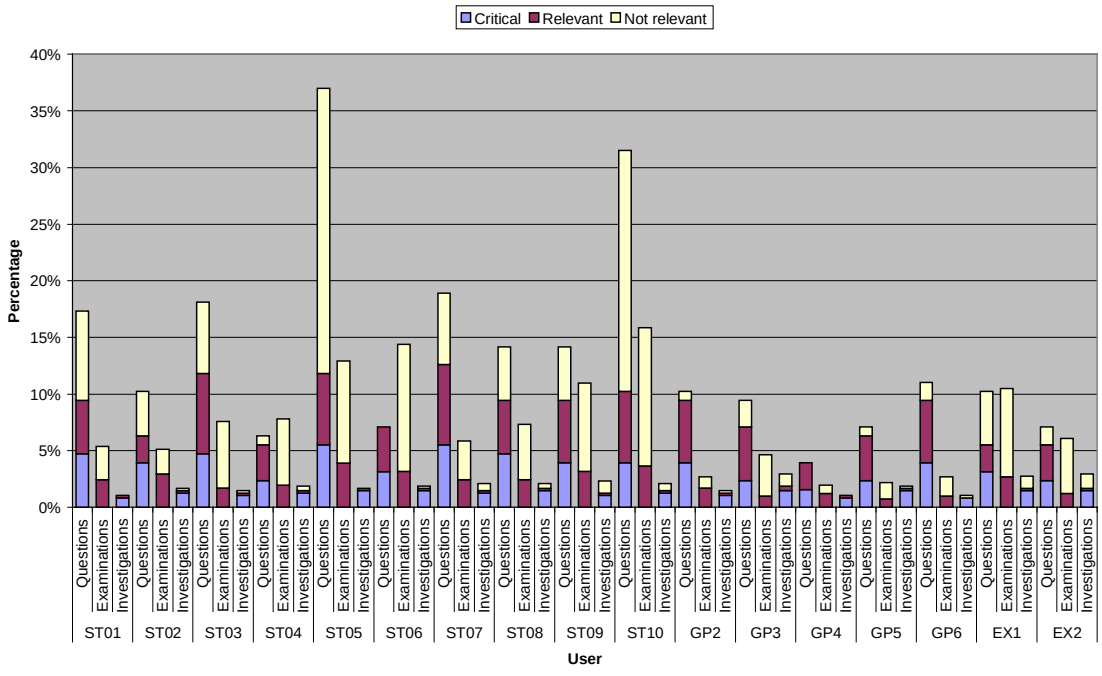

Figure 48: Number of questions asked, examinations performed, and investigations ordered in consultation one by each user, as in Figure 47, but also displaying the proportion of each action that are classified by the case author as being critical, relevant, and not relevant.

6.3.2.1 Usability and authenticity

From the questionnaire, the students and experts indicated that they found the software easy to use and navigate. In contrast, the general practitioners varied widely in their opinion. However, only one general practitioner indicated that it was not easy to learn how to use the application, or to navigate within it. Interestingly, however, other than the question and answer interface, which this individual gave a rating of 3 (on a scale of 1 to 5), this general practitioner did not give any particular interface element a low score. Also of note, was the comment by one of the specialists who stated, "I would have got stuck a few times if there were not verbal instructions." From the think-aloud observations, this comment related to the difficulty the user initially had in using the program. After they had been shown the necessary steps to continue, they were able to do so without difficulty.

Table 11: Summary of responses to the questionnaire.

Question		Student			General Practitioner			Expert		
		Mean	Range	n	Mean	Range	n	Mean	Range	n
1	I found it easy to learn how to use the application.	4.2	3-5	10	3.75	1-5	5	3.5	3-4	2
2	Navigation through the case was clear and intuitive.	4.1	3-5	10	3.25	2-5	5	3	2-4	2
3	I did not find it easy to ask the patient questions.	3.5	2-5	10	3.5	1-5	5	3	1-5	2
4	I preferred to ask questions by typing in my questions than selecting from the category lists.	3.7	2-5	10	3	2-4	5	3.5	3-4	2
5	The application was able to recognise most of the questions I asked.	3	2-4	10	3	2-5	5	3.5	3-4	2
6	I found it easy to examine the patient.	3.4	1-5	10	3	2-4	5	3.5	3-4	2
7	The specific action of each examination tool button was difficult to interpret.	2.4	1-5	10	2	1-3	5	2	1-3	2
8	It was easy to order investigations.	4.3	3-5	10	4.25	4-5	5	5	5-5	2

Question		Student			General Practitioner			Expert		
		Mean	Range	n	Mean	Range	n	Mean	Range	n
9	I found using a list box to select investigations was satisfactory.	3.8	2-5	10	4	3-5	5	4.5	4-5	2
10	It was easy to order management options.	3.8	3-5	9	2.75	1-5	5	3.5	3-4	2
11	Most of my management requests were recognised.	3.6	2-4	10	2.75	1-5	5	4	4-4	2
12	The review screen was difficult to use.	2.3	1-3	10	3.25	1-5	5	2.5	2-3	2
13	I found having to classify the importance of my questions and actions helped me to reflect on their usefulness.	3.7	2-5	10	3.5	2-5	5	4	4-4	2
14	I found having to classify the importance of my questions and actions frustrating.	3.2	2-5	10	3.25	1-5	5	3	2-4	2
15	The bar graph was easy to understand.	4.1	2-5	10	3	1-5	5	3	2-4	2
16	The bar graph provided useful information.	4.4	3-5	10	3.75	2-5	5	3	2-4	2
17	The pie graph was easy to understand.	3	1-4	6	3	3-3	2	3.5	2-5	2
18	The pie graph provided useful information.	3	1-4	6	3	2-4	2	3.5	3-4	2
19	The line graph was easy to understand.	2.6	1-5	5				2	2-2	2
20	The line graph provided useful information.	2.6	1-5	5				3	3-3	1
21	It was difficult to know what items were critical to the diagnosis or management of the patient from the graph generated in the feedback.	2.6	1-4	10	3	1-5	4	3.5	3-4	1
22	The review screens helped me reflect on the important diagnostic and management issues involved in this case.	3.8	3-5	10	3.25	2-5	5	4.5	4-5	2
23	Having multiple consultations made the case more realistic.	4.3	3-5	10	4	3-5	5	5	5-5	2
24	Having multiple consultations made the case tedious.	2.6	2-4	10	2	1-3	5	2	1-3	2
25	I have a greater understanding of the management of hypertriglyceridaemia having completed the case.	3.7	2-5	10	3.5	3-4	5	4.5	4-5	2

Question		Student			General Practitioner			Expert		
		Mean	Range	n	Mean	Range	n	Mean	Range	n
26	I have a greater understanding of the causes of hypertriglyceridaemia having completed the case.	3.5	1-5	10	3.5	3-4	5	3	2-4	2
27	The case was relevant to the understanding of hypertriglyceridaemia.	3.6	2-5	10	3.5	2-4	5	4	3-5	2
28	I found this was an interesting case.	3.9	3-5	10	4.5	4-5	5	4	3-5	2
29	I enjoyed using this virtual patient application	4.2	3-5	10	4.25	4-5	5	4	3-5	2
30	I would not use this simulation again	1.4	1-2	10	1.75	1-2	5	2	1-3	2

Question and Answer Interface

The ability of the software to understand the free text questions being asked by users was very limited. Data regarding the success rate of finding a suitable match during the first consultation for each user is summarized in Table 12. The median success rate was 50% with a minimum of 35% and a maximum of 67%. Only ST04 asked all their questions using the free text mode, and most questions that were asked by other users were selected from the various question lists. The questionnaire results reflect this finding, although there was considerable variation in the scores given by different individuals. Some indicated it was easy to ask questions, while others found it very difficult. There was no overall preference for asking questions using free text or selecting from lists of available questions. Furthermore, as illustrated in Figure 49, there was no clear association between the way people asked their questions and the difficulty they reported in asking questions. This suggests it is probably worth while to maintain both methods of interaction, at least until more advanced technologies, such as speech recognition and natural language processing have reached a more mature state.

