



## **Demetrios Prize 2019: Machine Learning as an Adjunct to Medical Decision Making**

Authors: Nicolas Woitzik  
Submitted: 14. July 2019  
Published: 16. July 2019  
Volume: 6  
Issue: 7  
Affiliation: Medizinische Fakultät, Albert - Ludwigs Universität Freiburg im Breisgau, Germany  
Languages: English  
Keywords: Demetrios Prize 2019, Winner, Medicine,  
Categories: Demetrios Project, Medicine, Life Sciences  
DOI: 10.17160/josha.6.7.586

### Abstract:

One of the winners, Nicolas Woitzik from Germany, evaluated in his thesis the textual clinical recommender system which represents a computerized clinical decision support system (CDSS). It analyses the patient's discharge summaries with the help of information retrieval and natural language processing methods. It provides the user with a similar patient case out of a database to include this information into the user's decision-making process. We conducted an experiment to validate the correlation between the computed similarities by the new CDSS and the similarity judgment of medical experts, junior doctors, and medical students. Taken together, the retrieval system still needs improvements, either based on an improved retrieval algorithm or by additional features. However, it is likely that the performance of the system will improve the more discharge summaries a database contains like it was shown in this thesis. Our data suggest that the simrec software might indeed become an important clinical tool to share clinical experience between hematologists and possibly also other medical specialties.

# JOSHA

[josha.org](http://josha.org)

**Journal of Science,  
Humanities and Arts**

JOSHA is a service that helps scholars, researchers, and students discover, use, and build upon a wide range of content

Aus dem Department Innere Medizin

Klinik für Innere Medizin I

Schwerpunkt: Hämatologie, Onkologie und Stammzelltransplantation

des Universitätsklinikums Freiburg im Breisgau

# **Machine Learning as an Adjunct to Medical Decision Making**

INAUGURAL - DISSERTATION

zur Erlangung des Medizinischen Doktorgrades

der Medizinischen Fakultät

der Albert – Ludwigs – Universität Freiburg im Breisgau

Vorgelegt 2019

von Nicolas Frank Philipp Woitzik

geboren in Freiburg im Breisgau



Dekan (kommissarisch): Prof. Dr. Norbert Südkamp

1. Gutachter: PD Dr. Reinhard Marks<sup>[1]</sup><sub>[SEP]</sub>

2. Gutachter: Prof. Dr. Frank Jäkel

Jahr der Promotion: 2019

*- to my family -*



## 1 Table of Content

<b>1 Table of Content .....</b>	<b>3</b>
<b>2 Introduction.....</b>	<b>5</b>
2.1 Preface .....	5
2.2 Aims and Objectives .....	8
<b>3 Materials and Methods .....</b>	<b>9</b>
<b>3.1 Patient Similarity - A Textual Recommender System for Clinical Data.....</b>	<b>9</b>
3.1.1 The Program – Clinicon SimRec.....	14
<b>3.2 The Similarity Experiment - Second Evaluation of the System .....</b>	<b>15</b>
3.2.1 Dataset of 489 CLL Patients Discharge Summaries – Structure, Content and Anonymization .....	15
3.2.2 Participants – Selection and Anonymization .....	16
3.2.3 Questionnaire about CLL and other Questions .....	17
3.2.4 Structure of the Experiment .....	19
3.2.5 Reference Letters - Selection Criteria .....	20
3.2.6 Procedure .....	22
<b>3.3 Biometric Aspects .....</b>	<b>22</b>
3.3.1 Hypothesis and Null Hypothesis .....	22
3.3.2 Statistics .....	24
<b>3.4 Literature.....</b>	<b>24</b>
<b>3.5 Software .....</b>	<b>24</b>
<b>4 Results.....</b>	<b>25</b>
<b>4.1 Data Analysis .....</b>	<b>25</b>
4.1.1 Data Computation for the Experiment .....	25
4.1.2 Participants Characteristics.....	26
4.1.3 Rating Analysis .....	26
4.1.4 Repeatability .....	28
4.1.5 Inter-rater Agreement.....	29
4.1.6 Testing against Chance .....	31
4.1.7 Correlation between Experts and the Program.....	32
4.1.8 Differences between Selected and Randomly Chosen Trials .....	34
4.1.9 Striking Trials .....	35



4.1.10 Individual Participant Correlation.....	36
4.1.11 Correlation of all Participants .....	37
4.1.12 Differences between Experts, Junior Doctors and Students.....	38
4.1.13 Similar or Not – Assessing Recommendation Quality .....	42
<b>4.2 Explorative Analysis.....</b>	<b>47</b>
4.2.1 Practical Usage.....	48
4.2.2 Usage for Medical Questions .....	49
4.2.3 Categories for Similarity Judgement .....	50
4.2.4 Potential for Improvement.....	52
<b>5 Discussion .....</b>	<b>53</b>
<b>6 Summary .....</b>	<b>58</b>
<b>7 Zusammenfassung.....</b>	<b>59</b>
<b>8 References .....</b>	<b>60</b>
<b>9 Table of Figures.....</b>	<b>62</b>
<b>10 Table of Tables .....</b>	<b>65</b>
<b>11 Acknowledgement.....</b>	<b>66</b>
<b>12 Curriculum Vitae.....</b>	<b>66</b>
<b>13 Appendix .....</b>	<b>67</b>
13.1 Glossary .....	67
13.2 Example Letters, Striking Trials .....	67
13.3 Declaration of Consent, Multiple Choice Test, Questionnaire.....	70
<b>14 Eidestattliche Versicherung .....</b>	<b>77</b>



## **2 Introduction**

### **2.1 Preface**

Technological advances and growing access to computer systems drive many health care innovations. In 2009, the United States authorized the Health Information Technology for Economic and Clinical Health (HITECH) Act. It aims to create a 21<sup>st</sup> century health care information system. One important step to achieve this goal is the expansion and adoption of electronic health records (Blumenthal 2010). These records consist of different patient characteristics, for example: diagnostic tests, like blood tests, biological information as well as social information and many other characteristics. Together, they build a specific pattern, unique for each patient. Computerized clinical decision support systems (CDSSs) use this electronic information to evolve recommendations for the health care staff. In a systematic review, Garg et al. (2005) investigated different CDSSs and categorized them into systems either for diagnosis, disease management, drug management, as well as reminder systems for prevention. Although many CCDSs have shown to improve practitioners' performance, outcome effects remain unstudied or inconsistent (Garg et al. 2005). In order to understand how CCDSs can improve clinical decision-making, it is important to understand how doctors approach a clinical decision. Medical decision-making research has a long-standing tradition and has developed over the last decades (Elstein, Shulman, and Sprafka 1990). Today, a dual-process model is the dominant universal model of decision-making and contains several different theories of decision-making (see Fig. 1.1), more specifically two different modes (Kahneman 2011; Evans 2008; Croskerry 2013).

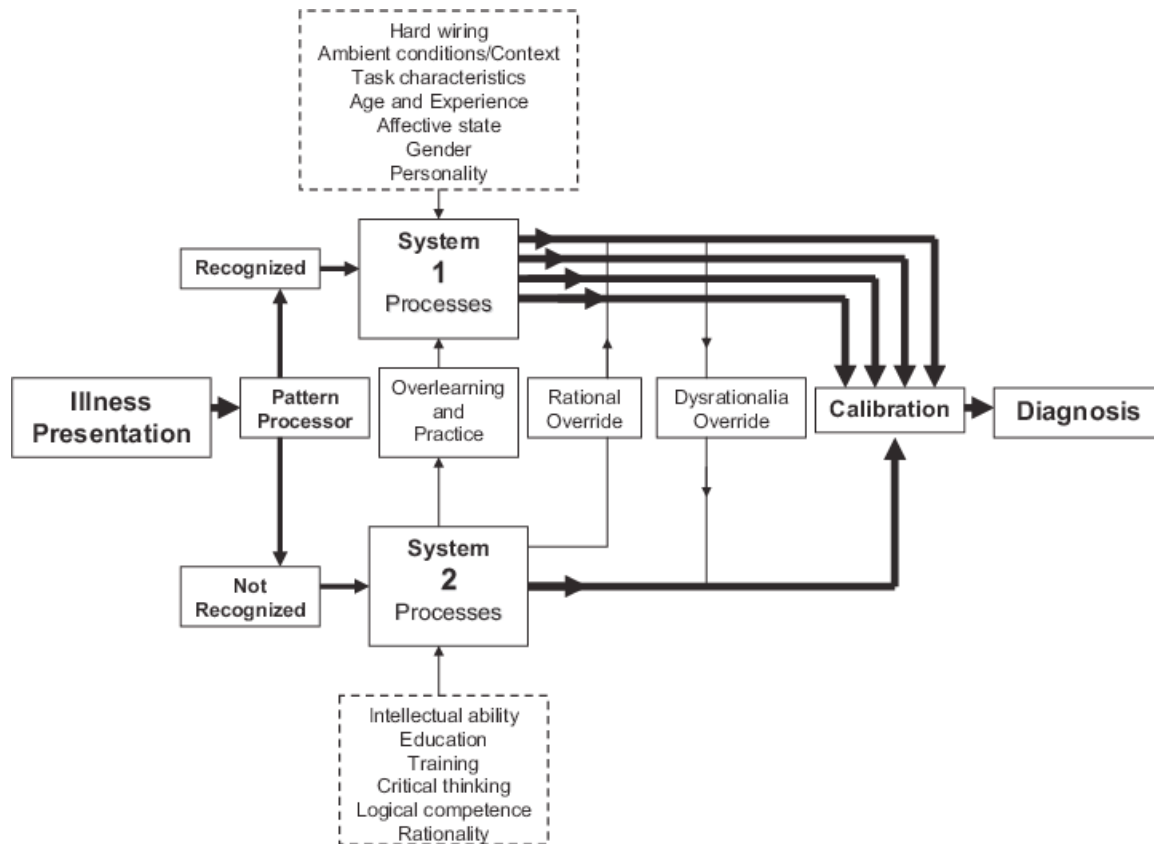


Fig 1.1: Medical decision-making model adapted from Crosskerry 2009

The first mode is called “System 1” and represents a non-analytical mode. It is intuitive, fast and performs similar to a reflex system without much effort. Furthermore, it works automatically. System 1 mode lends itself to common cases or in situations, which demand fast acting like in an emergency. Its quality depends on the clinical repertoire of different patterns and the physician's experience. Pattern recognition takes place at the very beginning of the medical decision-making process and therefore plays a central role. In case the patient's clinical pattern is not recognized the analytical thinking mode, called System 2 mode, is used. It is slow, conscious and usually effective and thus it can override System 1. Both systems can interact with each other, which is symbolized by the broken lines. For example, a repeated problem can become familiar and with enough practice System 1 can be applied instead of System 2.

A key component of System 1 is “experience”, while system 2 reflects “knowledge”. While it is relatively easy to employ computer-assisted searches in databases and in the relevant literature, there is no standard approach as yet to also take doctor and



patient-specific information obtained over a long course of a disease into account. The system of e.g. Tumor Boards, second opinions etc. are the most frequently used processes to share “experience” in addition to knowledge. In view of the increasing recognition of the importance of “precision medicine” or “personalized medicine” patient cohorts are subdivided into small and smaller subgroups. Thus “regular” patients become similar to patients with “rare diseases”, where classical clinical trials rarely exist or are even impossible to perform because of small numbers. Recognizing this challenge, case series might be a more practical way to explore diagnostic procedures, treatment responses, side-effects etc.. Another approach to tackle this challenge would be to screen unstructured medical records for patients similar to the one in front of us. We hypothesize that this approach, to use a text-mining software to screen large numbers of records for similarities, might indeed allow any physician to share the clinical experience of many other colleagues, ideally and eventually from many different medical centers.

In order to address this question we developed a novel Computerized clinical decision support system, that we termed Clinicon-SimRec.

This new CDSS uses the idea that the repeated confrontation with similar patient cases forms a prototype pattern. An algorithm provides examples of similar patient cases, which are normally retrieved from memory by the practitioner. Discharge summaries are very qualified for representing clinical cases, because they contain most of the important information of a given patient in a condensed form. The prototypical program uses information retrieval methods to retrieve similar patient’s discharge summaries out of a clinical database. The system described here was developed and tested with a dataset of 307 oncology cases including patients suffering from different cancer entities. Oncology experts confirmed the quality of the system with an experimental evaluation tool (Hummel et al. 2018). Further validation is needed as it is unclear whether the system can assign similar cases from a larger, more homogeneous database in a way experts would agree on. Therefore, we created a new dataset of 489 patients, all having a diagnosis of chronic lymphatic leukaemia (CLL) in common. CLL is the most common chronic lymphoproliferative disorder in the western world (Dores et al. 2007). Over the last decades, an increasingly complete picture of the genetic landscape, molecular mechanisms and the CLL genome emerged. These data





are crucial for understanding the disease pathogenesis and consequently to transfer the knowledge into new treatment strategies and therapeutic approaches (Fabbri and Dalla-Favera 2016). CLL patients show a high degree of biomolecular heterogeneity. As a consequence, CLL patients present a highly variable clinical course and the prognosis of each patient remains difficult to predict (van Oers 2016). CLL therapy has undergone a rapid evolution, especially in the last two decades due to the introduction of the first CD20 antibody (*rituximab*) and the application of a new generation antibodies as well as targeted agents, acting on the B-cell receptor signalling pathway, like *ibrutinib* (Hallek 2017). Taken together, CLL is a heterogeneous disease with a variable clinical course. Therefore, we are expecting CLL patient's cases to be suited for testing the clinical recommender system.

## 2.2 Aims and Objectives

We conducted an experiment to validate the correlation between the computed similarities by the new CDSS and the similarity judgement of medical experts, junior doctors and medical students. We hypothesized that experts rate the similarity between patients or rather their discharge summaries in a way that correlates with the computed similarity of the system. Experts and novices are usually using different strategies for clinical decision making and there is an evolution of clinician's diagnostic reasoning (Thammasitboon and Cutrer 2013). For this reason, we expect experts to assign different letters as similar to a reference letter than novices do, especially in such a heterogeneous disease like CLL. To testify our assumption, we asked medical oncology novices (junior doctors and medical students) to judge the similarity of several letter pairs. Therefore, expert's correlation should be higher than junior doctor's correlation, which again should be higher than student's correlation in this system.

Adaption, implementation, as well as user acceptance and clinical effectiveness are important issues for a new CDSS (Garg et al. 2005). As it is unclear how the program would act in a physician's daily working life, we also investigated some possibilities for integration. Additionally, we asked for which medical questions such a program could be used and we collected ideas for improvements from the different user groups. In addition, it is unclear which criteria participants prioritize in their judgement of similarity



between different patient's discharge summaries. We therefore collected feedback information from the study participants with a questionnaire in order to evaluate aspects for further improvement of the recommender system.

## **3 Materials and Methods**

### **3.1 Patient Similarity - A Textual Recommender System for Clinical Data**

A German IT company PSIORI GmbH, Freiburg, Germany, developed the CDSS evaluated in the current study in the framework of a Bachelor Thesis by Philipp Andreas Hummel from the Institute of Cognitive Science, University of Osnabrück, in cooperation with the oncology department of the University of Freiburg. The original thesis name is "A Textual Recommender System And Other Text Mining Applications For Clinical Data". A corresponding paper was published on the International Conference on Case-based Reasoning (Hummel et al. 2018). The following abstract gives an overview of the structure and function of the system.

#### *Dataset Processing and Similarity Measure*

The textual recommender system uses Natural Language Processing (NLP) and Information Retrieval methods to analyse patient's discharge summaries. It applies the so-called "The bag of words (BoW) model", which represents a text as a "bag" of words, disregarding grammar or word order, only using the information about how often a word or a term is present in a text (Manning, Raghavan, and Schütze 2008). This information is transferred into mathematical vectors, which can be used for text categorization. Before the BoW model can be applied to a text, the document passes through several processing steps, such as tokenization (extracting words out of a sequence), stemming ("patients" and "patient" have a common stem "patient") or removal of stop words (frequently appearing but uninformative words, like "a", "the", "for").



Examples of the BoW model:

Example 1

Document d1: "The patient with disease A."

Document d2: "The patients with disease B."

After processing, the text is represented as vectors in the BoW model:

$$\begin{array}{r}
 v_{d1} = \begin{array}{l} 1 \\ 1 \\ 1 \\ 0 \end{array} \quad v_{d2} = \begin{array}{l} 1 \\ 1 \\ 0 \\ 1 \end{array} \quad \begin{array}{l} \textit{patient} \\ \textit{disease} \\ A \\ B \end{array}
 \end{array}$$

Due to tokenization the "." has been removed from the text. Stop words like "the" and "with" are not taken into account and the word "patients" is represented as "patient" as a result of stemming.

The BoW model has several limitations. As shown above, it requires several processing steps to represent text in the BoW model. During this procedure, information gets lost. One problem is presented in Example 2:

Example 2

Document d3: "John is quicker than Mary"

Document d4: "Mary is quicker than John"

Based on the fact, that information is seen as the number of occurrences of each term, both documents are represented identically in the BoW model, although they have a completely different meaning. However, the two documents are still similar in content. In the BoW model all terms are weighted the same way, but not all words in a document are equally important to address the question of similarity and relevancy, even after removing stop words. Therefore, additional vector space models exist to address this problem. The term frequency - inverse document frequency (tf-idf) can be used for this



issue and is presented in more detail below, as it is the model of choice for the recommender system.

Duplicates and “follow-up” letters are irrelevant for the retrieval process. Thus, they can be sorted out by using their vector proximity from the BoW model. As similarity measure, the cosine similarity is used. The cosine similarity compensates the effect of different document length and is a well-established method to quantify the similarity between documents and their vector representations (Manning, Raghavan, and Schütze 2008). Despite the above-mentioned limitations, the BoW model is a useful model in practice for the retrieving task and a good tool to identify duplicate and “follow-up” letters.

#### *Term Frequency – Inverse Document Frequency (TF-IDF)*

The term frequency – inverse document frequency model (tf-idf) uses a specific scaling scheme, giving rare words more influence for the discrimination between texts than frequently used words (Manning, Raghavan, and Schütze 2008). Term frequency describes the number of occurrences of a term in a document. The document frequency is defined as the number of documents containing a term. Consequently, the inverse document frequency of a rare term in a set of documents is high. The tf-idf combines the two definitions. Certain terms have less discriminating power than others, for instance, words which often appear in a dataset of different documents. As an example, in our dataset the term “patient” would have a very high document frequency, hence, it is not eligible as a good discriminative feature. Words, which occur many times (high term frequency) in a small number of documents, (high inverse document frequency) are most appropriate for the discrimination task.