Table 12: Number of attempts to ask questions using free text or keywords and the number of times this was successful in finding a suitable match from the corpus of questions.

User	Free text attempts	Successful matches	Percent success	Percent free text	Total questions asked
ST01	4	2	50%	8%	24
ST02	5	2	40%	15%	13
ST03	12	6	50%	26%	23
ST04	13	8	62%	100%	8
ST05	20	10	50%	21%	47
ST06	10	6	60%	67%	9
ST07			N/A	0%	24
ST08	30	14	47%	70%	20
ST09	18	10	56%	91%	11
ST10	6	3	50%	8%	39
GP2	17	9	53%	69%	13
GP3	9	6	67%	50%	12
GP4	8	3	38%	75%	4
GP5			N/A	0%	9
GP6	23	11	48%	85%	13
EX1	20	7	35%	54%	13
EX2	6	4	67%	44%	9
Median	12	6	50%		13

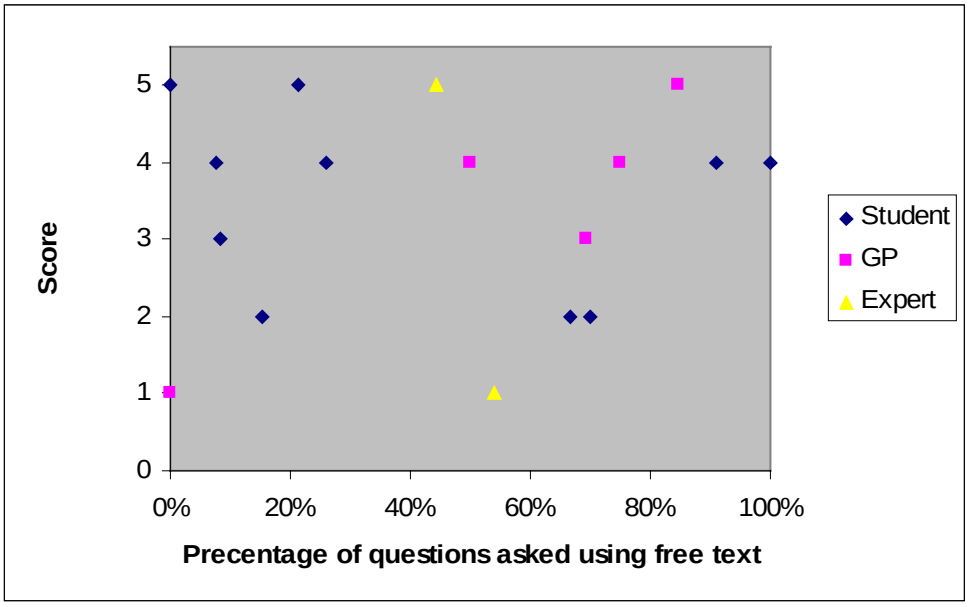

Figure 49: Percentage of questions asked using free text in contrast to selecting from a categorized list of questions versus the score for the statement, "I did not find it easy to ask the patient questions". A score of 1 strongly disagrees with the statement, while a score of 5 strongly agrees with the statement.

Physical Examination, Requesting Investigations, and Selecting Management

Most people were able to use the examination tool effectively, with one student stating, "The examination was very good and interactive, more real than I would have thought". On the other hand, there was one student and one general practitioner who did not find it easy to use. It was noted that a number of users performed a very large number of examinations (see Figure 44). However, as mentioned earlier, a number of users used the inspection tool over a large number of regions to get an overview of the patient because a full body image of the patient had not been provided.

All the students and general practitioners found it relatively easy to order investigations, with scores of 3 or above in the questionnaire. In contrast, one specialist indicated that it was difficult to order investigations. Unfortunately, no specific suggestions were given on how this might be improved.

Most of the users indicated that the ability to select management options was satisfactory. However, two of the general practitioners indicated that it was difficult to select the management they desired. The reasons for these difficulties are discussed in the following chapter in section 7.1.4.

6.3.2.2 Reflection

Figure 50 and Figure 51 show respectively, the proportion and absolute time each user spent in the reflective components during the first consultation. While GP3 spent a large proportion of his time reviewing the charts at the end of the first consultation, the absolute amount of time spent undertaking this activity was not much greater than that of other users. As previously noted, ST08 was very interesting, for while she spent little time involved in the reflective review, during the think-aloud sessions she was noted to

be much more willing to vocalize her thoughts and hypotheses while undertaking the case. This is reflected by the longer time this user spent interacting with the patient (Figure 39, Figure 50, and Figure 51).

Reviewing activity and assigning relevance scores

Figure 52 to Figure 55 show the relationship between the time over which each user reviewed the relevance of their activity, and the score they gave to the questionnaire components relating to this aspect of the review process. From Figure 52, it can be seen that most users found that having to classify their actions as critical, relevant or not relevant helped them to reflect on their activity. Only two users gave a negative response to this question. There was no obvious linear relationship between the time users spent reviewing their activity, and the degree to which they stated the activity supported reflection ($r = -0.185$, $p = 0.477$). This was also true if the result for ST05 (corresponding to the point in the upper left of the chart in Figure 52) was excluded from analysis, or if students, excluding ST05, were analysed alone. However, it is clear that the general practitioners (mean = 6.1 minutes) spent less time in this activity than the medical students (mean = 3.2 minutes) ($p = 0.002$). This was reflected in the data shown in Figure 54, with most participants believing the review screens in general, both relevance review and chart review, helped them to reflect on the important diagnostic and management issues. Again, there was no correlation between the extent to which they found the review screens helpful and the duration over which they spent reviewing their activity. Similarly, there was no clear relationship between the time users spent reviewing their activity and their stated level of frustration ($r = 0.151$, $p = 0.562$) (Figure 53). While there were some individuals who clearly found the process of reviewing their activity frustrating, this was not related to the total number of actions chosen by the user in the first consultation ($r = 0.154$, $p = 0.556$) (Figure 56). The total

number of actions includes the total number of questions asked, examinations performed, investigations ordered and management options selected in a single consultation. As illustrated in Figure 55, those individuals who found the activity most helpful (Question 13), also found the activity least frustrating (Question 14) ($r = -0.609$, $p = 0.009$). Together, these data illustrate that it is still possible to find it frustrating but helpful.

Inspection of Figure 53, Figure 54, and Figure 56 raised the possibility of a non-linear relationship between the time each user spent in the activity, or the number of actions selected, and the rating they gave for each question. In particular, the hypothesis was raised that there may possibly be some maximum value for the time each user spent reviewing the relevance scores, as well as the number of actions chosen by the user. To explore this possibility, the data for these comparisons were reanalysed after curve fitting with a sigmoid model based on Equation 1.

$$Y = a + \frac{b}{1 + e^{\frac{-(x-c)}{d}}} \qquad \textbf{Equation 1}$$

Where a, b, c, and d are curve fit parameters, and x is the question score.

After sigmoid transformation of the question scores, and calculation of Pearson's correlation coefficient, there were again, no significant correlations between firstly, the time spent reviewing the relevance of each user's actions and the stated level of frustration ($r = 0.315$, $p = 0.218$) (Figure 53), secondly, the time spent reviewing the relevance of each user's actions and the stated level of support for reflections ($r = 0.251$, $p = 0.331$) (Figure 54), and thirdly, total number of actions selected by each user and the stated level of frustration ($r = 0.382$, $p = 0.130$) (Figure 56). Note, a seven or

nine point Likert scale might have been more likely to turn up a significant non-linear correlation.

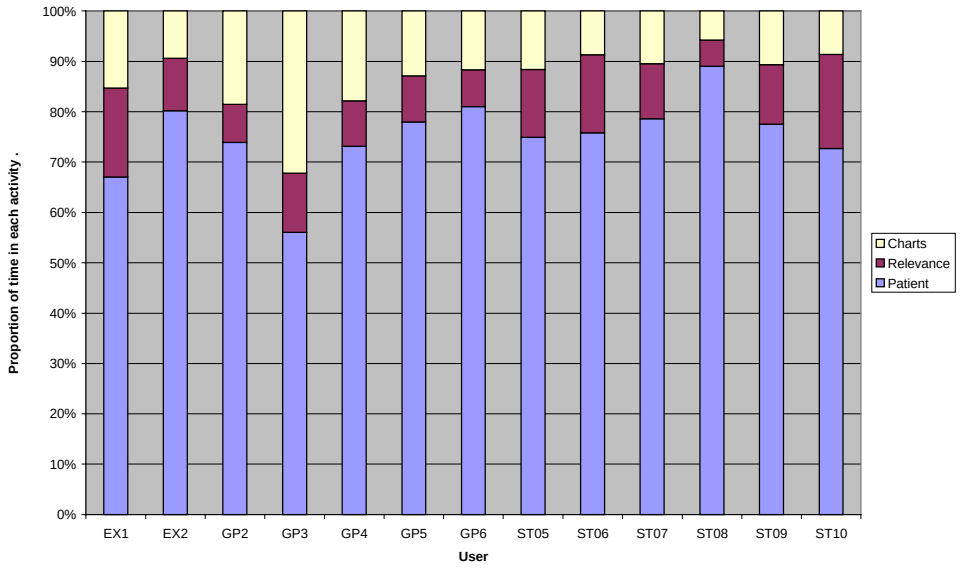

Figure 50: Proportion of time each user spent interacting with the patient, reviewing the relevance of their actions (reflection), and reviewing the charts in the first consultation. Data for ST01, ST02, ST03, and ST04 could not be included due to inadequate log data regarding time in chart review.

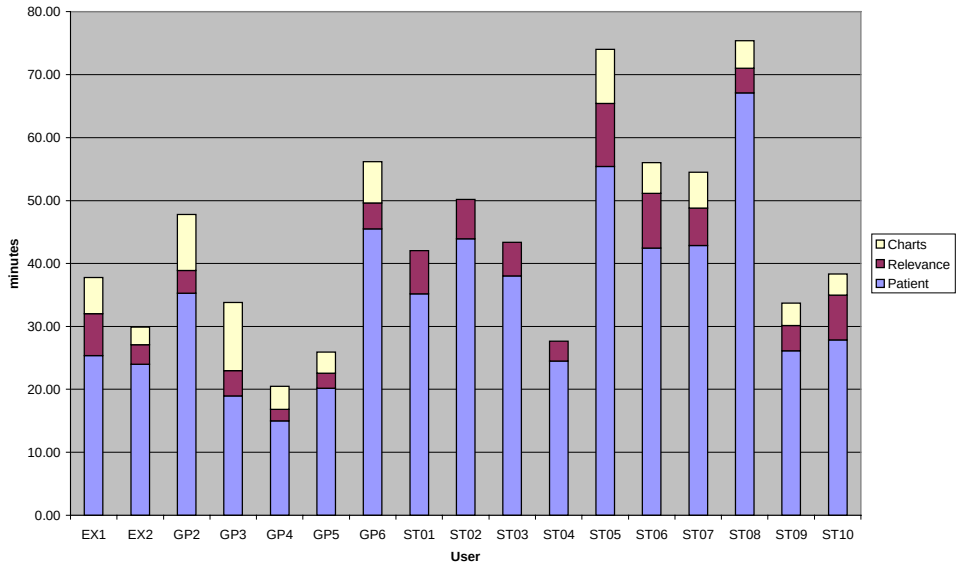

Figure 51: Total time each user spent interacting with the patient, reviewing the relevance of their actions (reflection), and reviewing the charts in the first consultation. Data on time in chart review for ST01, ST02, ST03, and ST04 could not be included due to inadequate data logging during the first four student evaluations.