#### *The Recommender System*

The algorithm computes, based on a reference letter, the most similar letter from the database. It creates a ranking of all other letters and the most fitting letter, according to the algorithm, is ranked as number one. The similarity was measured with the above-introduced cosine similarity of corresponding vectors. During the program's development different text embedding methods have been tested. A medical expert



supervised the similarity information. He categorized discharge summaries into 50 non-overlapping groups of similar patients. This information was used to adjust the algorithm. All methods worked high above chance level whereas the tf-idf method showed the best performance and is the embedding method of choice for the recommender system.

### *Evaluation and Results*

To verify the recommendation quality, an experimental setup was created. With the aid of a psychological experiment, the quality of the program was tested by a group of medical experts. Four oncology experts (at least 5 years of medical practice) and two advanced medical students participated. They were asked to rate the similarity between a “reference letter” and five other letters (“comparison letter”) from the dataset. 32 reference letters were selected. According to the computed similarity, the algorithm chose four out of five comparison letters. The last comparison letter was randomly selected. The participants gave a rating in the range of 1 (very dissimilar) to 7 (very similar) for each letter pair. The participants could choose their criteria for the rating task themselves. The inter-rater agreement was higher among experts (expert agreement: 0.76; student agreement: 0.59). Student-rating data was discarded for analysis. Follow-up pairs were also excluded because participants rate the similarity as expected as very high. Results are illustrated in Figure 3.1.1.

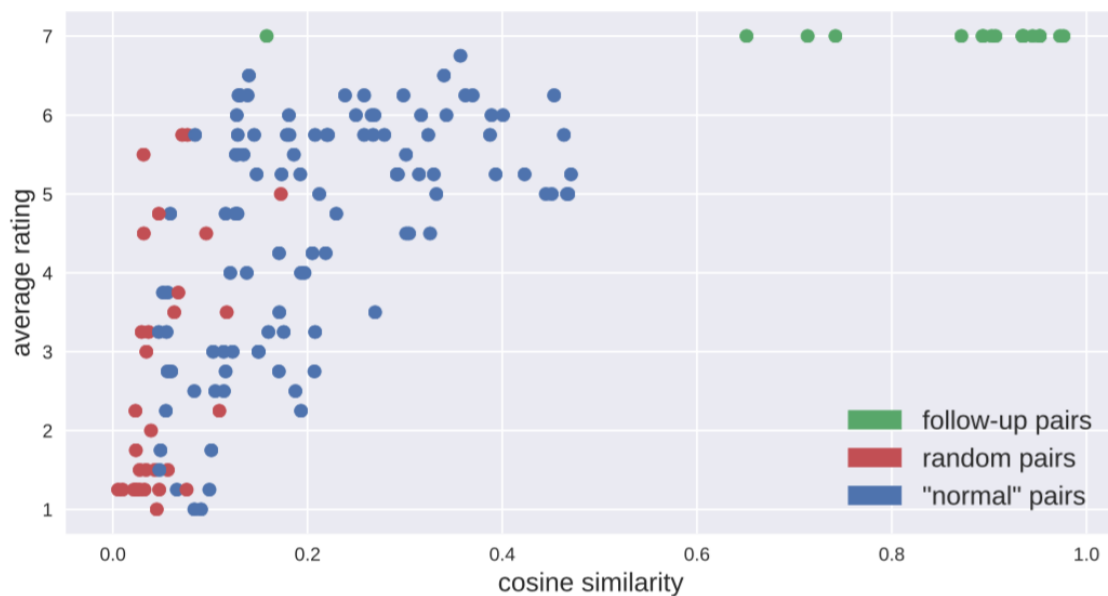


Figure 3.1.1: Cosine similarity vs. average rating of medical experts (Hummel et al. 2018)

Having a high cosine similarity, follow-up pairs received an average rating of 7 by all experts. Therefore, the recommender system can distinguish follow-up letters from non-follow-up pairs, due to their high cosine similarity. Relatively high cosine similarities go along with high expert's ratings. However, some discharge summaries with low cosine similarity are suitable for the retrieval task, whereas others are not. The letter can be seen as "false-positive" results. This distribution displays the difficulty to define a lower limit for the retrieval task.

Findings correlate with oncology expert's similarity judgement better than chance (Spearman coefficient of 0). According to Hummel et al. the correlation between the ranking given by the averaged expert rating and the ranking of the algorithm is 0.39 (95% CI [0.22,0.56]). The system's ranking is better than chance, however, compared to the inter-expert agreement of 0.71 (95% CI: [0.63, 0.79]) it is still below expert accordance.

### *Shortcomings*

The system was tested on a database with 307 oncology discharge summaries with different cancer entities. It is unclear if the system finds usable similarities in a more



homologous dataset. Due to the fact that it is the physician’s choice how they write the document and the circumstance, that other clinics have their own way to build up a patient letter, the system’s recommendation can vary. It remains to be determined if the system’s application can improve medical decision-making or the patient’s outcome.

### 3.1.1 The Program – Clinicon SimRec

Based on the above-described system, PSIORI developed a beta version of the program. The user interface is shown in Figure 3.1.1.1

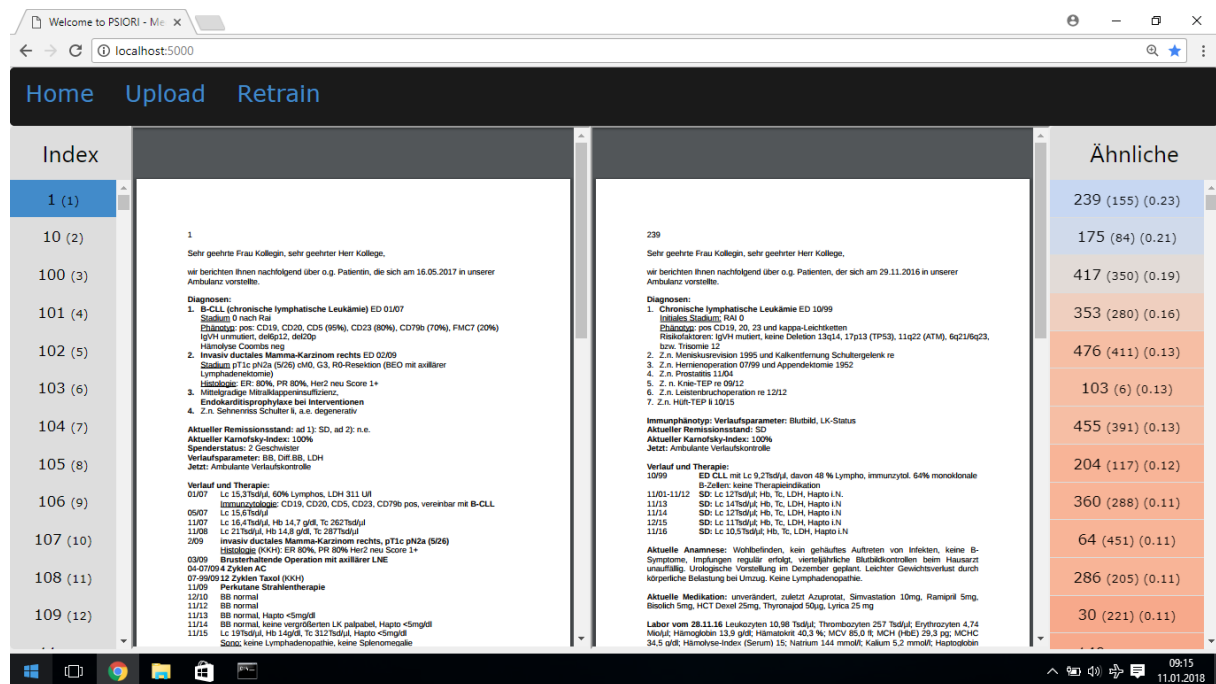


Figure 3.1.1.1: User interface of the program, “Home” view

On the left hand side of the “home” view the uploaded letters are listed (Index). The letter of a given patient is shown on the left, while there is a list of all letters, sorted by similarity towards the according Index letter, the most similar being first. The value in brackets is the computed cosine similarity. The coloured background illustrates the computed similarity. Blue coloured letters are more similar than red ones, according to the algorithm. The second tab “Upload” allows the user to upload new documents (only docx files) to the database, which has to be retrained (“Retrain” button) afterwards to



integrate the new letters. The program can only process letters with a XML data format, like it is used in word documents.

### **3.2 The Similarity Experiment - Second Evaluation of the System**

The system was originally tested on a database of 307 oncology discharge summaries. Patients had different types of cancer entities and the dataset contained “follow-up” letters. Based on the idea, that it is easy to distinguish between patients with different tumour entities it would be interesting to test the system with a new, larger dataset, containing only discharge letters of patients with one common disease. Therefore we created a new dataset of 489 anonymized discharge summaries. All patients have a CLL diagnosis in common. To evaluate the program’s capability with a previously more similar dataset, we designed a new experiment inspired by the studies, that were used previously to verify the recommendation quality. Our experiment is based on the dataset of 489 anonymized CLL patient discharge summaries. Due to a similar experimental composition towards the first evaluation, we expected that the results might be better comparable to each other.

A second issue we wanted to address was based on the fact that experts make their decisions in another way than novices do. For this reason, we wanted to investigate, if experts assess the quality of recommendation in differnt way than novices like young professionals or medical students do.

#### **3.2.1 Dataset of 489 CLL Patients Discharge Summaries – Structure, Content and Anonymization**

For our study, we manually converted 489 discharge letters of CLL patients to an anonymous version. All patients had a CLL diagnosis in common. We used an already established, structured CLL database of patients from the oncology department of the University of Freiburg to acquire CLL discharge letters. All letters are free text documents and written in German. Most patients have multiple letters from different departments. Over the time there are consecutive letters due to follow- up





presentations. “Follow-up letters” should naturally be similar to each other, whereas letters from different departments differ in structure and focused content. Therefore, we took the most current discharge letters from the oncology department for the anonymization, which ensures that there are no follow-up letters. The discharge summary’s form changed only minimally over the period of time, the letters were written. The oldest letters are from 1998 and the most recent from 2017.

#### *Structure and Content of the discharge letters*

Almost every letter is similarly built up in a way that is typical for medical letters, more precisely oncology letters. Beginning with a letterhead, including a greeting and short introduction with a date, followed by a list of the most important diagnosis in note form, important medical details, the predominately oncological therapy history, an actual medical history, a list of the actual patient’s medications and results of different investigations. The latter can include blood counts or results of imaging studies. The letter ends with a discharge summary, called “Epikrise” and a follow up. Wimsett et al. (2014) reviewed the key components for a good discharge summary: “discharge diagnosis, treatment received, results of investigations and follow-up required” (Manning, Raghavan, and Schütze 2008). All of these can be found in our set of discharge summaries. An example letter is shown in the appendix.

#### *Anonymization*

Patient’s information such as name, birth date and pathological sample numbers have been removed. We also excluded the official letterhead (name of the department, address, physician names), hospital names, physician’s names, telephone numbers and place names to ensure anonymization. The patient’s name is replaced by “the patient” and adjusted to German grammar. While we have lost some potentially relevant information, this loss of patient’s information shouldn’t be of consequence for the analysis of the discharge letters in view of the structure of the algorithm. Follow-up letters are from outpatients, discharge letters from the ward.

### **3.2.2 Participants – Selection and Anonymization**

To address the question whether experts judge similarity differently from novices, we decided to split our participants into three groups: experts, junior doctors (interns and



residents) and medical students. As the dataset investigated in this study contains hematologic patients with a CLL diagnosis, all doctors were recruited from the oncology department of the University of Freiburg. We consider a participant as an expert if the person is a specialist in oncology. Junior doctors undergo their medical and oncology training to become an oncologist. Medical students have finished their exam of internal medicine, to assure a certain level of knowledge and to be comparable with each other. All participants were asked personally if they would participate. All group members participated voluntarily. This selection is difficult in terms of external validation. In view of the risk of incomplete anonymization and to guarantee patients' privacy, we decided to ask only members of the University of Freiburg Medical Center. As they are all easily available, differently motivated and participate voluntarily, our sample has to be considered as a convenience sample ('Convenience Sample' 2008). Consequential disadvantages are lack of transferability and generalisation. However, the issue of generalization should be less relevant, because our study aims at this specific group of participants. Due to the explorative character of this study, we aimed to recruit at least five participants in each group, aware of the problem, that medical experts and junior doctors are always short on time. This selected participant size was based on the results of the first study from Hummel, which was done in the framework of the program's development (for more details see 3.1). To protect participant's anonymization, we asked experts and junior doctors only for the period of their practical experience.

### **3.2.3 Questionnaire about CLL and other Questions**

To get an idea about the different levels of knowledge about CLL of the participants, participants were asked to complete two multiple choice questionnaires about CLL, comprising 10 questions each. This certified questionnaires have been developed as a part of the CME (Continuing Medical Education) program (Hochstetter, n.d.; Bergmann and Wendtner, n.d.) and should be filled out before the beginning of the experiment to serve as a baseline. The questionnaires can be found in the appendix.



Additional to the similarity experiment and the questionnaires, we collected additional information with another questionnaire to address several issues, which are associated with the program use and implementation.

As we used a seven-level rating scale, it is not clear at what value participants would say a supposed patient is truly similar. Therefore it would be interesting to also convert our rating scale into a binary rating system. To answer this question we asked the following question:

1. From which value would you say two patients are similar according to their discharge summaries?

To get an idea in what situation such a system could be used in the future and how to integrate it into the daily clinical routine or teaching the following points were addressed

2. Do you think this program could help you in your daily practice?  
Strongly agree, agree, neutral, disagree, strongly disagree  
In which way?

Since we use a seven-level rating scale to rank similarity, nothing is known about how participants judge similarity.

3. On what categories do you focus regarding patients similarity?  
List these categories

It is unclear if this program can address medical problems or find medical characteristics other than well-established methods and programs. To find out if there might be specific applications for our program we asked the following questions:

4. For which medical question would you use such a program?

To record the participant's opinion, if this program could improve medical decision making, the following question was asked:

5. Do you think this program can improve medical decision making?  
Strongly agree, agree, neutral, disagree, strongly disagree



To improve the search algorithm in the future, we wanted to know how many discharge summaries a user would look at until he is satisfied with a search result?

6. After how many suggested similar discharge summaries there should be a satisfying result?
7. What possibilities are there to improve the system?

### **3.2.4 Structure of the Experiment**

In our experiment, participants have to compare two patients' discharge summaries for similarity. A letter pair consists of a so-called "reference-letter", which has to be compared to a "comparison-letter". In a practical setting the reference letter might be the actual patient about which one might find a similar patient out of the database. The comparison letter in our experiment represents this similar patient case. A trial consists of a reference letter, which has to be compared to five different comparison letters. Four of these comparison letters are the ones with the highest (cosine) similarity, computed by the program. In other words, these four discharge summaries are the most similar ones to the reference-letter, according to the program's algorithm. The final, fifth letter is randomly selected from our database to compare the algorithm against chance. The randomly selected letter mostly has a very low (cosine) similarity. The order of the reference letters and comparison letters is randomized and fixed afterwards. Participants rate each pair. A seven-level rating scale from 1 (very dissimilar) to 7 (very similar) is used for this task. We designed the experiment with 22 "reference-letters", thus we collect ratings of 110 letter pairs. After several pre-tests, this number turned out to be reasonable in terms of time required to perform this task. It took about 3 hours to complete the entire experiment, including the rating task and the questionnaires. At the beginning of the experiment, we chose two reference-letters, one with a particularly high computed similarity of the comparison-letters and a very low one. This should help participants to get a feeling for the experiment and the framework within which the experiment takes place. These first two trials are excluded from subsequent analysis, as the participants have to adapt their rating behaviour. The remaining 20 reference-letters were selected from our database, according to different criteria. 10 reference-letters were randomly selected; the remaining 10 were chosen,



based on different individual characteristics. A list of selection criteria can be found in the next section.

### 3.2.5 Reference Letters - Selection Criteria

Our experiment comprises 22 reference letters. The first two letters (see Table 3.2.5.1) are used to give the participants a feeling about the program’s functioning and capability. They should give the participants an idea in which similarity range the experiment takes place. Therefore, we chose a reference letter, whose most similar comparison letter, computed by the algorithm, has a high cosine similarity. This means, in accordance with the program, this letter should be very similar to the reference letter. As the maximal computed cosine similarity of a comparison letter in our dataset is about 0,6 and values over 0,3 should display a high similarity in the preliminary experiment (see Figure 3.1.2) we decided to choose a letter pair with similar cosine similarity. Again, it should be mentioned, that an absolute threshold for a high similarity is hard to define, because cosine similarity is computed based on the actual dataset and can vary, depending on the number and diversity of documents in the dataset. The second trial is one with a quite low computed similarity.

Reference Letter	Characteristic
020	High cosine similarity
096	Low cosine similarity

Table 3.2.5.1: Training trials

The remaining 20 reference letters are split into two subgroups of 10 each. The first ten pairs are chosen with a focus on several characteristics (see Table 3.2.5.2). These characteristics include conspicuous medical characteristics for CLL patients as well as special attention to computed similarity constellations.

Reference Letter	Characteristic
190	Highest cosine similarity; value leap
128	Multiple Myeloma; value leap
204	High cosine similarity
002	PBSCT; high cosine similarity



<b>054</b>	Ibrutinib
<b>463</b>	PBSCT, Richter Transformation
<b>140</b>	High cosine similarity
<b>286</b>	High cosine similarity
<b>89</b>	Richter Transformation, low cosine similarity
<b>168</b>	Stable disease

Table 3.2.5.2: Selected reference letters

A view letter-pairs are chosen with a focus on their cosine similarities, due to the fact, that this is the final parameter after which the program judges similarity. We wanted to ensure that there are letter-pairs with a big enough variety of different cosine similarity values in our experiment. Letter number 128 and 190 show relatively high cosine similarity value leaps between the most similar comparison letters, therefore it would be interesting if this leaps will be reflected by the similarity rating of the participants. Additionally, 190 is the reference letter, that has the most similar letter, according to the algorithm with a cosine similarity of 0,60. Number 2, 140, 204, 286, all have been selected because of high cosine similarity ratings. In contrast 89 has a quite low one. This selection should ensure the comparability between high and low cosine similarity. Except for interesting cosine similarities, some of the chosen letters also contain interesting medical features or have been selected solely because of a medical issue. Discharge summary 168 is from a patient, who never needed any interventions, regarding his CLL. 89 and 463 developed a so-called Richter transformation, which presents as a massive deterioration of the diseases course. Number 2, 463 and 54 were selected as they received a prominent treatment. 2 and 463 were treated with haematopoietic stem-cell transplantation and 54 with a relatively new drug, called ibrutinib. Patient 128 developed another haematopoietic cancer in addition to its CLL. The last ten reference letters are chosen by chance to reduce the bias of only selected recommendation letters (see Table 3.2.5.3). We used the online platform “random.org” (‘RANDOM.ORG - True Random Number Service’ n.d.) for our selection.