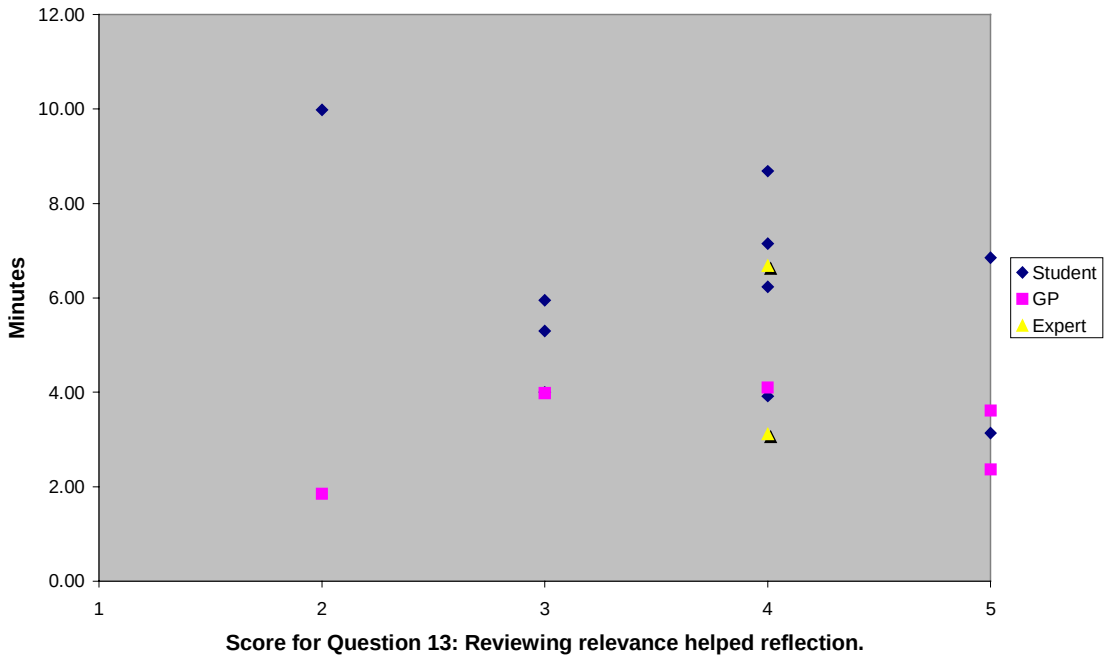

Figure 52: Relationship between the duration over which the user reviewed the relevance of their actions in consultation 1 and the score for question 13 of the questionnaire. Question 13 was, "I found having to classify the importance of my questions and actions helped me to reflect on their usefulness."

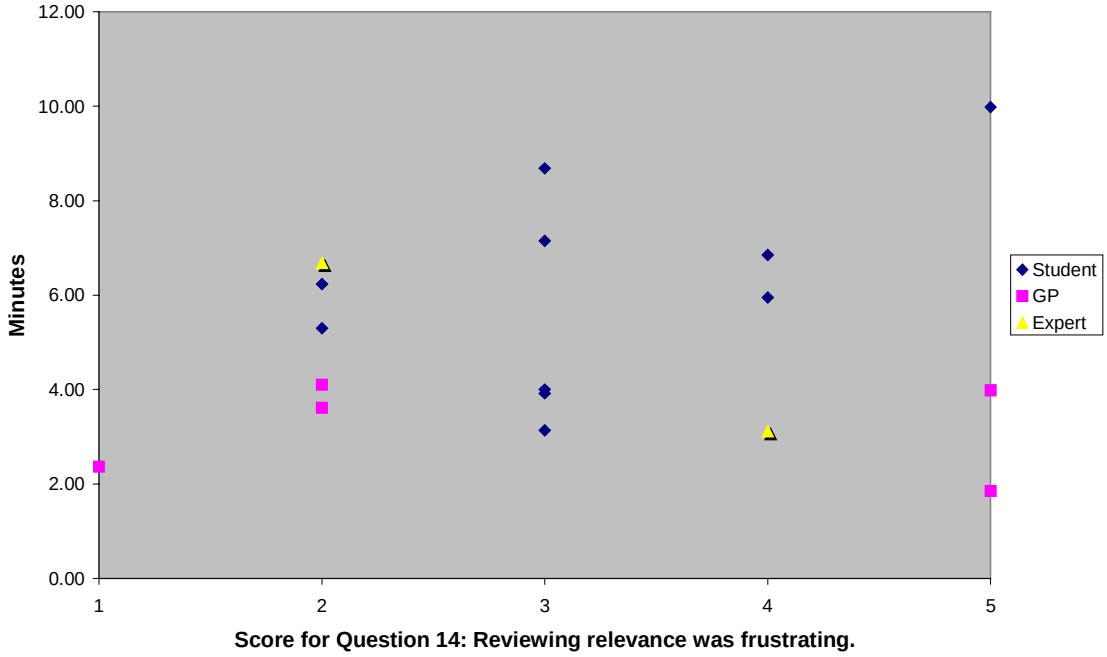

Figure 53: Relationship between the duration over which the user reviewed the relevance of their actions in consultation 1 and the score for question 14 of the questionnaire. Question 14 was, "I found having to classify the importance of my questions and actions frustrating."

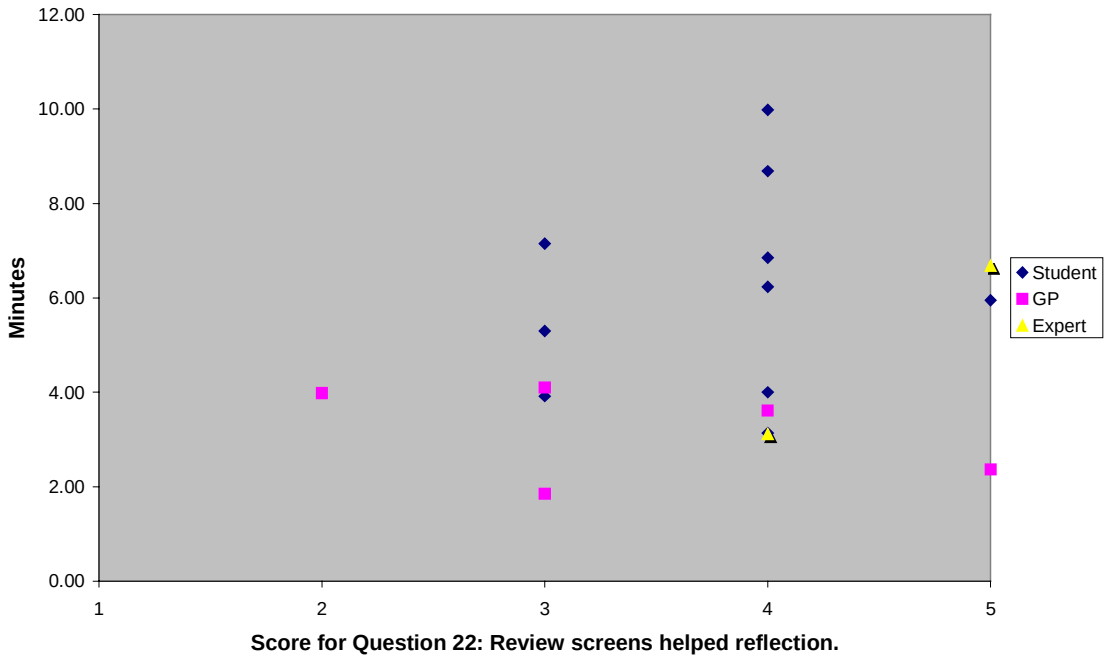

Figure 54: Relationship between the duration over which the user reviewed the relevance of their actions in consultation 1 and the score for question 22 of the questionnaire. Question 22 was, "The review screens helped me reflect on the important diagnostic and management issues involved in this case."

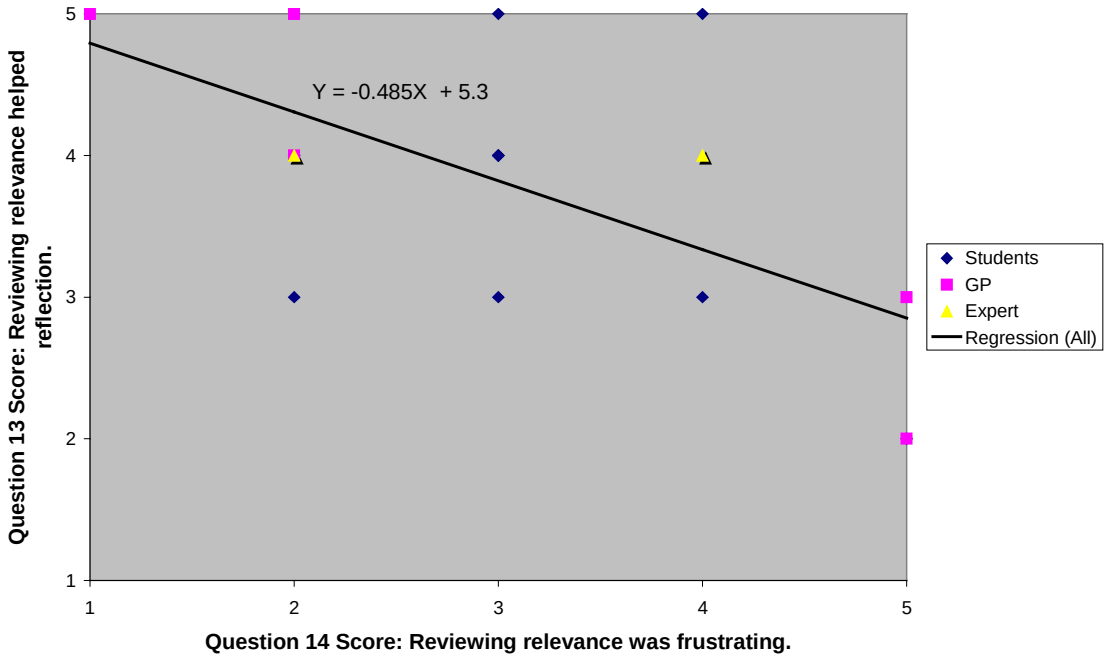

Figure 55: Relationship between Question 13, "I found having to classify the importance of my questions and actions helped me to reflect on their usefulness", and Question 14, "I found having to classify the importance of my questions and actions frustrating".

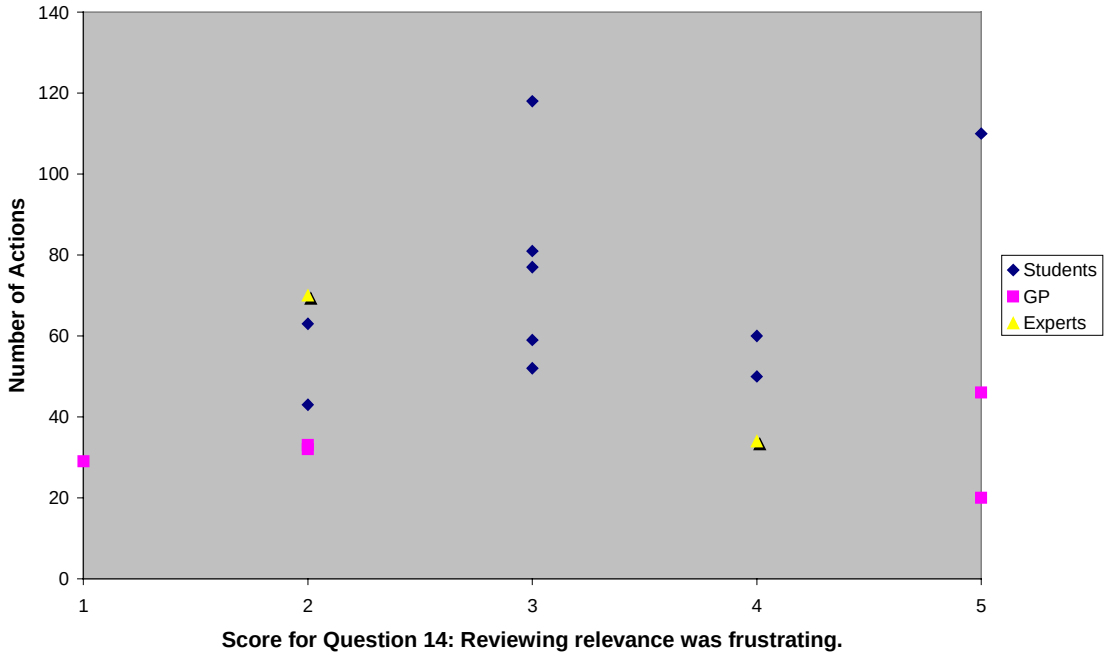

Figure 56: Relationship between the total number of questions asked, examinations performed, investigations ordered, and management orders selected in consultation one and the score for Question 14 of the questionnaire.

Chart Review

From the questionnaire data, with the exception of ST03, GP04 and EX2, users were able to use and interpret the bar chart without difficulty. However, more people expressed difficulty in determining which of their specific actions had been classified as critical, relevant, or not relevant, as well as which critical or relevant actions they had failed to perform (Figure 57). EX01 provided specific feedback on this, and suggested that this detailed information needed to be more readily accessible, rather only being visible after two mouse clicks. GP04 preferred the pie chart to the bar-chart. However, this was the reverse of the case for most other users, who found the bar-chart easier to understand. Very few people used the line-chart.

Reassuringly, there was no correlation between the duration over which the charts were reviewed in consultation one, and degree to which individuals had difficulty understanding the charts ($r = 0.120$, $p = 0.711$) (Figure 57). From Figure 58, there is also no clear association between the number of charts viewed and the difficulty which individual had in understanding the charts ($r = 0.184$, $p = 0.567$). After curve-fitting and sigmoid transformation, looking for a plateau response, there was still no significant correlation ($r = 0.350$, $p = 0.264$). While most users stated the review component helped them reflect on the important diagnostic and management issues, there was no correlation between the extent to which they held this opinion and the time spent reviewing charts ($r = -0.291$, $p = 0.336$) (Figure 59), or the number of different charts they reviewed ($r = 0.195$, $p = 0.522$) (Figure 60) at the end of consultation one.

Figure 61 compares the time each user spent reviewing the relevance of their actions with the time they spent reviewing the charts and associated information at the end of consultation one. Overall, there was no clear association between these times. However, the data suggests that within the student and general practitioner user groups, there is a linear association between the times spent in each of these activities. Statistical analysis confirmed this association (Students: $r = 0.666$, $p = 0.148$, General Practitioners: $r = 0.827$, $p = 0.084$) although, due to the small numbers in each of these subgroups, this did not reach statistical significance. Inspection of Figure 61 clearly indicates that the students and experts spent relatively more time reviewing the relevance of their actions than the general practitioners. Only six students are represented in Figure 61, as a result of being unable to estimate the chart review time for the first four students. As previously noted at the beginning of section 6.3, this arose through inadequate data logging at the time of evaluation. The students spent 6.1 minutes ($n = 10$) reviewing

their activity, while the general practitioners participated in this activity for an average of 3.2 minutes (n = 5) (p = 0.003). In contrast, the students (5.1 minutes, n = 6) and general practitioners (6.7 minutes, n = 5) reviewed the charts for similar periods of time (p = 0.37). This relationship is clearly reflected in the difference in the slope of these two regression lines. The implications of this are discussed in Chapter 7.