<b>Random Reference Letters</b>	<b>069,353,259,302,196,171,007,049,344,037</b>
---------------------------------	--

Table 3.2.5.3: Random reference letters



### **3.2.6 Procedure**

At the beginning, participants had to consent in writing, which can be found in the Appendix. Before the experiment had started, participants were asked to do a multiple-choice test about CLL. Afterwards, they got an introduction and explanation about the specific course of the experiment, which had to be made on a Laptop computer.

The experiment consists of 22 trials, in which participants have to compare letter pairs for similarity. There are 22 reference letters and participants have to rate five assigned comparison letters to each reference letter. A seven-level rating scale from 1 (very dissimilar) to 7 (very similar) is used for each pair. After ranking a reference letter with its five assigned comparison letters, the results have to be saved before skipping to the next reference letter. Modifications are possible over the entire time but have to be saved. The introduction draws attention to the fact, that all patients have a CLL diagnosis in common. There are no time constraints, but participants are recommended to deal with discharge summaries in a way they would do in their practical daily life. No further advice on how to rate similarity is given to the participants, but demands to the investigator are always possible. The investigator is present or reachable during the experiment. After the experiment, the participants were pleased to answer a questionnaire, containing a list of mainly explorative questions.

## **3.3 Biometric Aspects**

### **3.3.1 Hypothesis and Null Hypothesis**

The first hypothesis, this theses address, is following:

1. Experts rate similarity in a way that correlates with the computed similarity of the system.

As we expect a non-linear positive correlation between the measured similarity and the according cosine similarity, a suitable way to compute the relationship between the two variables (average rating and cosine similarity) is to use the Spearman's rank



correlation coefficient ( $r_s$ ). Therefore the hypothesis  $H_{a1}$  and the corresponding null hypothesis  $H_{01}$  can be illustrated as follows:

$$H_{a1} \quad r_s > 0$$

$$H_{01} \quad r_s \leq 0$$

The second hypothesis is:

2. By the system suggested similarity correlates stronger the more experienced a participant is.

Again, the Spearman's rank correlation coefficient should be used to compare the different correlations between the average rating of the three groups (experts, junior doctor and student) and the computed similarity. To answer the question if there is a difference between the groups we have to compare each group individually.

$$H_{a2.1} \quad r_s (\text{expert}) > r_s (\text{junior doctor})$$

$$H_{02.1} \quad r_s (\text{expert}) \leq r_s (\text{junior doctor})$$

$$H_{a2.2} \quad r_s (\text{expert}) > r_s (\text{student})$$

$$H_{02.2} \quad r_s (\text{expert}) \leq r_s (\text{student})$$

$$H_{a2.3} \quad r_s (\text{junior doctor}) > r_s (\text{student})$$

$$H_{02.3} \quad r_s (\text{junior doctor}) \leq r_s (\text{student})$$

Another possibility to investigate the difference between experts and novices is to calculate the correlation between the individual results of the multiple-choice test and the individual correlation between the rating and the system. If experience is considered as the period of practical working the correlation between the time and the rating correlation could be used. As significance level ( $\alpha$ ) we choose 0,05.





### 3.3.2 Statistics

The Spearman's rank correlation coefficient ( $r_s$ ) is suitable to compute the correlation between the average similarity rating and the computed cosine similarity after they have been converted to ranks. The results of the preliminary experiment have pointed to (see Figure 3.1.2) a non-linear correlation. Therefore, a similar distribution can be expected and consequently the Spearman's rank correlation coefficient is more suitable for this task. The standard error of the coefficient ( $\sigma$ ) for a sample size  $n$  is ('Spearman's Rank Correlation Coefficient' 2018):

$$S(r_s) = \frac{0.6325}{(\sqrt{n-1})}$$

$n = 100 \rightarrow$  95% CI is 0.12459453795

### 3.4 Literature

We used as main databases *Pubmed* and *MEDLINE* to look for literature. "Google scholar" was used to find a wider spectrum of publications. Another tool we used was the *KatalogPlus* from the Universitätsbibliothek Freiburg, a college library database.

### 3.5 Software

For data analysis and illustrations we used different Python libraries: Pandas ('Python Data Analysis Library — Pandas: Python Data Analysis Library' n.d.), Numpy ('NumPy — NumPy' n.d.), SciPy ('SciPy.Org — SciPy.Org' n.d.), Matplotlib ('Matplotlib: Python Plotting — Matplotlib 2.2.2 Documentation' n.d.) and Seaborn ('Seaborn: Statistical Data Visualization — Seaborn 0.9.0 Documentation' n.d.).



## 4 Results

### 4.1 Data Analysis

#### 4.1.1 Data Computation for the Experiment

The algorithm computed the according cosine similarities for each letter pair (reference letter and comparison letter). A trial in the experiment consists of a reference letter, which has to be compared to five comparison letters. The fifth comparison letter is randomly chosen and therefore normally has a low cosine similarity. The results for training trials (Table 4.1.1.1), selected trials (Table 4.1.1.2) and random trials (Table 4.1.1.3) are shown below. The order in the experiment was randomized.

Reference Letter	Comparison Letter 1.	2.	3.	4.	Random Letter	Characteristic
020	396 (0,48)	454 (0,38)	081 (0,36)	138 (0,34)	107 (0,03)	High cosine similarity
096	123 (0,10)	092 (0,08)	343 (0,08)	223 (0,07)	169 (0,07)	Low cosine similarity

Table 4.1.1.1: Data computation for training pairs; computed cosine similarity in brackets

Reference Letter	Comparison Letter 1.	2.	3.	4.	Random Letter	Characteristic
190	011 (0,60)	256 (0,24)	189 (0,17)	060 (0,13)	177 (0,03)	Highest cosine similarity; value leap
128	486 (0,24)	144 (0,17)	402 (0,13)	226 (0,12)	155 (0,03)	Multiple Myeloma; value leap
204	030 (0,37)	286 (0,37)	340 (0,35)	353 (0,35)	231 (0,02)	High cos sim
002	441 (0,39)	438 (0,35)	009 (0,35)	369 (0,27)	165 (0,02)	PBSCT; high cos sim
054	012 (0,29)	210 (0,24)	025 (0,21)	080 (0,21)	204 (0,09)	Ibrutinib
463	184 (0,19)	009 (0,19)	397 (0,18)	002 (0,18)	448 (0,02)	PBSCT, Richter Transformation
140	340 (0,33)	103 (0,30)	318 (0,30)	204 (0,30)	002 (0,06)	High cos sim
286	030 (0,57)	064 (0,39)	204 (0,37)	171 (0,36)	292 (0,11)	High cos sim,
89	463 (0,12)	418 (0,11)	184 (0,10)	300 (0,10)	349 (0,05)	Richter Transformation, low cos sim
168	057 (0,30)	039 (0,20)	416 (0,19)	021 (0,17)	314 (0,01)	Stable disease



Table 4.1.1.2: Data computation for selected reference letters; computed cosine similarity in brackets

Reference Letter	Comparison Letter 1.	2.	3.	4.	Random Letter
069	404 (0,19)	345 (0,14)	300 (0,13)	205 (0,12)	299 (0,08)
353	030 (0,37)	204 (0,35)	286 (0,34)	476 (0,34)	393 (0,02)
259	202 (0,10)	039 (0,08)	418 (0,08)	222 (0,08)	417 (0,02)
302	290 (0,30)	441 (0,27)	002 (0,21)	010 (0,21)	088 (0,04)
196	220 (0,13)	288 (0,12)	415 (0,12)	210 (0,11)	027 (0,03)
171	030 (0,46)	286 (0,36)	057 (0,35)	064 (0,23)	417 (0,10)
007	085 (0,12)	225 (0,09)	262 (0,08)	312 (0,07)	012 (0,02)
049	354 (0,14)	353 (0,12)	445 (0,12)	342 (0,11)	073 (0,04)
344	185 (0,15)	402 (0,14)	046 (0,13)	068 (0,12)	279 (0,04)
037	446 (0,21)	074 (0,12)	229 (0,11)	391 (0,09)	364 (0,02)

Table 4.1.1.3: Data computation for randomly chosen reference letters; computed cosine similarity in brackets

#### 4.1.2 Participants Characteristics

Table 4.1.2.1 is giving an overview about the participant's characteristics. All participants meet the chosen conditions. All medical students are enrolled at the University of Freiburg and passed the exam of internal medicine. Junior doctors (assistants) are passing their training as oncology specialists. All experts are specialists for haematology. Approximately half of the participants are women.

	Average Experience	Average MC Test Result (max. 20 Points)
Students	6.1 years (study time)	8.6
Junior Doctors	1.9 years (working experience)	14.2
Experts	22 years (working experience)	17.2

Table 4.1.2.1: Average participant's characteristics

#### 4.1.3 Rating Analysis

To get a first impression of how participants judged the recommender capability of the system we want to find out how participants rated the letter pairs. Figure 4.1.3.1 is



showing the individual rating behaviour of each participant for all letter pairs, except the first two reference letters, which served as familiarization letters. Remember, a seven-level rating scale was used to assess similarity. It is notable that some participants (e.g. assistant 3, expert2) avoided extreme responds categories, an issue, which is known as the central tendency bias.

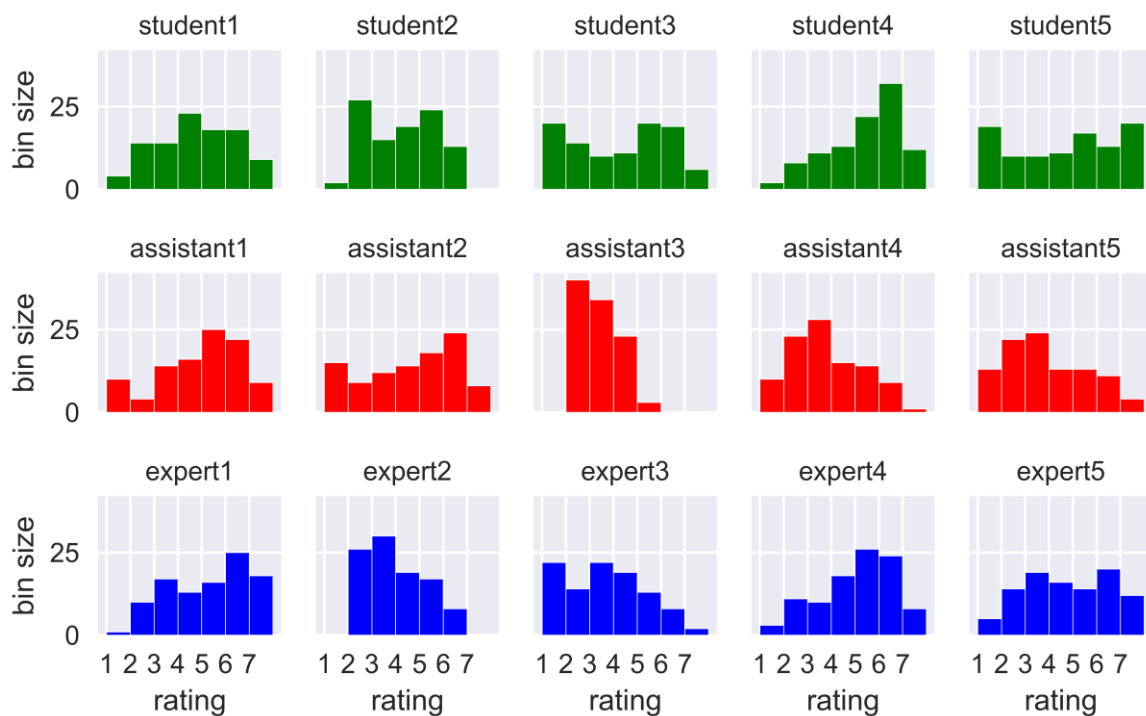


Figure 4.1.3.1: Histograms of rating behaviour of each participant, including random letters. Central tendency bias can be seen in several participants, e.g. assistant3, expert2

Figure 4.1.3.2 shows the rating behaviour of each group in a stacked histogram.

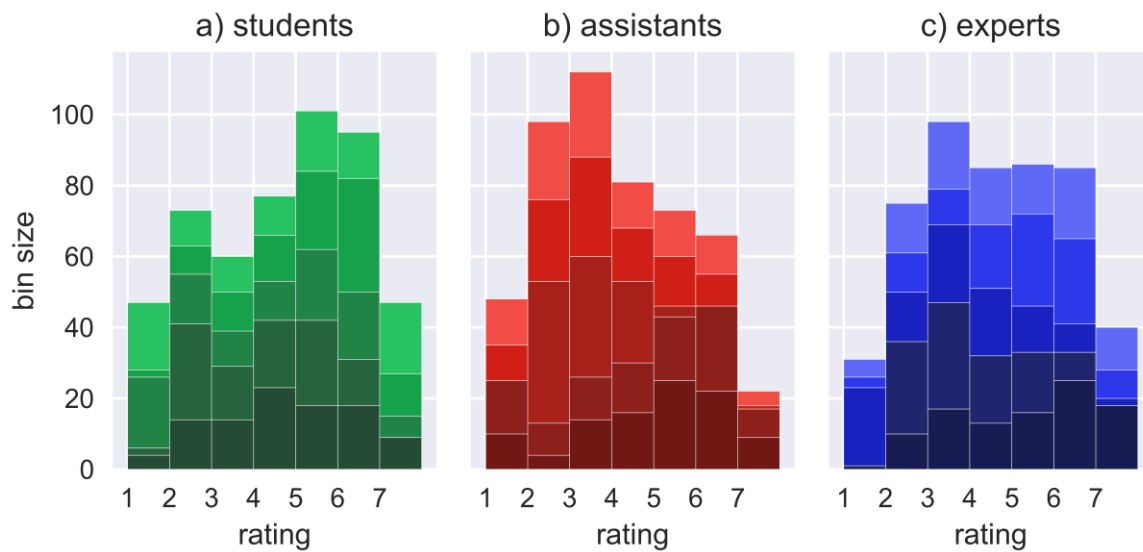


Figure 4.1.3.2: Rating behaviour of each group, individual ratings are stacked upon each other a) students b) assistants c) experts

Without random selected letters, the median rating of the first four similar computed letters was 4 in the expert group, 3 in the assistant group and 4 in the students group.

#### 4.1.4 Repeatability

In our experiment two letter-pairs had to be rated two times. The pairs are shown in Table 4.1.4.1. The position in the experiment is shown, as well as the individual ratings. The first five ratings are the ratings of the experts, followed by assistant's rating and student's rating. The Cronbach's alpha (tau-equivalent reliability) is used to compute the internal consistency ('Cronbach's Alpha' 2018).

Letterpair	Position	Ratings	Cronbach's alpha
204 – 286	35	[4, 3, 3, 5, 3, 4, 4, 2, 3, 3, 4, 3, 6, 5, 7]	0.781 (0.7 < alpha < 0.8 <b>acceptable</b> internal consistency)
286 – 204	84	[5, 4, 5, 5, 2, 5, 5, 3, 2, 2, 3, 2, 4, 5, 6]	
171 - 286	57	[6, 6, 6, 6, 5, 4, 3, 2, 5, 3, 7, 3, 2, 5, 6]	0.570



<b>286 - 171</b>	82	[4, 5, 5, 6, 1, 6, 5, 2, 5, 3, 3, 3, 2, 3, 5]	(0.5 < alpha < 0.6 <b>poor</b> internal consistency)
------------------	----	--	--

Table 4.1.4.1: Repeatability: Internal consistency for double letter pairs using Cronbach’s alpha

The internal consistency for the letter-pair “204-286” is 0.781, that can be interpreted as acceptable and the second letter-pair “171-286” has a poor internal consistency, according to a general interpretation scale of Cronbach’s alpha. Reasons for the rating variations might be a changing focus on different aspects of similarity during the experiment in general and a changing focus on different aspects of the same patient. For example a participant might first focus on the patient’s diagnosis and later more on his treatment plan. Another possible explanation might be the different comparison letters in each trial. Participants tended to rate the similarity for each pair in relation to the other comparison letters.

#### 4.1.5 Inter-rater Agreement

We calculated the inter-rater agreement (Table 4.1.5.1) by using Spearman’s rho to calculate the pairwise inter-rater agreement. The pairwise correlation between each pair of participants is illustrated in Figure 4.1.5.1. To compute the average level of agreement for each group we calculated the mean of the correlations of each group. The average of all possible combinations of Spearman rank coefficients is also referred to as Kendall’s coefficient of concordance or Kendall’s W. Kendall’s W is a common measure for assessing agreement among raters.

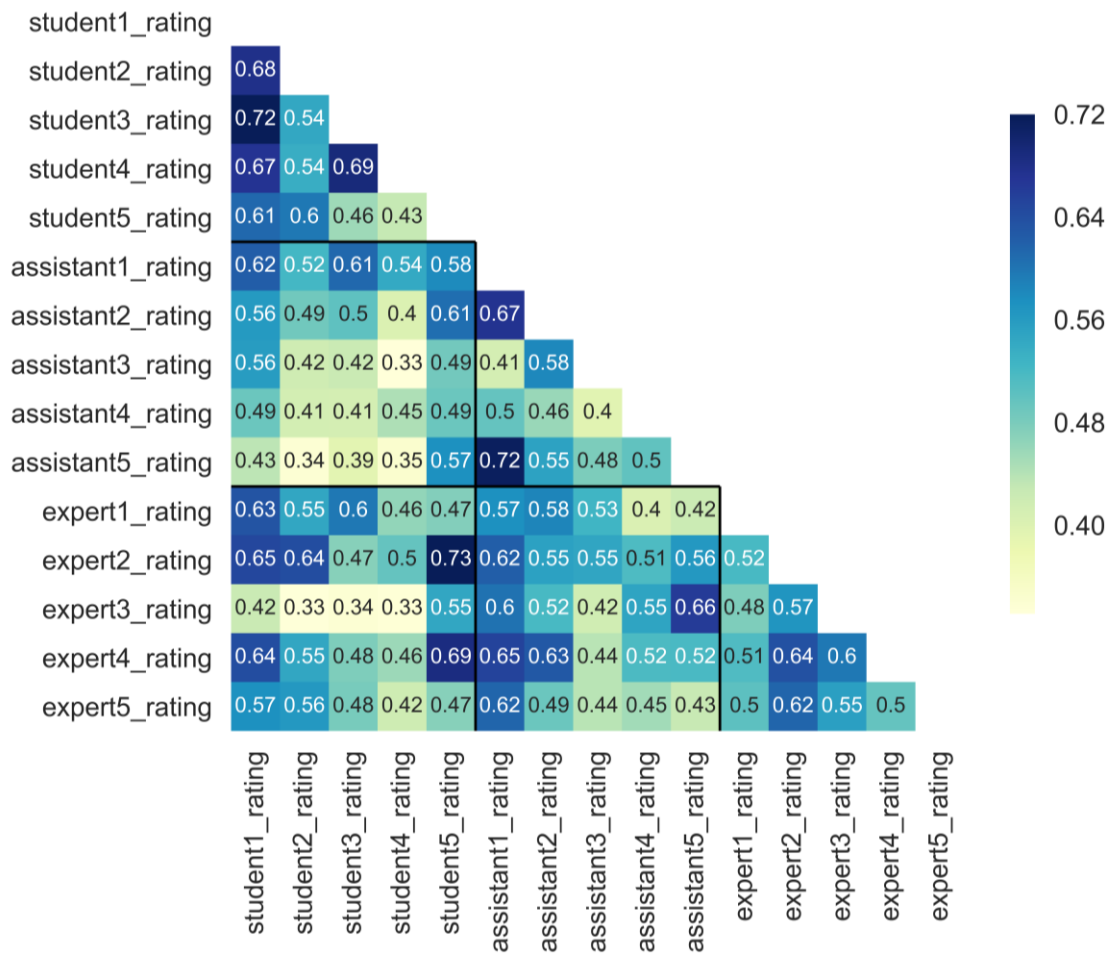


Figure 4.1.5.1: Inter-rater agreement between each participant

The maximal and minimal correlation between two raters is 0.728 (expert2 and student5) and 0.331 (student4 and assistant3), respectively.