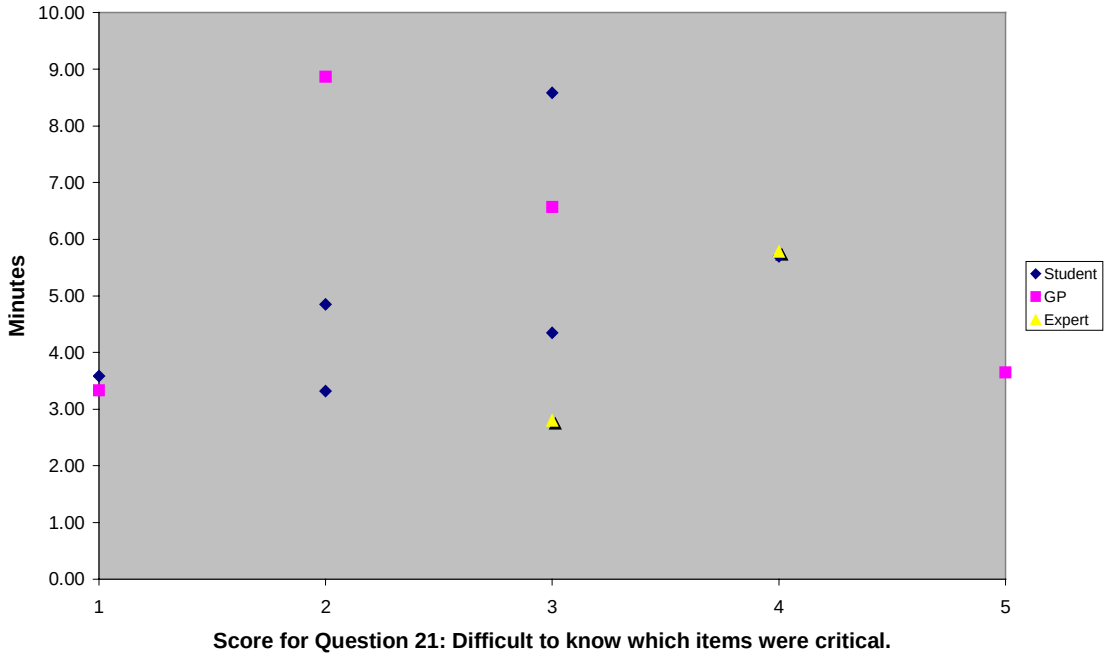

Figure 57: Relationship between the time spent reviewing the charts and the score for question 21, "It was difficult to know what items were critical to the diagnosis or management of the patient from the graph generated in the feedback".

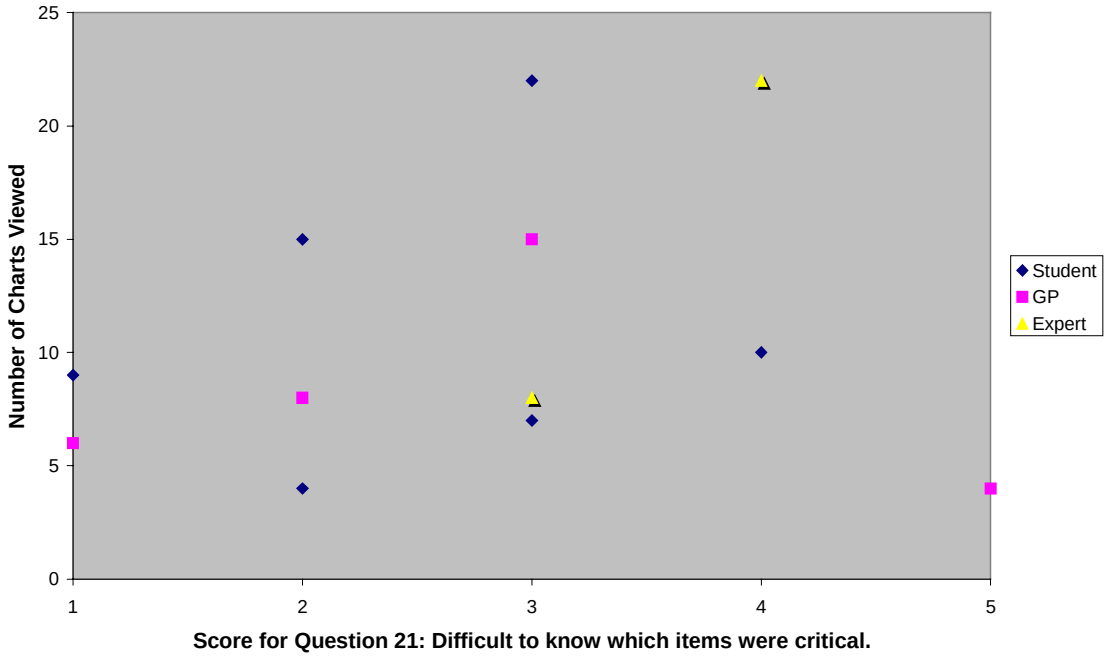

Figure 58: Relationship between the number of charts viewed and the score for Question 21

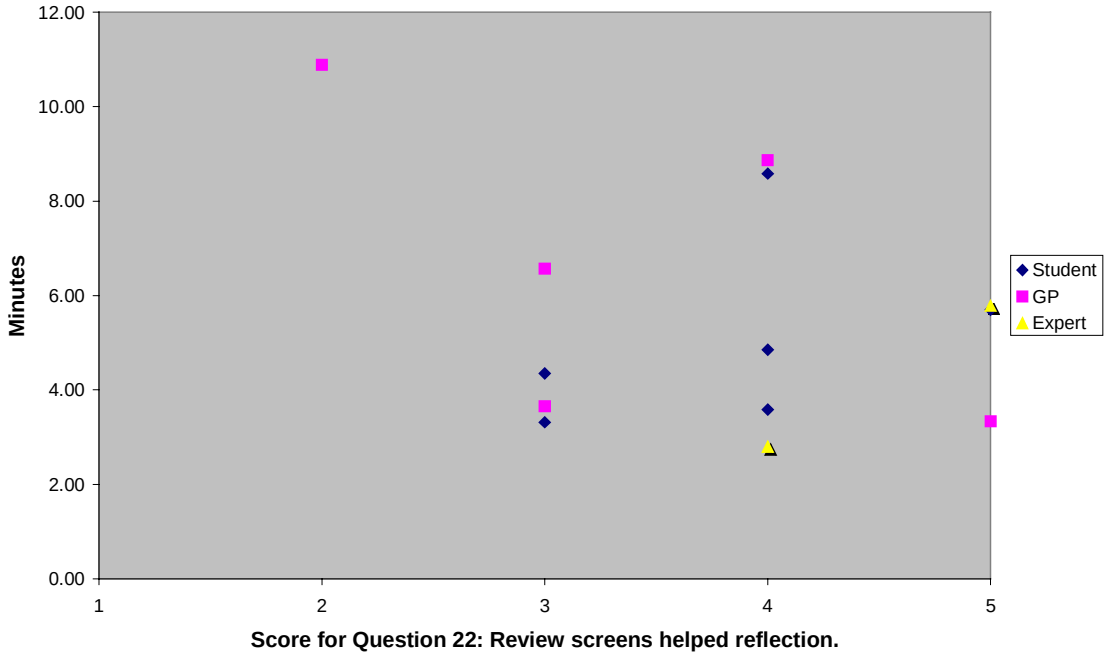

Figure 59: Relationship between the time spent reviewing the charts and the score for question 22, "The review screens helped me reflect on the important diagnostic and management issues involved in this case".

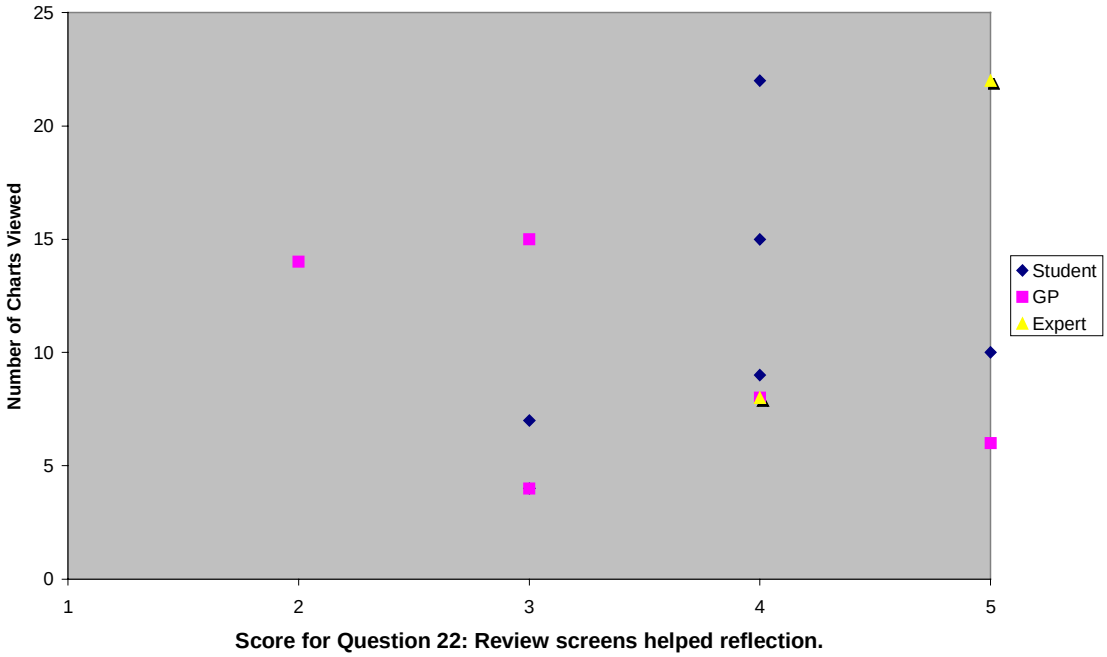

Figure 60: Relationship between the number of charts viewed and the score for question 22.

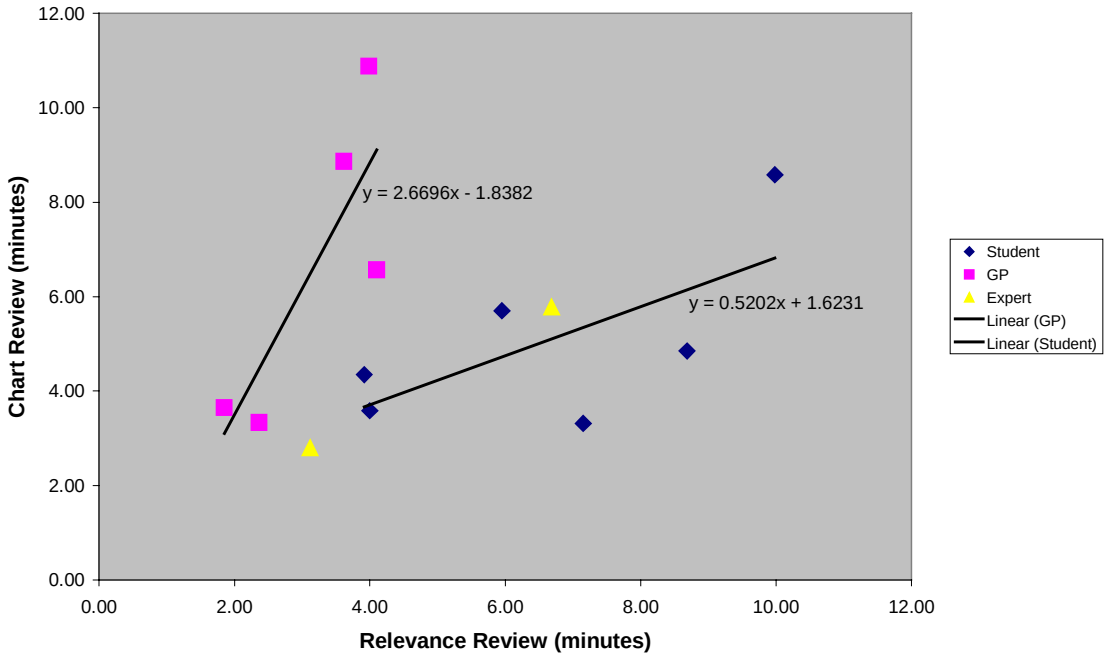

Figure 61: Time each user spent reviewing the relevance of their actions versus the time they spent reviewing the charts.

6.3.2.3 Multiple consultations

The majority of users found that having multiple consultations made the simulation more realistic and did not make it tedious. Only two medical students gave the statement that, “having multiple consultations made the simulation tedious”, a score of more than three. From Table 8, Table 9 and Table 10 it is evident that the first consultation was by far the longest for all users. Subsequent consultations required relatively quick review of the patient, and instigation of changes to management based on observed changes, or lack of changes. This may have contributed to the perception that the simulation was not overly tedious. On the other hand, there were users such as ST09 who found the case repetitive. In this instance the repetitiveness was most likely a consequence of their management strategy, which had not been anticipated when authoring the case. This issue is discussed further in the following chapter.

6.3.2.4 Understanding and Interest

With respect to understanding the causes and management of hypertriglyceridaemia, there were large variations in the scores. Some students found that the simulation had improved their understanding, while others did not find that it added to their understanding. The general practitioners gave it a more favourable rating than the students. Surprisingly, both experts said it had improved their understanding of the management of hypertriglyceridaemia, and one expert said it improved their understanding of the causes of hypertriglyceridaemia. One expert indicated that he thought the case was relevant to the understanding of hypertriglyceridaemia, while the other expert indicated that it was not so relevant, giving this question a score of 3. The general practitioners were divided in their opinion, with scores ranging from 2 to 4 with an average of 3.5. The medical students were also divided in their opinion, although only one student scored this question less than three. Having noted the above responses

to the directed questions regarding hyperlipidaemia, it is still clear, from the additional comments they provided on what they learnt (see Appendix I on page 266), that a number of medical students and general practitioners focused on the specifics of Type 1B Glycogen Storage Disease (GSD). On the other hand, ST06 was able to list a number of more generally applicable issues, more in line with the original goals of the case author. These included: “adverse effects of fibrates, Glycogen Storage Diseases, how to run a follow up consultation, and management of triglyceridaemia”.