Inter-rater agreement	Students	Assistants	Experts	All participants
Mean inter-rater Spearman correlation (Kendalls W)	0.594	0.527	0.549	0.523

Table 4.1.5.1 Average inter-rater agreement



#### 4.1.6 Testing against Chance

Although the systems superiority against chance has been already shown in the preceding evaluation experiment from Hummel et al, we wanted to confirm this finding, as it is unclear if it is applicable for our dataset as well. Therefore, we compare the ratings of the as most similar retrieved letter, (ranked at first place according to the algorithm) with the randomly chosen letter. Figure 4.1.6.1 illustrates the pairwise rating difference and compares the results of experts, junior doctors and students. The results are also shown in Table 4.1.6.1.

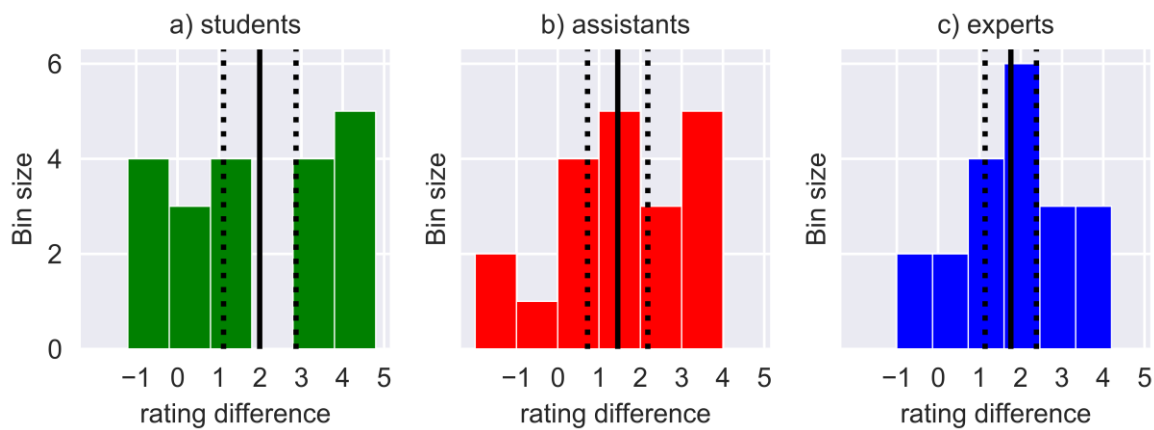


Figure 4.1.6.1: Rating difference between the most similar letter, according to the algorithm and the randomly assigned letter, including the mean difference and the 95% confidence interval. a) Students mean rating difference 1.99 (95%CI: [1.11, 2.87]) b) junior doctors mean rating difference 1.45 (95%CI: [0.72, 2.18]) c) experts mean rating difference 1.76 (95%CI: [1.14,2.38])

	Students	Assistants	Experts
<b>Rating difference (most similar – random)</b>	1.99 (95%CI: [1.11, 2.87])	1.45 (95%CI: [0.72, 2.18])	1.76 (95%CI: [1.14,2.38])

Table 4.1.6.1: Rating difference between the most similar letter, according to the algorithm and the randomly assigned letter

As we assume experts to be the medical gold standard, their rating behaviour is the benchmark for the comparison between the retrieved letters and the random letters. For Experts, the mean rating difference is 1.76 (95% CI: [1.14,2.38]). The confidence interval was calculated with the standard error of the mean (SEM). This assumes a





normal distribution of rating means, which could be questionable since these means are calculated for 20 letter pairs. Another way to compare the randomly chosen comparison letters to the computed ones is to illustrate the rating results in a scatterplot (Figure 4.1.6.2). Randomly chosen comparison letters have a lower computed similarity.

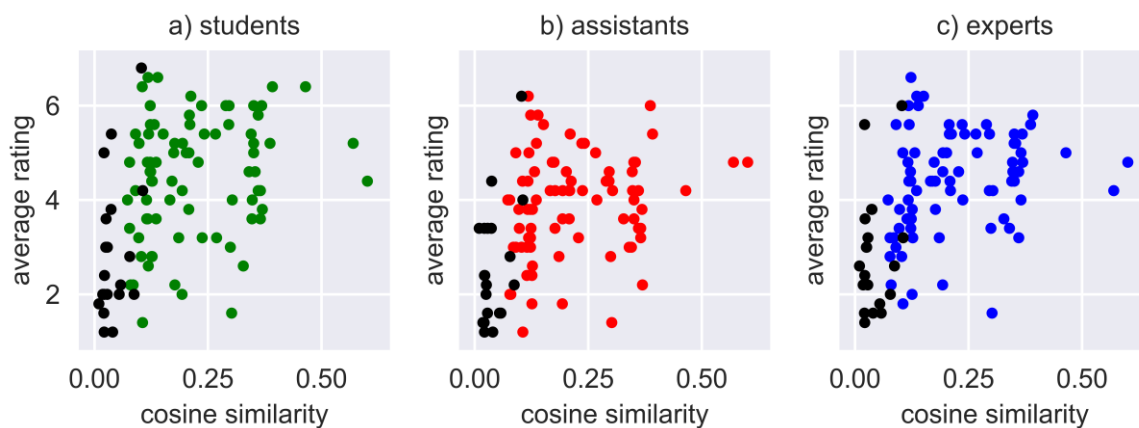


Figure 4.1.6.2: Average rating versus the computed cosine similarity of all letter pairs. Letter pairs with random comparison letters are black. a) Students b) assistants c) experts

Although the random letters have low cosine similarity values some of them are rated as similar.

#### 4.1.7 Correlation between Experts and the Program

For the quality of the clinical recommender system it is crucial that there is an agreement between the retrieved patients and the rating of similarity by medical experts of these patients. Figure 4.1.7.1 shows a scatterplot of the computed cosine similarity versus the average expert rating. To calculate the correlation between the experts rating and the program we used the Spearman rank correlation coefficient ('Spearman's Rank Correlation Coefficient' 2018), a measure for assessing nonlinear relationships of rank correlation.

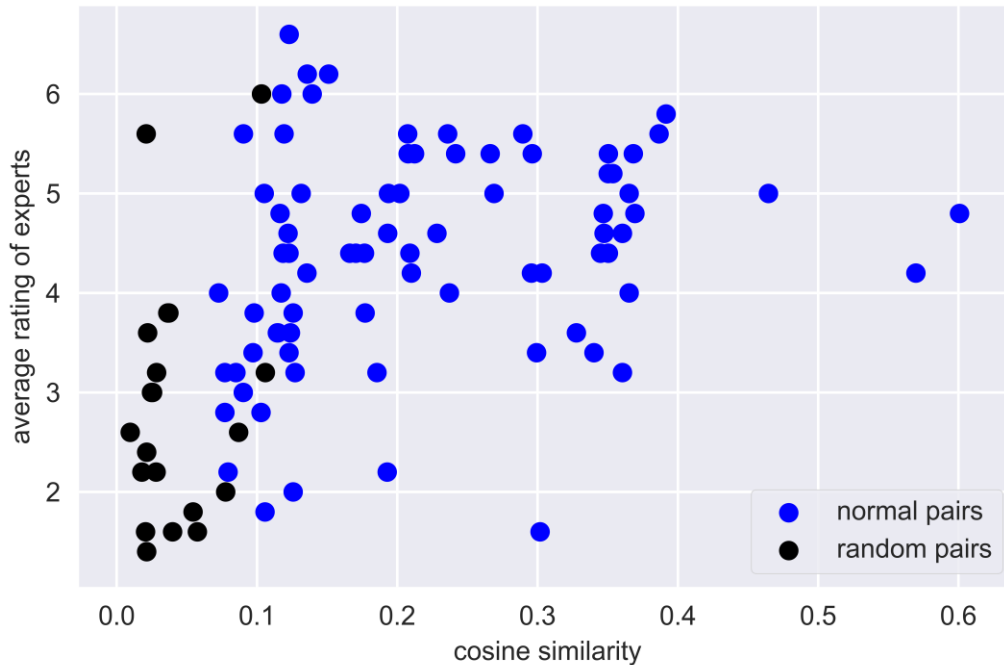


Figure 4.1.7.1: Cosine similarity versus average expert rating

We calculated the Spearman correlation with the average expert rating for each letter pair by using the Python libraries Pandas and SciPy. Both reach the exact same results. The Spearman correlation between the system and the average expert rating is 0.512 (95% CI: [0.387,0.637]). Another possibility is to compute the correlation for each expert and afterwards to compute the mean of these correlations. For comparison we calculated the more familiar Pearson correlation for linear relationships, which is 0.426. All results are shown in Table 4.1.7.1

Correlation	Experts
Spearman correlation ('Pandas.DataFrame.Corr — Pandas 0.23.3 Documentation' n.d.), computed with <u>average ratings</u>	0.512
Spearman correlation ('Scipy.Stats.Spearmanr — SciPy v0.14.0 Reference Guide' n.d.), computed with <u>average ratings</u>	0.512
<u>Mean Spearman correlation</u>	0.406
Kendall's tau coefficient ('Pandas.DataFrame.Corr — Pandas 0.23.3 Documentation' n.d.), computed with average ratings	0.366



<b>Pearson correlation ('Pandas.DataFrame.Corr — Pandas 0.23.3 Documentation' n.d.), computed with <u>average ratings</u></b>	0.426
---	-------

Table 4.1.7.1: Spearman correlation coefficient, Kendall's tau coefficient and Pearson correlation for experts, Pandas and SciPy provide the same results

In the following sections of the thesis, we calculate the Spearman correlation coefficient of each group with the average ratings. Therefore the correlation of a group can be higher than the individual correlations.

#### 4.1.8 Differences between Selected and Randomly Chosen Trials

After dismissing the first two trials for familiarisation reason, 20 trials remained for analysis. A trial consists of a reference letter and the related four best fitting letters, according to the algorithm, plus a randomly chosen letter. The average rating of all participants for the ten selected and the ten randomly chosen reference trials is shown in Figure 4.1.8.1. Table 4.1.8.1 shows the according correlations.

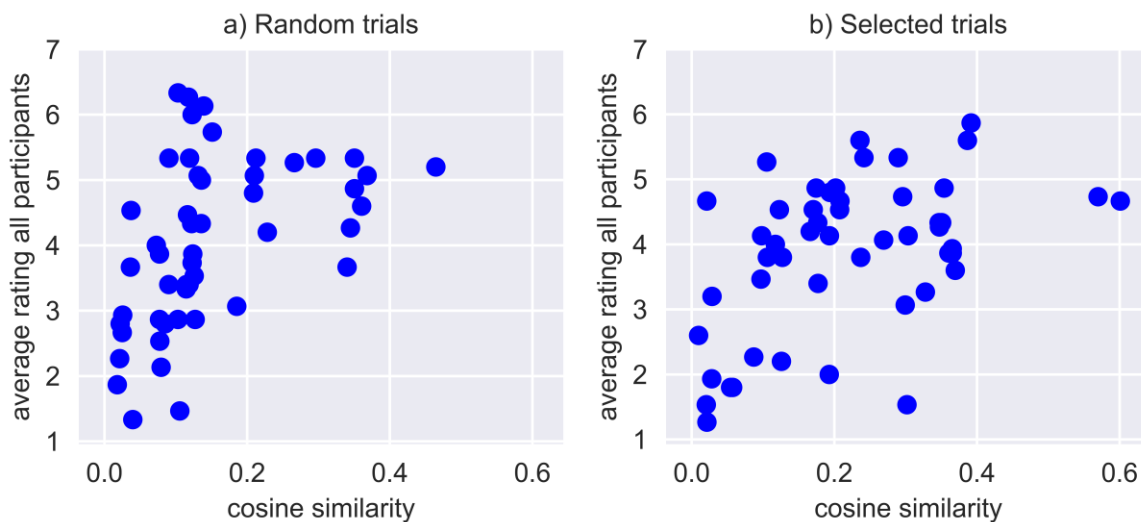


Figure 4.1.8.1: Cosine similarity versus average rating of all participants for a) random trials and b) selected trials

<b>Spearman correlation random trials (all participants)</b>	<b>Spearman correlation selected trials (all participants)</b>
--	--



0.596 (95% CI: [0.471, 0.721])	0.440 (95% CI: [0.315, 0.565])
--------------------------------	--------------------------------

Table 4.1.8.1: Correlation of random trials and selected trials.

The correlation of random trials of all participants was 0.596 (95% CI: [0.471, 0.721]) and the correlation of the selected trials was 0.440 (95% CI: [0.315, 0.565]).

### 4.1.9 Striking Trials

During the data analysis we discovered some outliers. Letter-pairs, who should be similar according to the algorithm, but who are rated as very dissimilar and vice versa letter-pairs with a low computed similarity and a comparatively high participant rating. The striking trials are marked in figure 4.1.9.1.

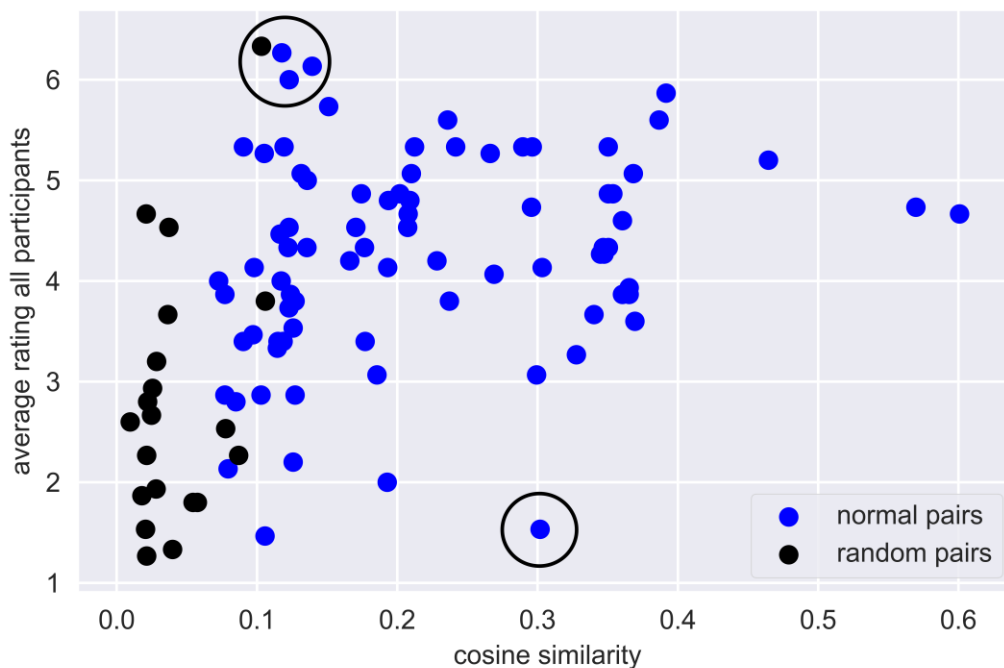


Figure 4.1.9.1: Striking trials are circled.

Reference letter	Comparison letter	Cosine similarity	Average rating of all participants
37	74	0.123	6.000
171	417	0.103	6.333
49	445	0.118	6.267
344	402	0.140	6.133



140	318	0.302	1.533
-----	-----	-------	-------

Table 4.1.9.1: Striking trials

The letter numbers are listed in table 4.1.9.1. We examined the outliers manually and found good reasons, which support the rating behaviour of the participants. The letter pair 171 – 417 can be found in the appendix. Both patients presented themselves for a follow-up appointment. They never got any medical treatment and always had a stable disease.

#### 4.1.10 Individual Participant Correlation

Figure 4.1.10.1 illustrates the individual rating versus the cosine similarity for each participant to get another impression of the people’s rating behaviour.

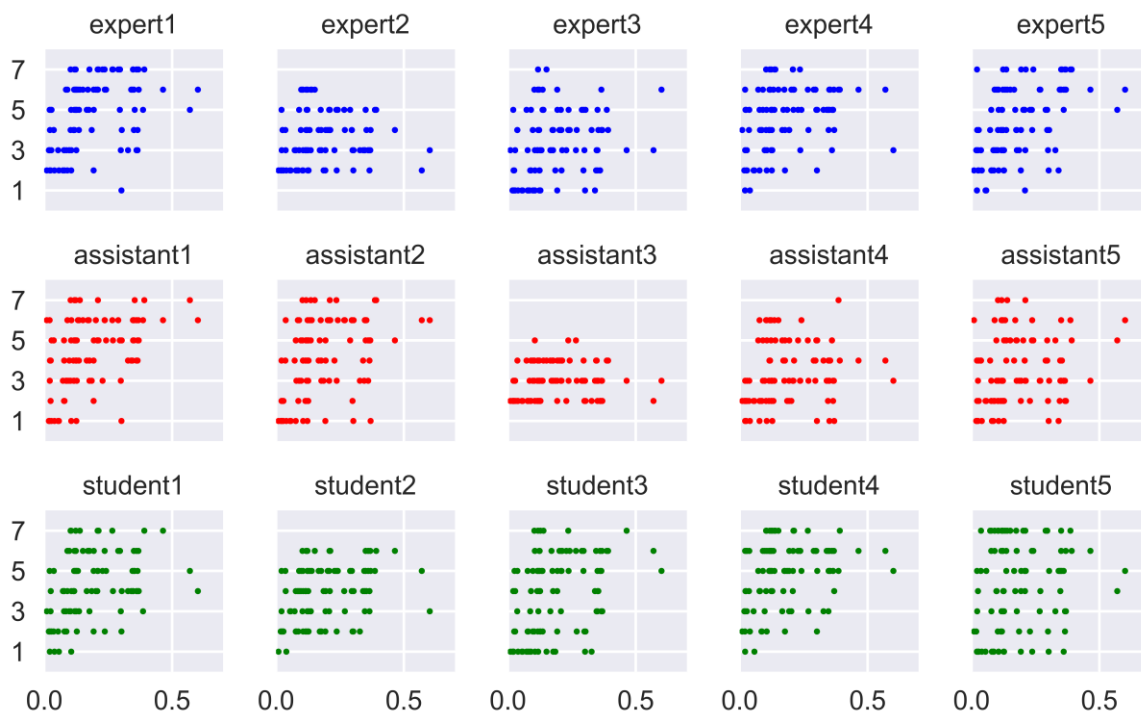


Figure 4.1.10.1: Individual rating versus the computed cosine similarity for each participant.

Table 4.1.10.1 lists the Spearman correlation for each participant towards the system. The lowest correlation is 0.256 (95% CI: [0.131, 0.381]) (expert2). The highest correlation is 0.554 (95% CI: [0.429, 0.679]) (expert5).



Participant	Spearman correlation
Student1	0.435 (95% CI: [0.310, 0.560])
Student2	0.423 (95% CI: [0.298, 0.548])
Student3	0.425 (95% CI: [0.300, 0.550])
Student4	0.307 (95% CI: [0.182, 0.432])
Student5	0.256 (95% CI: [0.131, 0.381])
Assistant1	0.435 (95% CI: [0.310, 0.560])
Assistant2	0.460 (95% CI: [0.335, 0.585])
Assistant3	0.258 (95% CI: [0.133, 0.383])
Assistant4	0.312 (95% CI: [0.187, 0.437])
Assistant5	0.265 (95% CI: [0.140, 0.390])
Expert1	0.484 (95% CI: [0.359, 0.609])
Expert2	0.256 (95% CI: [0.131, 0.381])
Expert3	0.345 (95% CI: [0.220, 0.470])
Expert4	0.390 (95% CI: [0.265, 0.515])
Expert5	0.554 (95% CI: [0.429, 0.679])

Table 4.1.10.1: Spearman correlation for each participant

#### 4.1.11 Correlation of all Participants

Additionally, we calculated the correlation of the mean rating from all fifteen participants. The correlation for all fifteen participants is 0.502 (95% CI: [0.377, 0.627]) and the related distribution is shown in figure 4.1.11.1

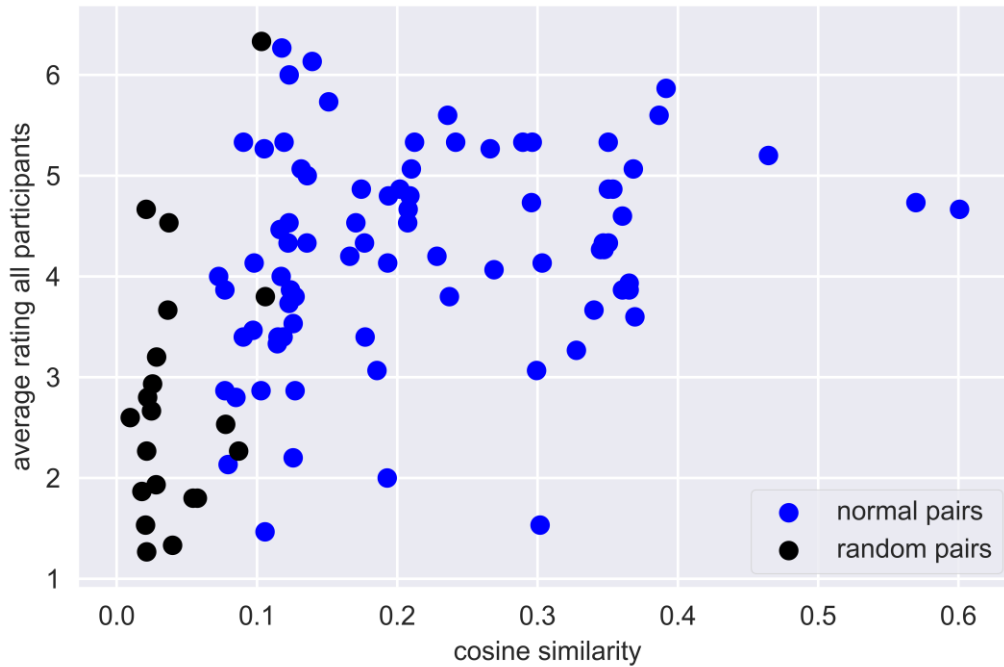


Figure 4.1.11.1: Mean rating of all fifteen participants versus computed cosine similarity.

#### 4.1.12 Differences between Experts, Junior Doctors and Students

We expected differences in the rating behaviour between the groups due to the different level of medial experience. We asked students, junior doctors (assistants) and experts to make our experiment and hypothesised that expert rating correlates more strongly with the computed similarity. The Spearman correlation for each group is calculated between the average rating of each group and the computed cosine similarity (figure 4.1.12.1 and table 4.1.12.1).

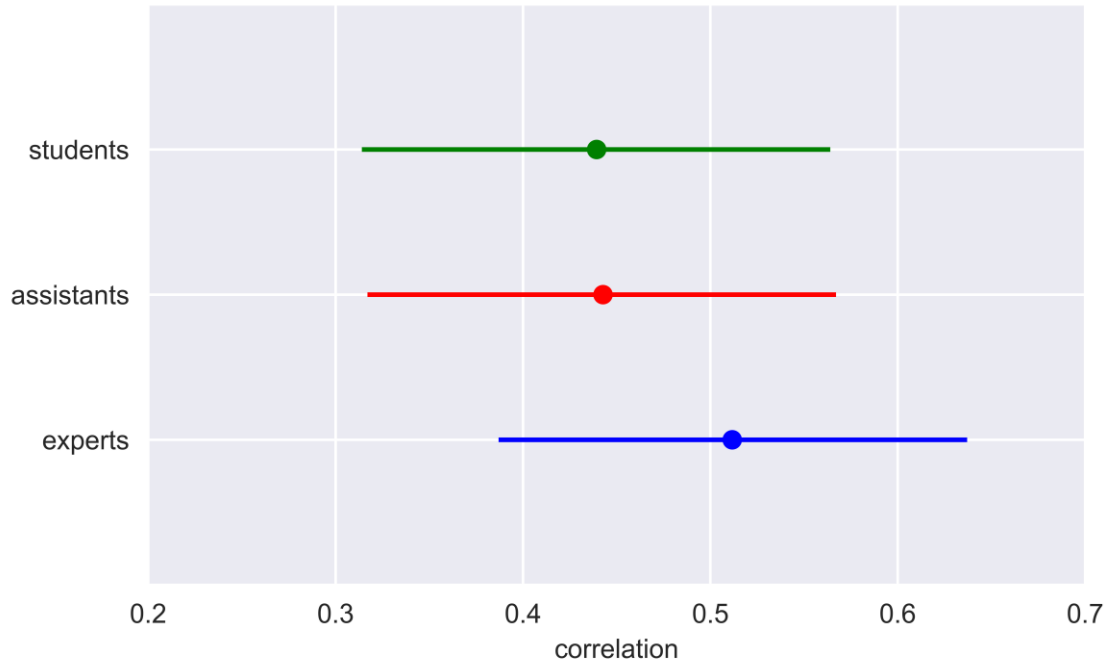


Figure 4.1.12.1: Correlation between the average rating of each group and the computed similarity for each group with the related 95% Confidence intervals.

Correlation	Students	Assistants	Experts
<b>Spearman</b>	0.439 (95% CI: [0.314,0.564])	0.443 (95% CI: [0.318,0.568])	0.512 (95% CI: [0.387,0.637])
<b>Pearson</b>	0.391	0.377	0.426

Table 4.1.12.1: Correlation of each group

Figure 4.1.12.1 is illustrating that there is (most likely) no significant difference between the three groups. We fail to reject the null hypothesis, that the correlation of experts is higher than the correlation of assistants or students.

### *Correlation versus Experience*

The transformation from a novice into a clinical expert takes time. The correlation of each participant and each group towards the system over the practical work experience can be seen in figure 4.1.12.2.



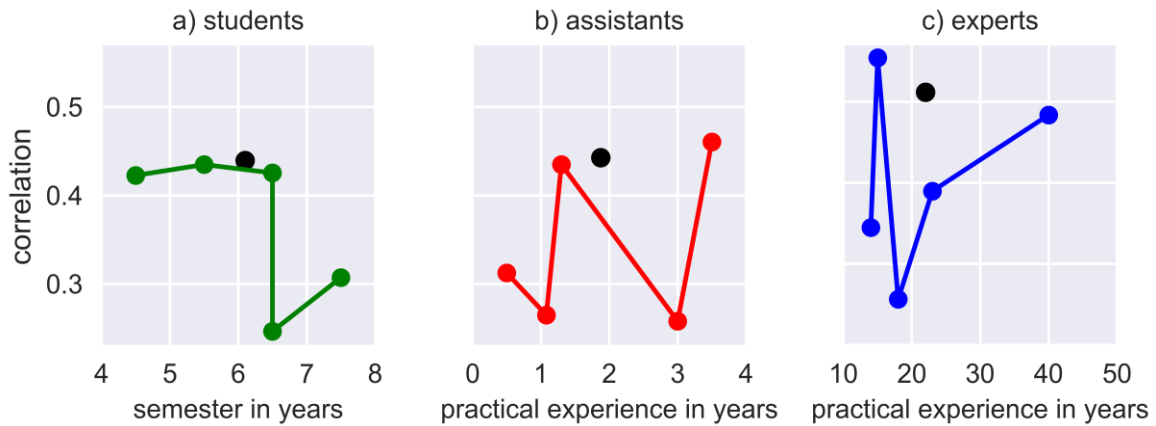


Figure 4.1.12.2: Correlation versus experience, development over time. Notice that the correlation for each group is computed with the average rating. a) students b) assistants c) experts

A remarkable fact is that the correlation computed with the average rating of each group is higher than the individual rating correlation of each participant. We expected an increasing correlation with increasing practical experience. Figure 4.1.12.3 is showing the relation between experience and correlation in a single graph for all participants. The distribution indicates no correlation between practical experience and the correlation of each participants rating towards the system.

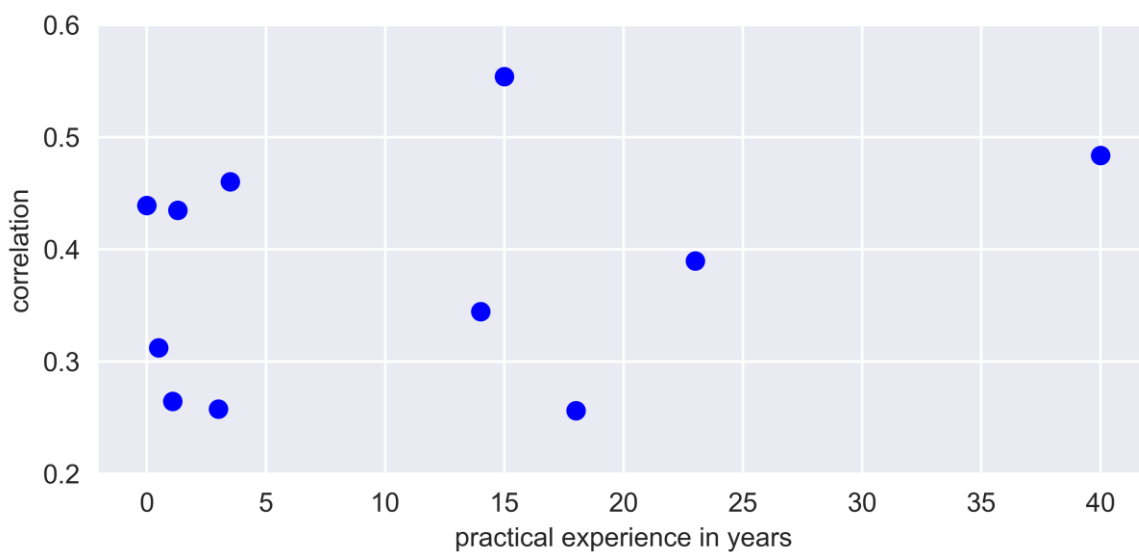


Figure 4.1.12.3: Correlation versus experience; students correlation is at the starting point and represents the null value



### *Influence of Disease Knowledge*

To get an idea about the participant's initial disease knowledge, we asked participants to fill in two validated CME multiple-choice tests about CLL. An absolute score of 20 was the maximum. Normally, such tests try to cover actual guideline knowledge and refer to the related article. Therefore generalization in term of general disease knowledge is difficult and the correlation between the test results and the rating result has to be seen very carefully. Especially more experienced doctors tend to differ from guidelines, as they know a higher variety of treatment options. Nevertheless, a tendency can be seen between the groups of novices and in practice, you would only agree with the results of a doctor if he also knows the actual guidelines, even if he might differ for good reasons. Figure 4.1.12.4 is showing the results of the multiple-choice test versus the rating correlation towards the program for each group.

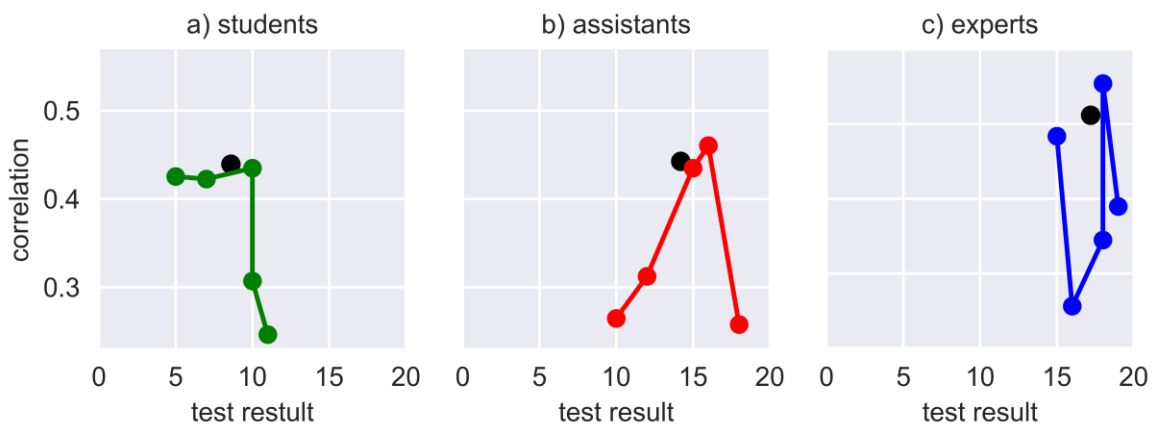


Figure 4.1.12.4: Correlation versus result multiple choice test, correlations of average ratings of each group are marked in red a) students b) junior doctors c) experts

Students had an average test result of 8.6, junior doctors of 14.2 and experts of 17.2, respectively. The relation between the test results and the individual participants correlation regardless of the different groups is shown in figure 4.1.12.5. The computed spearman correlation is 0.07, saying that there is no connection between the tested disease knowledge and the correlation towards the program, although there are



differences in the test results between the three groups. This finding is possibly explainable by the fact that the multiple-choice test is only testing guideline knowledge, but participants used their experience and other categories for the rating task. We address the question, which categories the participants have used in a following section.

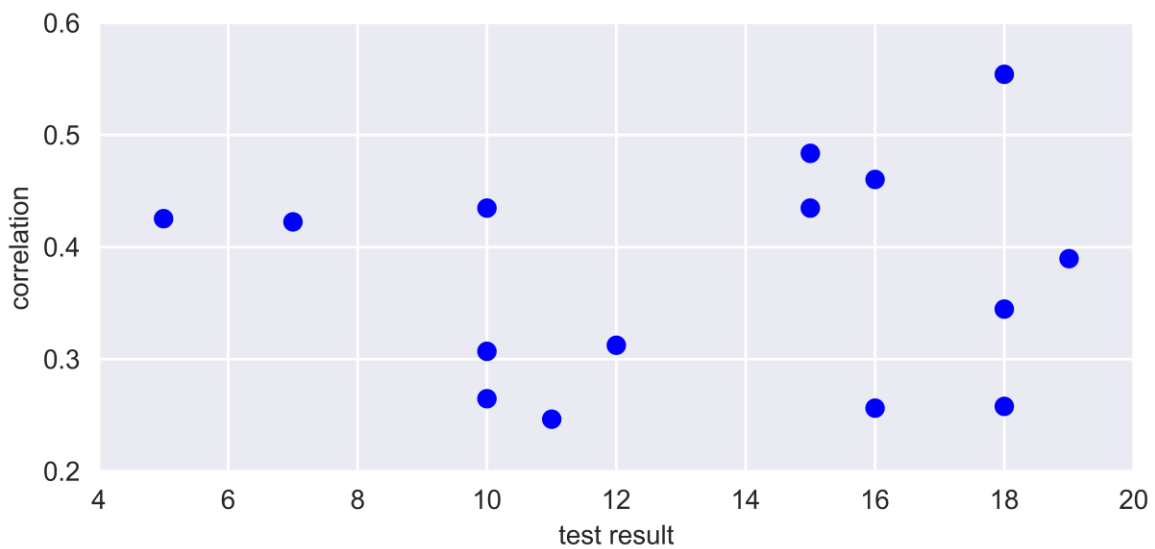


Figure 4.1.12.5: Correlation versus multiple choice test result

#### 4.1.13 Similar or Not – Assessing Recommendation Quality

In the end, for the program's user it might be only important to know whether a presented discharge summary is similar or not. Thus, we asked the participants to name their individual cut off to translate the seven-level rating scale into a binary scale (Figure 4.1.13.1).