All users indicated that they found the case interesting. Two students and one expert gave it a score of three out of five. All other users rated it more favourably. Similar results were obtained for the question regarding enjoyment when using the application. Interestingly, the people who rated the application less favourably with respect to interest, did not necessarily rate it less favourably with respect to enjoyment. With the exception of one expert, all users disagreed (score of 4) or strongly disagreed (score 5) with the statement that they, “would not use the simulation again”. This same expert had otherwise rated the simulation very favourably, so one wonders if this was an error when reading the question, which had been asked in the negative.

User “Exploration” of Key Concepts of the Case

With respect to the diagnosis and management of the patient used in evaluation of SIMPRAC, a list of the concepts to be explored by the user can be found on page 46 in Chapter 3, and is repeated here for convenience.

1. Identify that the patient had massive hypertriglyceridaemia, and to a lesser extent, hypercholesterolaemia secondary to Type 1B GSD.
2. Exclude other secondary causes of hyperlipidaemia, including excessive ethanol use, renal failure, liver disease, obesity, and diabetes.

3. Consider cardiovascular risk factors, such as smoking and hypertension.
4. Consider primary causes of hypertriglyceridaemia, as might be suggested by a family history of disease.
5. Manage the GSD with strategies to maintain blood glucose. Especially the use of low glycaemia index supplements, such as uncooked cornstarch.
6. Consider the patients diet from the point of view of GSD and hypertriglyceridaemia
7. Institute lifestyle changes such as regular exercise.
8. Commence an appropriate lipid lowering agent, with the preferred agent belonging to the fibrate therapeutic class.
9. Recognize the side-effects of the therapeutic agents, and monitor the patient accordingly.

The table at Appendix K on page 277 lists those questions and actions chosen by each user that relate to the goals listed above. This data was extracted from the log files for each user. Referring to this table it is clear that most of the key history taking questions were asked by all users. All users, except GP4, asked about the patient's past medical history and learnt that the patient had Type 1B GSD (concept 1). Although GP2 had to be prompted about this information during the think-aloud session. All users also asked about the patient's diet.

The secondary causes of hypertriglyceridaemia were addressed to a varying extent (concept 2). Interestingly, both users who were considered experts in the management of hyperlipidaemia failed to ask whether the patient smoked cigarettes (concept 3). With respect to the problem at hand this was a minor issue. However, as a whole of life issue, and these are probably better addressed by General Practitioners, it is (in the author's

opinion) of some importance. Blood pressure and the patient's weight were considered to be important as they can increase the risk of cardiovascular disease, and obesity is associated with higher triglyceride levels. As with smoking, blood pressure is less important in the acute setting. Three students and one expert failed to ask about the patient's alcohol intake. ST10 and GP2 were the only users not to assess the patient's renal function by measuring the serum creatinine, and only three students did not check the patient's liver function tests (concept 2).

Only EX1 failed to ask about a family history of disease (concept 4). This was a general question and the reason users asked the question was not known. It could have been asked to consider the possibility of a primary lipid disorder. However, the likelihood of this being the case in the context of this patient is remote. Alternatively, the question may have been asked to see if other family members had Glycogen Storage Disease. As a rare autosomal recessive disorder, the chances of this being the case, except for siblings of the patient, would be remote.

Six students failed to measure the triglyceride concentration in the first consultation, despite its being the primary reason the patient presented to them (concept 1). This may have been because they failed to realise the importance of confirming a triglyceride level, measured with a point-of-care device, with more accurate and precise laboratory methods before instituting treatment. Two students, ST05 and ST10 failed to measure the triglyceride at all. In the case of ST05, this was because he was more focussed on lowering the cholesterol concentration, as evidenced by his use of an HMG CoA Reductase Inhibitor (statin). In the latter student's case, he treated the triglyceride with a

fibrate without first confirming the hypertriglyceridaemia, and without further monitoring of response to treatment.

Although not flagged as a relevant or critical investigation by the case author, both specialists requested thyroid function tests including thyroid-stimulating hormone (TSH). This is an interesting observation, as hypothyroidism is a possible cause of hyperlipidaemia (concept 2). However, it tends to elevate the cholesterol concentration in serum more than the triglyceride concentration. Nevertheless, it is frequently part of investigation protocols within clinics devoted to the management of hyperlipidaemia (personal experience). The fact that both specialists requested this test may be an indication of a scripted approach to the management of these disorders. Alternatively, it may be that they considered that it was important to exclude this as a co-morbidity that would accentuate the lipid abnormalities seen in GSD.

Seven students, two general practitioners, and both specialists advised the patient to have a diet low in saturated fat (concepts 6 and 7). During the think-aloud session, GP6 indicated she wanted the patient to consider dietary choices with a low glycaemic index, as an alternative to cornstarch for maintaining the blood glucose level. Only four students and EX1 suggested to the patient she should consider lifestyle changes such as exercise to reduce their lipid levels (concept 7).

With respect to the patient's management, seven students and two general practitioners used uncooked cornstarch, as recommended by the information sheet (concept 5). Neither of the specialists used this treatment modality. Six students used a statin as the primary form of treatment for the patients lipid disorder, suggesting that they were

focussing on the moderately elevated cholesterol rather than massively elevated triglycerides. Three students, all the general practitioners, and both specialists used a fibrate as the primary treatment for the patient's lipid disorder (concept 8). One student initially used a statin and then changed to a fibrate, while EX2 started with a fibrate and added a statin.

Gemfibrozil was, until recently, the only fibrate routinely used in Australia. Statins, as a class of drugs, including agents such as Simvastatin, Pravastatin, or Atorvastatin, can cause muscle pain (myalgia), muscle inflammation (myositis), or more rarely, muscle breakdown (rhabdomyolysis). The severity of the inflammation or injury can be estimated by measuring the creatine kinase activity in serum or plasma. Of the 16 users who used a fibrate or a statin to treat the patient, only five students checked the creatine kinase activity (concept 9). In the case of ST06 this was done after first seeing that this was marked by the case author as being a relevant investigation, and then learning subsequently that myalgia and myositis were side-effects of Gemfibrozil, as well as statins. Neither of the specialists checked the activity of this enzyme. Other than being unaware of the value in performing this test, the users may have considered that without a history of myalgia, the chances of their being any significant muscle injury would be low. On the other hand, some individuals can have drug induced elevation of creatine kinase without myalgia.

This chapter has reported the qualitative and quantitative results arising from the end-user evaluations. The quantitative results have been considered in terms of usability and authenticity, support for reflection, the use of multiple consultations, and evidence that the users have been aware of the core concepts in diagnosing and managing the patient.

The next chapter discusses these results in detail, with a focus on the core concerns of the thesis.