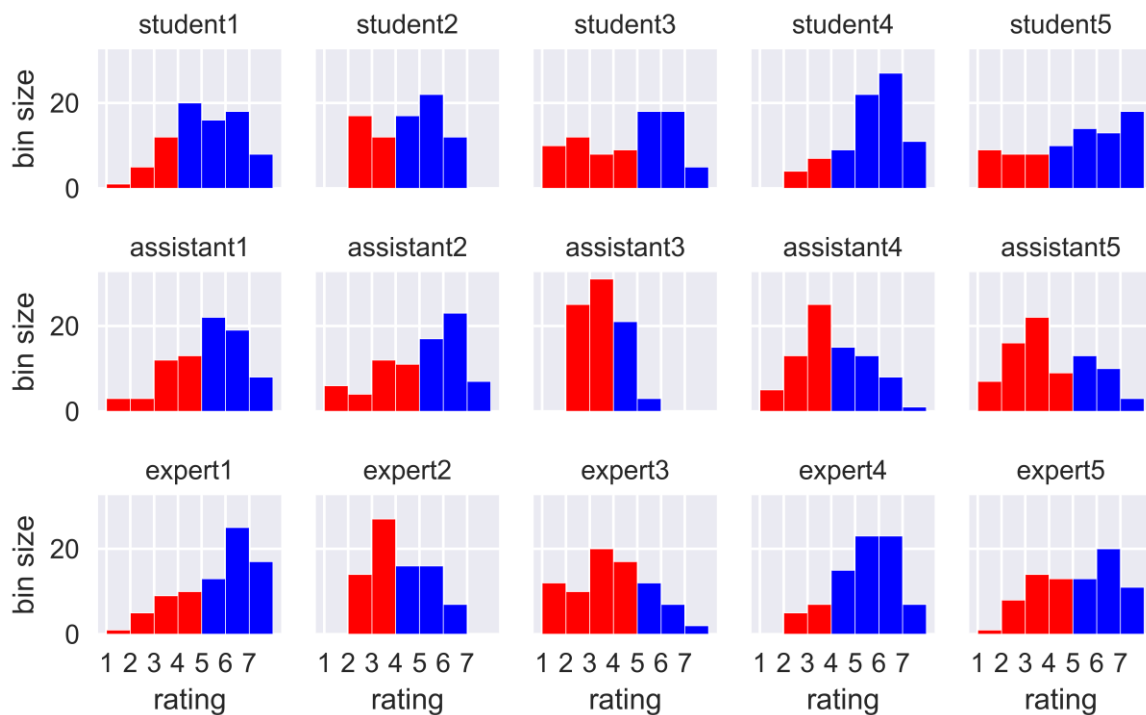


Figure 4.1.13.1: Individual rating analysis with individual cut off to translate the seven-level rating scale into a binary scale. Red bars are showing dissimilar letters, blue bars are representing similar ones.

About half of the participants have chosen a rating of 5 (7/15) as threshold for similarity and the other half have chosen a rating of 4 (8/15).

Figure 4.1.13.2 is showing the binary rating of each participant for the trials with the best fitting letters (rank1).

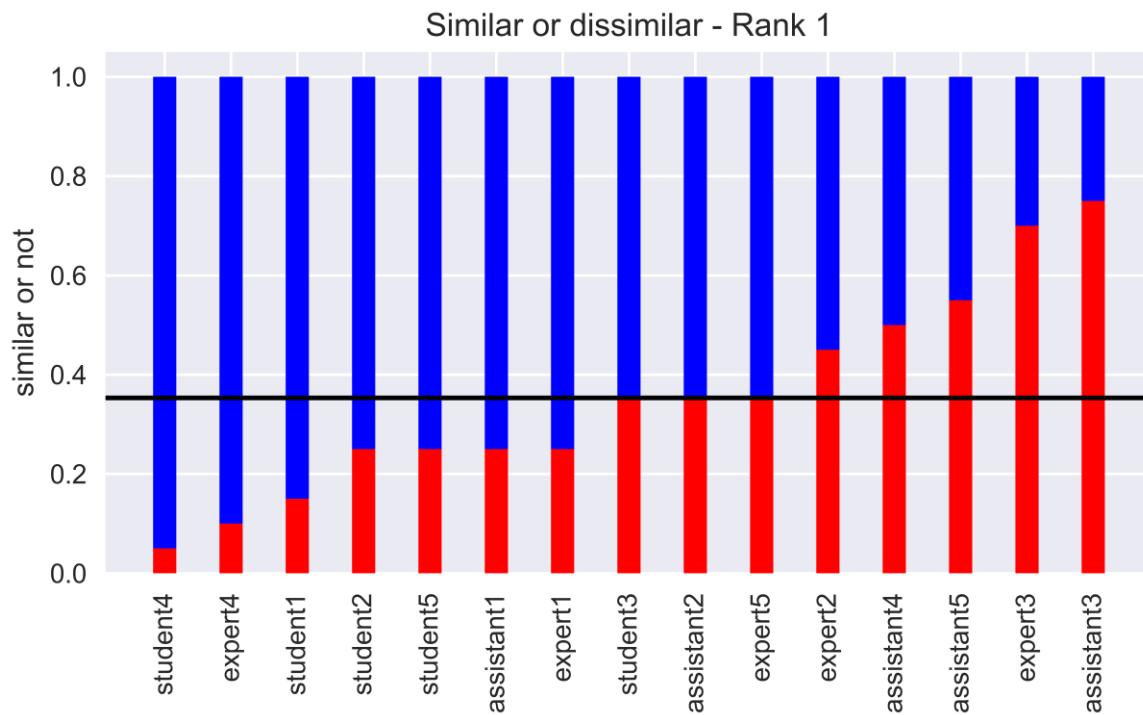


Figure 4.1.13.2: Normalized binary rating for the 20 most similar (rank1) letter-pairs. Blue bars are representing similar pairs, red bars dissimilar. The black line is the average binary rating.

Approximately two-thirds of the as most similar retrieved letters are similar. The average binary rating of Figure 4.1.13.2 for rank 1 is also shown in Figure 4.1.13.3, which is illustrating the average binary rating of all participants and of all experts for rank 1 to 4 and the random letter. It looks like that there are no differences between all participants and only experts.

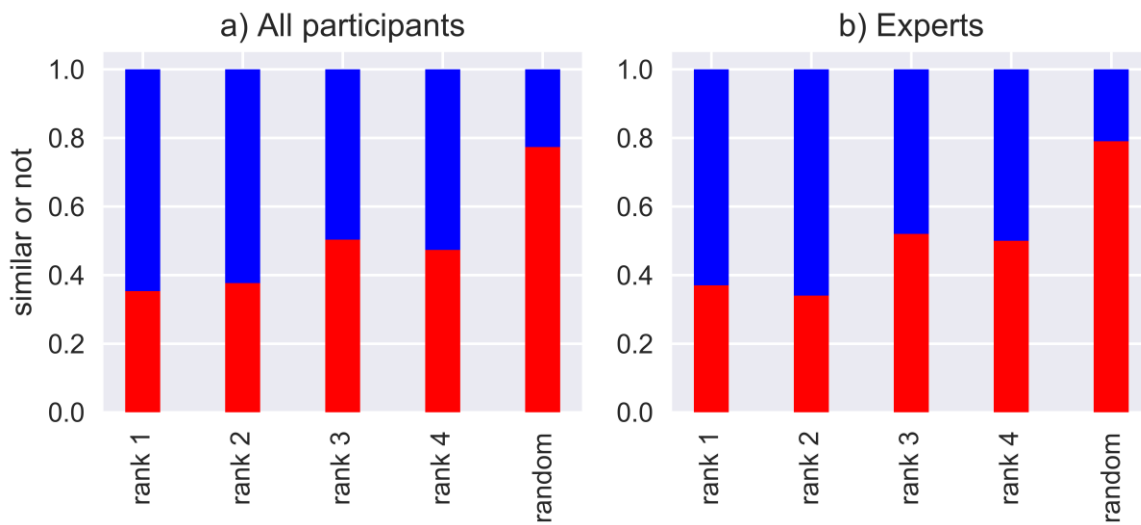


Figure 4.1.13.3: Average binary rating for each rank and the random letter. Similar letters are in blue, dissimilar ones in red. a) All participants b) Experts

It is interesting to know how much letters a user have to look up until retrieving at least one similar letter (Figure 4.1.13.4).

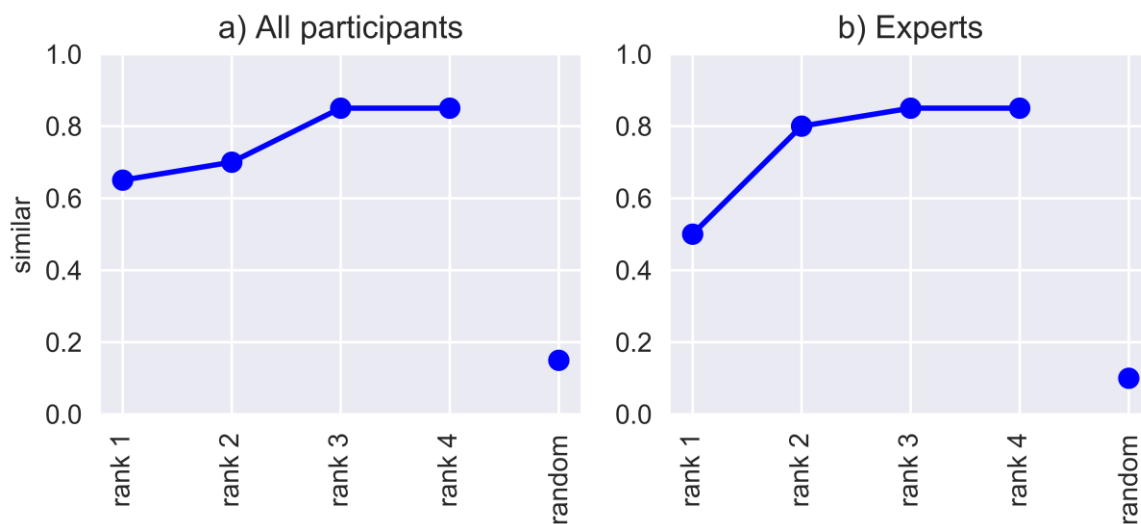


Figure 4.1.13.4: Probability to retrieve at least one similar letter after looking at letters, ranked one to four. The probability to find a similar letter by looking at a random letter is shown on the right. a) all participants b) experts

After looking at the best three retrieved letters the probability to find at least one similar letter is about 85%. A fourth letter doesn't seem to improve this probability.



Therefore, presenting the best three letters to the user seems to be enough. This finding goes well in line with the view of the participants, who mainly said that they would look at two or three letters until they expect to see a fitting patient.

The measurement of similarity is the cosine similarity, which is between 0 and 1. Consequently, the higher the computed similarity the more similar the retrieved letters should be. Figure 4.1.13.5 is showing the relationship between the cosine similarity and the binary rating for all participants and experts. Random letters are included for the analysis.

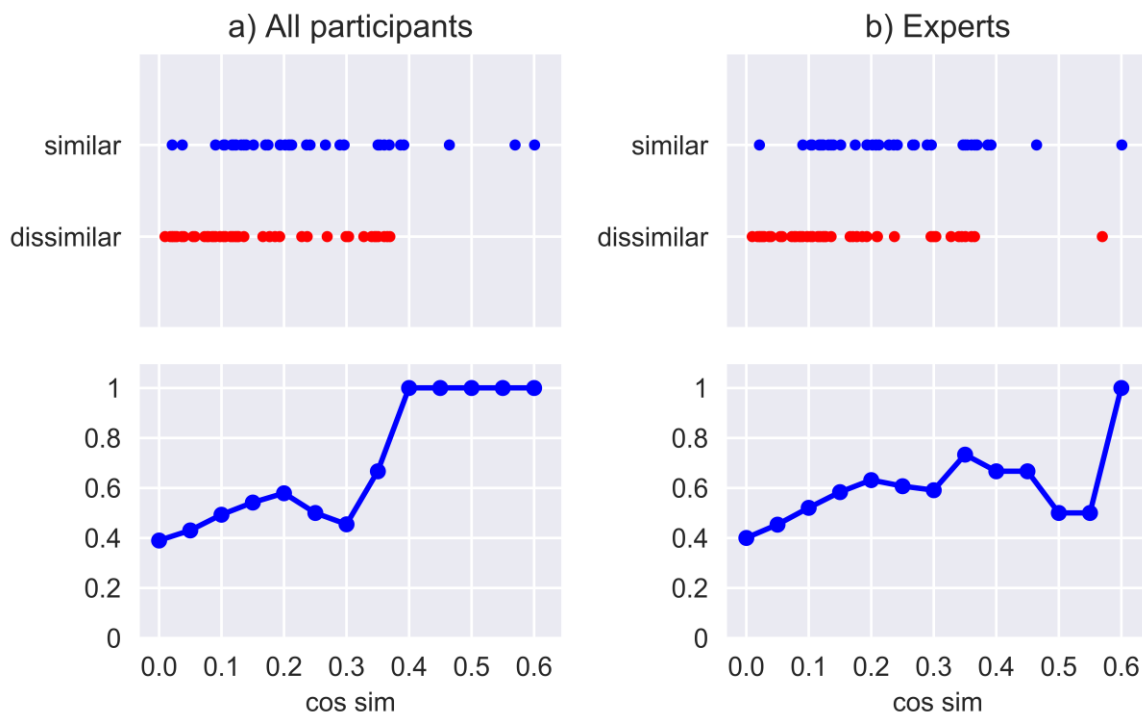


Figure 4.1.13.5: Probability of retrieving a similar letter plotted against the cosine similarity. Upper and lower plots share the same x-axis. The probability is the proportion between all similar letters and all letters, who are higher than a certain cos sim value. a) All participants b) Experts

As there are only little letter-pairs with a high cosine similarity, the probability for high cosine similarity values have to be seen critically. Especially the expert's plot indicates a possible relation between the computed cosine similarity and the probability of retrieving a similar letter.

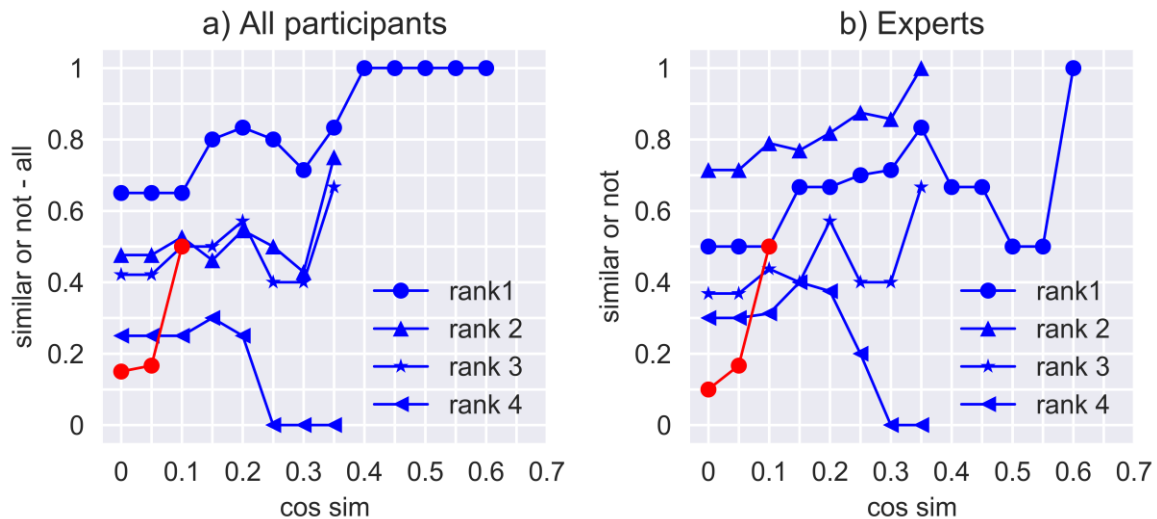


Figure 4.1.13.6: Probability of retrieving a similar letter plotted against the cosine similarity for each rank and random letter-pairs. Random letters are in red. The probability is the proportion between all similar letters and all letters, who are higher than a certain cos sim value. a) All participants b) Experts

Figure 4.1.13.6 is plotting the probability for retrieving a similar letter against the cosine similarity for each rank and random letter-pairs. For higher cosine similarity values the results have to be seen more carefully, because less letters meet the criteria. 5 letters of rank 4 have a higher cosine similarity than 0.25 and only two random letters are higher than 0.1. The plot indicates that a higher rank in combination with a higher computed cosine similarity is resulting in a higher probability of finding a similar letter.

Taken together, after looking at the best three retrieved letters the probability of finding at least one similar letter is about 85%. The recommender quality might be even better by only looking at letters with a certain cosine similarity.

## 4.2 Explorative Analysis

The system was developed to support physicians during their daily work life. It is unclear in which clinical situation it is the most useful. Experts probably use the program in another situation or context than junior doctors or students do. Likely, they





address different medical questions and therefore have different demands for improvement and modifications of the system. To tackle these issues, participants had to fill out the explorative questions from the above-described questionnaire. The results of the three groups are presented parallel and will be discussed afterwards in each section.

#### 4.2.1 Practical Usage

We asked participants if they could imagine working with the system practically. We used a five level rating scale (5 = “strongly agree” to 1 = “strongly disagree”). Additionally, participants had the possibility to describe more precise in which situation they could imagine using the program. Results are shown in Table 4.2.1.1

	Rating	Usage Situation
<b>Experts</b>	4	“Routine outpatient”
	4	“After usual therapy line”, “for complications”
	4	“Rare entities, aimed questions”
	4	“case-by-case decisions”
	5	
<b>Junior Doctors</b>	3	“Therapy decision, complications”
	3	“Unusual overall situation, comorbidities”
	3	“Actual unclear, if similarity is present”
	4	“To find comparable patients with extraordinary courses/ problems”, “retrieving previously seen patients”
	5	“Patients from extern hospitals with previous therapy”
<b>Students</b>	4	“Writing own discharge summaries, especially for patients with unknown clinical pictures” “How to treat a patients with a rare symptom pattern”
	4	“In similar, often occurring situations, e.g. follow-up in patients with watch and wait strategy”
	5	“Practical year and junior doctor residency”
	2	“For Patients with unusual courses”

Table 4.2.1.1: Practical Usage, Rating of different groups if they could imagine to work with the system. (5 level rating scale, 5 = “strongly agree” to 1 = “strongly disagree”)

In general it could be assumed that physicians and medical students can imagine working practically with the system, as the median overall rating was 4.



Potential users can imagine using such a program especially in uncommon patient constellations (“Unusual overall situation, comorbidities”, “Extraordinary courses/problems”, “Patients with rare symptom pattern”, “Patients with unusual courses”). Overall, participants consider the system as useful if the patient is more complex, unusual or exceptional. We are expecting, , even quite rare constellations or unique patients can be retrieved out of this database by integrating into a larger database.

#### 4.2.2 Usage for Medical Questions

Another question we asked our participants was for what medical issues such a program could be used. Statements are listed in Table 4.2.2.1.

Table 4.2.2.1: Statements of the three different groups about medical issues, which could be addressed

	<b>Usage situations for medical issues</b>
<b>Experts</b>	“Education and training for doctors” “After usual therapy line”, “for complications” “Combination with .... or search function” “Rare entities or unconventional courses”, “for comparison of similar therapies” “Case-by-case decisions”
<b>Junior Doctors</b>	“Therapy decision” “Setting of diagnosis” “Comparing patients with extraordinary problems/disease courses” “Comparing therapy standards at different times” “Therapy planning of different patient groups, especially with respect to previous therapy”
<b>Students</b>	“Treatment of patients with similar disease course” “Response to Therapy and influence to treatment of actual patient” “Similarities in patients treatment, which hasn’t been detected yet” “Comparison of on therapy decision with others” “Easier dictation at the end” “For therapy decision in difficult cases”, “To compare disease courses (not to forget potential risks)” “For patients with unusual courses”

with the program.



### 4.2.3 Categories for Similarity Judgement

Doctors and medical students rated the similarity between two patients discharge summaries by using a seven-level rating scale (1 = “very dissimilar” to 7 = “very similar”). Therefore, it remains unclear which categories participants use to assess the similarity. We asked the participants to list the categories, which they increasingly used for this task. The answers are listed in Table 4.2.3.1 Figure 4.2.3.1 is showing a clustermap of the applied categories of each participant and each group. We clustered the used the different participants depending on their used categories.

	Used Categories
<b>Student1</b>	Regression stadium, disease course, medicaments, stem cell transplantation, reason for first presentation, special symptoms, procedure
<b>Student2</b>	Therapy, medicaments, epicrisis, disease course, mutations, death, symptoms
<b>Student3</b>	Diagnosis outpatient vs. inpatient, karnofsky – index, date of first presentation, actual presentation, actual anamnesis, medicaments, epicrisis
<b>Student4</b>	Diagnosis, stadium, progress (especially Richter transformation and therapy indication), CLL therapy success, medicaments, primary diagnosis, risk factors
<b>Student5</b>	Stadium at diagnosis, stadium actual, immunophenotyping, duration from first diagnosis to first chemotherapy, medicaments, therapy changes, disease course, side effects of chemotherapy
<b>Assistant1</b>	Primary diagnosis (CLL+ other neoplasia, CLL+ Richter-Transformation), stadium at diagnosis, stadium actual, therapy course, chosen therapy, stem cell transplantation success/ failure, death
<b>Assistant2</b>	Previous illness history, previous hemato-oncological disease, stem cell transplantation, previous therapy, medicaments, epicrisis, clinical status
<b>Assistant3</b>	Disease course, comorbidities, chosen therapy, response to therapy, duration of therapy response, symptoms
<b>Assistant4</b>	Stadium at diagnosis, therapy course, chosen therapy, comorbidities, patients age, patients fitness (Karnofsky-Index), risk profile, therapy associated problems, disease associated problems, actual state of remission
<b>Assistant5</b>	Medicaments, therapy course, chosen therapy, epicrisis, medical history, future planning



<b>Expert1</b>	Diagnosis, therapy, clinical status, epicrisis
<b>Expert2</b>	Prognostic factors (Del 17q, mutations, Del 13q14, Zap70), stadium of CLL, therapy course, patients age, patients sex, death, stem-cell transplantation fam-allog./autolog Tx
<b>Expert3</b>	Patients age, cytogenetics, chosen therapy, special features (e.g. AIHA, splenectomie, therapy duration), risk factors
<b>Expert4</b>	Therapy course, moleculargenetics, patients age, complications, side effects, other tumour disease, previous illness history
<b>Expert5</b>	Disease factors, Richter- Syndrom,cytogenetics, mutations, significant therapy (e.g. allo-Tx, Ibrutinib), patients age

Table 4.2.3.1: Categories used by each participant

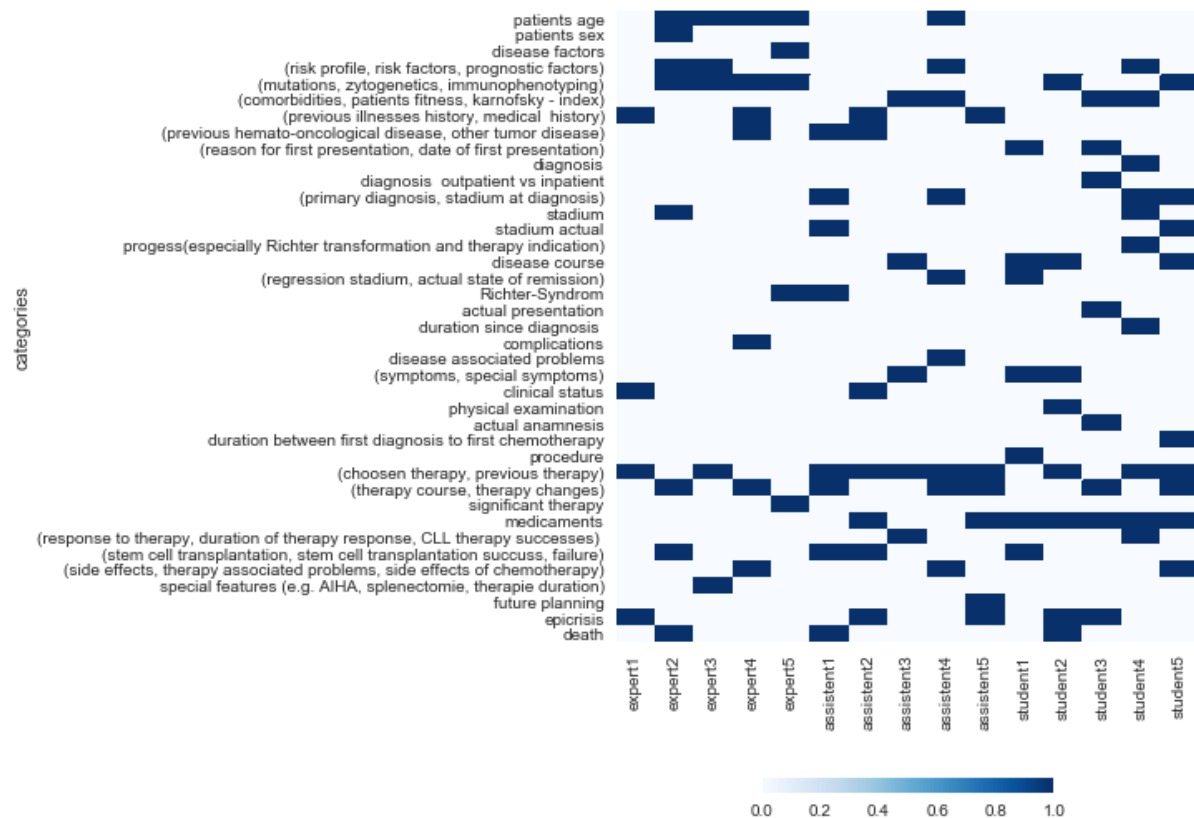


Figure 4.2.3.1: Heatmap of used categories for similarity judgement of each participant



#### 4.2.4 Potential for Improvement

The results above indicate that the actual program still has room for improvements, not only in terms of improving the retrieval process but also in terms of application. For this reason, we asked our participants for their ideas to improve the system or to propose some new features. Statements are listed in Table 4.2.4.1.

	Suggestions for Improvement
<b>Experts</b>	“Combination with other types of data”, “more patients, more diagnosis”, “combination with filter for key words” “Automatic ranking of accordance” “Narrowing with multiple keyword combinations”
<b>Junior Doctors</b>	“Categorization of characteristics (e.g. stadium)” “Creating patient fact sheet” “Better comparability in terms of letter structure” “Integrated search function, using a search word” “Program integration into digital medical record program”
<b>Students</b>	“Sorting of parts into a standardized format for better comparability” “Colour highlighting of common features” “Learning from difficulties/ mistakes of earlier treatments (compare Balint-groups)” “Letters should have a more similar format, to speed up the comparison, (e.g. tabulating)”

Table 4.2.4.1: Ideas for improvement and additional options for the program.



## 5 Discussion

The textual clinical recommender system we evaluated in this thesis presents a text mining based computerized decision support system (CDSS), which proposes similar patients by using their discharge summaries. To do so, it uses information retrieval and natural language processing methods. Using unstructured text data like patient discharge summaries has several advantages. Text documents are easily available, cheap and contain the most important information in a condensed form. Therefore, they should be suited to represent the entire patient. With our experiment we wanted to testify the assumption that the program retrieves similar patients in a way medical experts would agree with for a new dataset. Hummel et al. showed that the system is able to suggest appropriate patient cases for a dataset of 307 discharge summaries of patients suffering from different cancer entities. They used the same experimental structure as we did and found a computed (Spearman) correlation between the rating of four medical experts and the system of 0.39 (95% CI: [0.22, 0.56]).

We evaluated the system by using a more homogenous and larger dataset, containing 489 anonymized discharge letters of patients, all suffering from CLL. We considered medical experts (all haematologists) as the gold standard for the evaluation of the program. The Spearman correlation between the computed (cosine) similarity and the average rating of five experts is 0.512 (95% CI: [0.387,0.637]). This result confirms the significant correlation between the computed similarity and the expert's rating. Compared to the first evaluation from Hummel et al., our correlation is even higher. This is remarkable, as our dataset is more homogenous than the first one. This result suggests that the program is capable to differentiate subgroups of patients within a cohort of CLL patients. We had expected the program to have more issues distinguishing between patients with a shared cancer diagnosis than between patients with different cancer entities. However, the algorithm seems to be able to propose reasonable patients from a homogenous dataset. This homogenous dataset presents an even more valid setting since a user is likely to look for and compare similar patients with each other.



We calculated the inter-rater agreement by using the mean Spearman rank coefficient, also called Kendall's  $W$ , for each group. The inter-rater agreement for experts is 0.549, 0.527 for junior doctors and 0.594 for medical students. Students tend to agree stronger among each other. The only moderate agreement is possibly explainable with the participant's freedom to choose the criteria for the rating and the difficulty to distinguish within a homogeneous dataset. Under these circumstances the agreement is appropriate. Hummel et al. found an inter-rater agreement of 0.71, which is higher than the one we found, possibly explainable by the necessity to recognize and rate subtleties. The moderate inter-rater agreement should encourage modifying the experimental setup in the future. As another quality criteria we calculated Cronbach's alpha (tau-equivalent reliability) for the internal consistency of the experiment. Two letterpairs ("204-286" and "171-286") appeared twice during the experiment, first at position 35 and 57 and the second time at position 84 and 82. The Cronbach's alpha for the first double pair is 0.781, which can be described as an acceptable internal consistency and 0.570 (poor internal consistency) for the second. The low internal consistency may be due to the fact that the participants adjust their internal rating scheme over the course of the experiment. Another possible explanation might be the fact that the other comparison letters in each trial differ and participants make their decisions in relation to the other comparison letters.

Hummel et al. showed that the system acts above chance. They found a rating difference of 1.93 (95% CI: [1.17, 2.70]) between the best fitting and the random comparison letter. We could confirm the superiority above chance. In our experiment, the overall rating difference between the most similar retrieved letter, (ranked at first place according to the algorithm) and the randomly chosen letter was 1.76 (95%CI: [1.14,2.38 ])for experts, 1.45 (95%CI: [0.72, 2.18]) for assistants and 1.99 (95%CI: [1.11, 2.87]) for medical students.

We postulated that the correlation for medical experts is higher than for junior doctors and the correlation for junior doctors is higher than for medical students. The correlation for each group is: students: 0.439 [0.314,0.564], junior doctors 0.442 [0.317,0.567], experts 0.512 [0.387,0.637]. This result is illustrated in figure 4.1.12.1.



The figure demonstrates that there is no significant difference between the groups, but experts appear to correlate stronger with the system than novices do. To confirm the assumption that there is a higher correlation between experts and the algorithm than between novices and the algorithm, a higher number of participants would be needed. Although the results indicate that there is a slight difference between experts and novices in the way they rate similarity this question should be addressed in more detail in a future trial.

The experiment comprised 22 trials, each consisting of one reference letter and five related comparison letters. After dismissing the first two trials for familiarization reasons 20 trials remained. Ten trials were chosen randomly and ten were selected with respect to different criteria. Figure 4.1.8.1 shows the rating of the randomly chosen trials and the selected trials. The computed correlation of the average rating of all participants for the randomly chosen trials is 0.596 (95% CI: [0.471, 0.721]) and 0.440 (95% CI: [0.315, 0.565]) for the selected trials. A selection bias seems to be unlikely as the correlation for the randomly chosen trials is higher than the correlation for the selected trials.

During the data analysis we discovered some outliers. Letter pairs, which should be similar according to the algorithm, but which are rated as very dissimilar and vice versa letter pairs with a low computed similarity and a comparatively high participant rating. We manually examined these cases and tried to understand why the algorithm and the participants reached different results. We could not identify any obvious reasons for these differences. Studying the documents by 2 experts, the participants rating appeared more plausible. However, this difference in similarity rating between medical doctors and the SimRec system would not jeopardize its clinical usefulness, since the doctor would select the “useful” records and discard the other one, not considered helpful. Additionally, this problem illustrates that the algorithm works like a “black box” and the process is neither comprehensible nor approachable for the user. Even if people and algorithms effect the same performance it doesn't say that they acted the same way, which is a common issue in machine learning (Radermacher 2015).





The participants used a seven-level rating scale to rate the similarity between two discharge summaries. They were free to choose their own criteria for this task. Afterwards, we asked them, which categories they used. We illustrate the results in a heatmap (figure 4.2.3.1). Experts seem to focus more on pathobiological features and patient's characteristics. The chosen therapy, the therapy course and changes are important for all groups. This seems reasonable, as experts are familiar with the different therapy options and as a next step they can concentrate on the pathobiological features of a given patient. Since the preferred parameters differed between users, this might explain the moderate inter-rater agreement as every participant used a different category pattern for the rating task. We were aware of this issue before deciding to use the experiment. We wanted to display the clinicians "gut feeling" with this experimental structure, therefore we chose this general approach.

We used a questionnaire to get an idea what potential users think about the program or what modifications or improvements they would like to implement. Novices tend to use such a program to optimize therapy decisions, whereas experts are more focussed on rare patient constellations. Participants would like to restrict the results with a keyword or search function, like in a classic search engine, possibly because it is hard to understand the unusual way the program retrieves its results. Novices suggested presenting the text information in a more condensed form, like a cue sheet and in a more similar format to speed up the comparison process.

Taken together, the system suggests similar patients even when analyzing a homogeneous dataset. Regarding the way the system computes similarity it is reasonable to assume that the performance will further improve by including more patient letters in the database. Therefore, one next step would be to test the program with a larger dataset. This dataset could contain patient letters from other medical disciplines to collect overall more information about the capability of the algorithm. Another possible area of application would be to test the performance with other free text documents like textbooks or scientific literature. We envisage that the algorithm suggests a fitting patient to a specific chapter of a course book.



Due to the unspecific and diversified approach, the system can be used to answer a variety of questions. It might help making therapy decisions and answering diagnostic questions. The detection and presentation of similar patients is the key idea of how this program can improve the medical decision making process. First, presenting similar patients may help building different patient pattern, which can be used by System 1 for the decision making process. This is typically done through repeated experience, but could be accelerated by using the system (Croskerry 2013, 2009). Second, presenting similar patients might help considering alternatives. As a result the doctor is forced to rethink a patient's case or pausing a diagnostic process. Because of this double check, faults by System 1 could be avoided (Thammasitboon and Cutrer 2013). This holistic approach automatically implies disadvantages. Showing similar patient cases might lead to the so-called bias of *Representativeness* and errors in estimation the probability of disease (Elstein and Schwarz 2002; Kahneman 2011). Especially if a rare constellation is present, the comparison with other cases might distort the feeling for probability and can lead to an overestimation of probability. To avoid this issue an integrated reminder of the programs limitations and shortcomings would be desirable.

In the future we will ask doctors to use a prototype during their day-to-day working life. As with any CDSS, the recommender program has to be further evaluated before it can be integrated into clinical practice, as the effects of CDSSs on patient health often remain unstudied (Garg et al. 2005).

We need to come up with new experiments, which will answer the question, if the program can improve clinical care, patient's outcome or other end points, like finding the right diagnosis or consider new treatment plans. After a testing period the system can be integrated into the clinic's documentation system, where it can fall back to a large database. The recommended, similar patients can be automatically presented to a doctor, while writing a new letter or collecting information about a new patient. We are feeling certain that the recommender system will demonstrate its impact on medical decision-making.



## 6 Summary

The textual clinical recommender system we evaluated in this thesis represents a computerized clinical decision support system (CDSS). It analyses patient's discharge summaries with the help of information retrieval and natural language processing methods. It provides the user a similar patient case out of a database to include this information into the users decision-making process. We conducted an experiment to validate the correlation between the computed similarities by the new CDSS and the similarity judgement of medical experts, junior doctors and medical students. We expected that experts rate similarity between patients or rather their discharge summaries in a way that correlates with the computed similarity of the system. We assumed that the more experienced a participant is, the higher is the correlation. We created a new dataset of 489 CLL patient's discharge summaries for the verification. The participants rated the similarity of letter pairs, using a seven-level rating scale. Five participants per group (experts, junior doctors and students) took part. We could show that for this bigger, more homologues dataset the computed similarity correlates with the expert rating (Spearman correlation 0.512 (95% CI: [0.387, 0.637])). The experiment confirmed the already demonstrated superiority over chance (rating difference between best ranked letter and random letter 1.76 (95% CI: [1.14, 2.38] for experts). Furthermore we investigated differences between the three groups. The correlation was higher for experts than for assistants (0.443 (95% CI: [0.318,0.568])) and students (0.439 (95% CI: [0.314,0.564])), but no significant difference could be found. Additionally, we asked participants to fill in a questionnaire for explorative analysis to gather information about future application areas in working life or for medical issues, as well as possibilities for improvement.

Taken together, the retrieval system still needs improvements, either based on an improved retrieval algorithm or by additional features. However, it is likely that the systems performance will improve the more discharge summaries a database contains, like it was shown in this thesis. Our data suggest that the simrec software might indeed become an important clinical tool to share clinical experience between haematologists and possibly also other medical specialties.



## 7 Zusammenfassung

Das auf Textanalyse basierende Empfehlungssystem, das wir in dieser Arbeit evaluieren möchten, stellt ein computergestütztes System zur Unterstützung der medizinischen Entscheidungsfindung dar. Es analysiert Entlassbriefe von Patienten mit der Hilfe von Information Retrieval und Natural Language Processing Methoden. Dem Benutzer werden ähnliche Patientenfälle aus einer Datenbank angezeigt. Diese Information kann in den Entscheidungsfindungsprozess einbezogen werden. Wir führten ein Experiment zur Bestätigung der Korrelation zwischen der berechneten Ähnlichkeit des Systems und der Ähnlichkeitsbewertung von medizinischen Experten, Assistenzärzten und Medizinstudenten durch. Wir nahmen an, dass je erfahrener ein Teilnehmer ist, desto höher ist die Korrelation. Wir schufen einen neuen Datensatz von 489 Arztbriefen von an CLL erkrankten Patienten. Die Teilnehmer bewerteten die Ähnlichkeit von Briefpaaren und nutzen dazu eine sieben stufige Bewertungsskala. Fünf Teilnehmer pro Gruppe nahmen teil. Wir konnten zeigen, dass für diesen größeren, homologeren Datensatz die berechnete Ähnlichkeit mit der Expertenbewertung korreliert (Spearman Korrelation 0.512 (95% CI: [0.387, 0.637])). Das Experiment bestätigte die bereits gezeigte Überlegenheit des Systems gegenüber dem Zufall (Bewertungsunterschied zwischen ähnlichstem und zufälligem Brief von 1.76 (95% CI: [1.14, 2.38]) für Experten. Außerdem untersuchten wir Unterschiede zwischen den drei Gruppen. Die Korrelation war höher für Experten als für Assistenzärzte (0.443 (95% CI: [0.318,0.568])) und Studenten (0.439 (95% CI: [0.314,0.564])), es konnte aber kein signifikanter Unterschied gefunden werden. Zusätzlich baten wir die Teilnehmer einen Fragebogen für eine explorative Analyse auszufüllen, um Informationen über Einsatzmöglichkeiten im Arbeitsleben oder für medizinische Fragestellungen, aber auch für Verbesserungsmöglichkeiten zu sammeln. Insgesamt bedarf es für das Empfehlungssystem weiterer Verbesserungen. Entweder durch einen verbesserten Suchalgorithmus oder durch zusätzliche Funktionen. Es ist jedoch wahrscheinlich, dass die Leistung des Programms sich mit zunehmender Größe der Datenbank weiter verbessert, wie es durch diese Arbeit gezeigt werden konnte.



## 8 References

- Bergmann, Manuela, and Clemens-Martin Wendtner. n.d. 'Diagnostik Und Therapie Bei CLL Individualisiertes Vorgehen Zur Optimierung Des Behandlungserfolges'. 02.09.2016. <https://doi.org/10.1007/s15004-016-5206-2>.
- Blumenthal, David. 2010. 'Launching HITECH'. *New England Journal of Medicine* 362 (5): 382–85. <https://doi.org/10.1056/NEJMp0912825>.
- 'Cronbach's Alpha'. 2018. *Wikipedia*.  
[https://en.wikipedia.org/w/index.php?title=Cronbach%27s\\_alpha&oldid=832164637](https://en.wikipedia.org/w/index.php?title=Cronbach%27s_alpha&oldid=832164637).
- Croskerry, Pat. 2009. 'Clinical Cognition and Diagnostic Error: Applications of a Dual Process Model of Reasoning'. *Advances in Health Sciences Education* 14 (1): 27–35. <https://doi.org/10.1007/s10459-009-9182-2>.
- . 2013. 'From Mindless to Mindful Practice — Cognitive Bias and Clinical Decision Making'. *New England Journal of Medicine* 368 (26): 2445–48. <https://doi.org/10.1056/NEJMp1303712>.
- Dores, Graça M., William F. Anderson, Rochelle E. Curtis, Ola Landgren, Evgenia Ostroumova, Elizabeth C. Bluhm, Charles S. Rabkin, Susan S. Devesa, and Martha S. Linet. 2007. 'Chronic Lymphocytic Leukaemia and Small Lymphocytic Lymphoma: Overview of the Descriptive Epidemiology'. *British Journal of Haematology* 139 (5): 809–19. <https://doi.org/10.1111/j.1365-2141.2007.06856.x>.
- Elstein, Arthur S, and Alan Schwarz. 2002. 'Clinical Problem Solving and Diagnostic Decision Making: Selective Review of the Cognitive Literature'. *BMJ: British Medical Journal* 324 (7339): 729–32.
- Elstein, Arthur S., Lee S. Shulman, and Sarah A. Sprafka. 1990. 'Medical Problem Solving: A Ten-Year Retrospective'. *Evaluation & the Health Professions* 13 (1): 5–36. <https://doi.org/10.1177/016327879001300102>.
- Evans, Jonathan St B. T. 2008. 'Dual-Processing Accounts of Reasoning, Judgment, and Social Cognition'. *Annual Review of Psychology* 59: 255–78. <https://doi.org/10.1146/annurev.psych.59.103006.093629>.
- Fabbri, Giulia, and Riccardo Dalla-Favera. 2016. 'The Molecular Pathogenesis of Chronic Lymphocytic Leukaemia'. *Nature Reviews Cancer* 16 (3): 145–62.



<https://doi.org/10.1038/nrc.2016.8>.

Garg, Amit X., Neill K. J. Adhikari, Heather McDonald, M. Patricia Rosas-Arellano, P. J. Devereaux, Joseph Beyene, Justina Sam, and R. Brian Haynes. 2005. 'Effects of Computerized Clinical Decision Support Systems on Practitioner Performance and Patient Outcomes: A Systematic Review'. *JAMA* 293 (10): 1223–38.

<https://doi.org/10.1001/jama.293.10.1223>.

Hallek, Michael. 2017. 'Chronic Lymphocytic Leukemia: 2017 Update on Diagnosis, Risk Stratification, and Treatment'. *American Journal of Hematology* 92 (9): 946–65.

<https://doi.org/10.1002/ajh.24826>.

Hochstetter, Manuela. n.d. 'Chronische Lymphatische Leukämie Interaktive Fälle Zur Aktuellen DGHO-Leitlinie'.

Hummel, Philipp, Frank Jäkel, Sascha Lange, and Roland Mertelsmann. 2018. *A Textual Recommender System for Clinical Data: 26th International Conference, ICCBR 2018, Stockholm, Sweden, July 9-12, 2018, Proceedings*.

[https://doi.org/10.1007/978-3-030-01081-2\\_10](https://doi.org/10.1007/978-3-030-01081-2_10).

Kahneman, Daniel. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

'Matplotlib: Python Plotting — Matplotlib 2.2.2 Documentation'. n.d. Accessed 17 July 2018. <https://matplotlib.org/>.

'NumPy — NumPy'. n.d. Accessed 17 July 2018. <http://www.numpy.org/>.

Oers, Marinus H J van. 2016. 'Analysis of Prognosis in CLL: Collaboration Makes the Difference'. *The Lancet Oncology* 17 (6): 691–92. [https://doi.org/10.1016/S1470-2045\(16\)30052-3](https://doi.org/10.1016/S1470-2045(16)30052-3).

'Pandas.DataFrame.Corr — Pandas 0.23.3 Documentation'. n.d. Accessed 16 July 2018. <https://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.corr.html>.

'Python Data Analysis Library — Pandas: Python Data Analysis Library'. n.d. Accessed 17 July 2018. <https://pandas.pydata.org/>.

Radermacher, F. J. 2015. 'Algorithmen, maschinelle Intelligenz, Big Data'.

*Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz* 58 (8): 859–



65. <https://doi.org/10.1007/s00103-015-2189-3>.
- 'RANDOM.ORG - True Random Number Service'. n.d. Accessed 11 January 2018. <https://www.random.org/>.
- 'SciPy.Org — SciPy.Org'. n.d. Accessed 17 July 2018. <https://www.scipy.org/>.
- 'Scipy.Stats.Spearmanr — SciPy v0.14.0 Reference Guide'. n.d. Accessed 16 July 2018. <https://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.spearmanr.html>.
- 'Seaborn: Statistical Data Visualization — Seaborn 0.9.0 Documentation'. n.d. Accessed 17 July 2018. <https://seaborn.pydata.org/>.
- 'Spearman's Rank Correlation Coefficient'. 2018. *Wikipedia*. [https://en.wikipedia.org/w/index.php?title=Spearman%27s\\_rank\\_correlation\\_coefficient&oldid=841994032](https://en.wikipedia.org/w/index.php?title=Spearman%27s_rank_correlation_coefficient&oldid=841994032).
- Thammasitboon, Satid, and William B. Cutrer. 2013. 'Diagnostic Decision-Making and Strategies to Improve Diagnosis'. *Current Problems in Pediatric and Adolescent Health Care* 43 (9): 232–41. <https://doi.org/10.1016/j.cppeds.2013.07.003>.
- Wimsett, Jordon, Alana Harper, and Peter Jones. 2014. 'Review Article: Components of a Good Quality Discharge Summary: A Systematic Review'. *Emergency Medicine Australasia: EMA* 26 (5): 430–38. <https://doi.org/10.1111/1742-6723.12285>.

## 9 Table of Figures

- Fig 1.1: Medical decision-making model adapted from Crosskerry 2009
- Figure 3.1.1: Cosine similarity vs. average rating of medical experts (Andreas Hummel et al. 2018)
- Figure 3.1.1.1: User interface of the program, "Home" view
- Figure 4.1.3.1: Histograms of rating behaviour of each participant, including random letters. Central tendency bias can be seen in several participants, e.g. assistant3, expert2
- Figure 4.1.3.2: Rating behaviour of each group, individual ratings are stacked upon each other a) students b) assistants c) experts



Figure 4.1.5.1: Inter-rater agreement between each participant

Figure 4.1.6.1: Rating difference between the most similar letter, according to the algorithm and the randomly assigned letter, including the mean difference and the 95% confidence interval. a) Students mean rating difference 1.99 (95%CI: [1.11, 2.87]) b) junior doctors mean rating difference 1.45 (95%CI: [0.72, 2.18]) c) experts mean rating difference 1.76 (95%CI: [1.14,2.38])

Figure 4.1.6.2: Average rating versus the computed cosine similarity of all letter pairs. Letter pairs with random comparison letters are black. a) Students b) assistants c) experts

Figure 4.1.8.1: Cosine similarity versus average rating of all participants for a) random trials and b) selected trials

Figure 4.1.9.1: Striking trials are circled.

Figure 4.1.10.1: Individual rating versus the computed cosine similarity for each participant.

Figure 4.1.11.1: Mean rating of all fifteen participants versus computed cosine similarity.

Figure 4.1.12.1: Correlation between the average rating of each group and the computed similarity for each group with the related 95% Confidence intervals.

Figure 4.1.12.2: Correlation versus experience, development over time. Notice that the correlation for each group is computed with the average rating a) students b) assistants c) experts

Figure 4.1.12.3: Correlation versus experience; students correlation is at the starting point and represents the null value

Figure 4.1.12.4: Correlation versus result multiple choice test, correlations of average ratings of each group are marked in red a) students b) junior doctors c) experts

Figure 4.1.12.5: Correlation versus multiple choice test result

Figure 4.1.13.1: Individual rating analysis with individual cut off to translate the seven-level rating scale into a binary scale. Red bars are showing dissimilar letters, blue bars are representing similar ones.

Figure 4.1.13.2: Normalized binary rating for the 20 most similar (rank1) letter-pairs. Blue bars are representing similar pairs, red bars dissimilar. The black line is the average binary rating.





Figure 4.1.13.3: Average binary rating for each rank and the random letter. Similar letters are in blue, dissimilar ones in red. a) All participants b) Experts

Figure 4.1.13.4: Probability to retrieve at least one similar letter after looking at letters, ranked one to four. The probability to find a similar letter by looking at a random letter is shown on the right. a) all participants b) experts

Figure 4.1.13.5: Probability of retrieving a similar letter plotted against the cosine similarity. Upper and lower plots share the same x-axis. The probability is the proportion between all similar letters and all letters, who are higher than a certain cos sim value. a) All participants b) Experts

Figure 4.1.13.6: Probability of retrieving a similar letter plotted against the cosine similarity for each rank and random letter-pairs. Random letters are in red. The probability is the proportion between all similar letters and all letters, who are higher than a certain cos sim value. a) All participants b) Experts

Figure 4.2.3.1: Heatmap of used categories for similarity judgement of each participant



## 10 Table of Tables

Table 3.2.5.1: Training trials

Table 3.2.5.2: Selected reference letters

Table 3.2.5.3: Random reference letters

Table 4.1.1.1: Data computation for training pairs; computed cosine similarity in brackets

Table 4.1.1.2: Data computation for selected reference letters; computed cosine similarity in brackets

Table 4.1.1.3: Data computation for randomly chosen reference letters; computed cosine similarity in brackets

Table 4.1.2.1: Average participant's characteristics

Table 4.1.4.1: Repeatability: Internal consistency for double letter pairs using Cronbach's alpha

Table 4.1.5.1 Average inter-rater agreement

Table 4.1.6.1: Rating difference between the most similar letter, according to the algorithm and the randomly assigned letter

Table 4.1.7.1: Spearman correlation coefficient, Kendall's tau coefficient and Pearson correlation for experts, Pandas and SciPy provide the same results

Table 4.1.8.1: Correlation for random trials and selected trials.

Table 4.1.9.1: Striking trials

Table 4.1.10.1: Spearman correlation for each participant

Table 4.1.12.1: Correlation of each group

Table 4.2.1.1: Practical Usage, Rating of different groups if they could imagine to work with the system. (5 level rating scale, 5 = "strongly agree" to 1 = "strongly disagree")

Table 4.2.2.1: Statements of the three different groups about medical issues, which could be addressed with the program.

Table 4.2.3.1: Categories used by each participant

Table 4.2.4.1: Ideas for improvement and additional options for the program.



## 11 Acknowledgement

First, I would like to thank the initiators of the project: Prof. em. Dr. Drs. h.c. Roland Mertelsmann and Dr. Sascha Lange. I deeply appreciate that they gave me the opportunity to be part of the team. I thank Roland for his continuous support, intellectual input and his great encouragement. Furthermore, I would like to thank Dr. Reinhard Marks, who supervised my thesis and who was a permanent discussion partner.

I would like to express my sincere gratitude to Philipp Hummel and Prof. Frank Jäkel. Without their extraordinary assistance and support, this thesis couldn't be done.

I thank my girlfriend Laura and my friends for their advice, patience and encouraging words.

Last but not least, I would like to thank my family: my parents, my grandparents and my brother and my sister for supporting me not only throughout writing this thesis. They are always there for me and support me in every aspect of life.

Thank you very much, everyone!

---

## 12 Curriculum Vitae

Geboren am 08. Juli 1989 in Freiburg im Breisgau

### Kurzprofil

- Weiterbildung zum Facharzt für Anästhesiologie am Ortenauklinikum Offenburg – Gengenbach
- Studium der Humanmedizin an der Albert-Ludwigs-Universität Freiburg
- Auslandsstudium an der Université Claude Bernard in Lyon und am Imperial College in London
- ERASMUS-Stipendium der Europäischen Union



## 13 Appendix

### 13.1 Glossary

BoW	Bag of Words
CDSS	Computerized clinical decision support system
CLL	chronic lymphatic leukaemia
HITECH	Health Information Technology for Economic and Clinical Health
TF – IDF	Term Frequency – Inverse Document Frequency

### 13.1 Example 13.2 Example Letters, Striking Trials

Letter Number 171 and 417 are shown as example letters. They are also listed in section 4.1.9 as examples for striking trials.



Organomegalie palpabel. Lymphknoten: Cervikal, nuchal und supraklavikulär: keine Lymphknoten tastbar.

Diagnostik: Labor vom 24.05.2017: Hämato-logie: Leukozyten 10,32 Tsd/µl; Thrombozyten 199 Tsd/µl; Erythrozyten 4,17 Mio/µl; Hämoglobin 12,7 g/dl; Hämokrit 35,3 %; MCV 84,7 fl; MCH (HbE) 30,5 pg; MCHC 36,0 g/dl; RDW (Erv. Verteilungsbreite) 12,1 %; Retikulozyten 1,30 %; Reti abs. 54 Tsd/µl; Reti-Prod.-index 0,7; Retikulozyten- Hämoglobin 33,4 pg; Hypochrome Erythrozyten 0,1 %; Klinische Chemie: Natrium 140 mmol/l; Kalium 5,0 mmol/l; Calcium 2,31 mmol/l; Magnesium 0,86 mmol/L; Harnstoff 34 mg/dl; Kreatinin 1,01 mg/dl; GFR-Abschätzung(MDRD) 73 ml/min/1.73qm; CKD-EPI GFR geschätzt 75 ml/min/1.73qm; Harnsäure 6,8 mg/dl; Glukose 109 mg/dl; LDH 190 U/l; GPT (ALT) 35 U/l; Alk. Phosphatase 63 U/l; Gamma-GT 28 U/l; Bilirubin gesamt 0,5 mg/dl; C-reaktives Protein < 3 mg/l; Eiwweiß (gesamt) 6,7 g/dl; Immunoglobulin G 747 mg/dl; Differentialblutbild: Stabkernige 1 %; Neutrophile Segm. 16 %; Eosinophile 3 %; Monozyten 4 %; Lymphozyten 63 %; Atyp.Lymphphoz.vermutl.neoplas. 1 %; Kernschatten 12 %;

Epikrise: Die ambulante Vorstellung des Patienten erfolgte zur erneuten Verlaufskontrolle bei bekannter, 2007 erstdiagnostizierter B-CLL. Seit Erstdiagnose war bei stets konstanten hämatologischen Parametern eine watch & wait-Strategie verfolgt worden. Der Patient berichtete erfreulicherweise weiterhin von einem sehr guten Allgemeinbefinden; ebenso ergaben sich in der klinischen Untersuchung und laborchemisch keine Hinweise auf einen Progress der CLL, sodass wir eine Fortführung des abwartenden Vorgehens mit dem Patienten vereinbarten.

Ein unkomplizierten Verlauf vorausgesetzt, bitten wir um eine Wiedervorstellung in einem Jahr. Der Patient wird den Termin telefonisch vereinbaren.

Wir danken für die Vorstellung und verbleiben mit freundlichen kollegialen Grüßen

171 Sehr geehrter Herr Kollege, wir berichten Ihnen nachfolgend über o.g. Patienten, der sich am 24.05.2017 in unserer Ambulanz vorstellte.

Diagnosen: 1. B-CLL, ED 2007 a. Stadium RAI 0 b. IgVH-Status: mutiert, Deletion 13q14 günstige Prognose 2. 05/2010Rotatorenmanschettenruptur 3. Arterielle Hypertonie

Aktueller Remissionsstand: SD Aktueller Karnorsky-Index/Gewicht/Größe: 100%/168cm/78,5kg Impfstatus: 2015 Grippe, Pneumovax (Jahr nicht erinnernlich) Jetzt: Ambulante Verlaufskontrolle

Vorsorgeuntersuchungen: Regelmäßige kardiol. Vorstellung im Herzzentrum, Koloskopie 11/07

Verlauf und Therapie: erstmaliger Nachweis einer Leukozytose von 14 Tsd/µl Koloskopie im Rahmen der Krebsvorsorge, weiterhin Leukozytose Leukozytose 14,7 Tsd/µl, Lymphozytose Leukozytose 15,6 Tsd/µl, Lymphozytose Leukozytose 16,0 Tsd/µl, Lymphozytose Leukozyten 16,8 Tsd/µl, watch and wait Leukozyten 18,1 Tsd/µl, watch and wait Leukozyten 18,1 Tsd/µl, watch and wait Sturz auf Schulter rechts, zunehmende Beschwerden Rotatorenmanschettenruptur: große SSP-/ISP-Ruptur (Bateman 4, Patte 3) (R), Outlet-impingement-Syndrom der Schulter (R) arthroskopische subakromiale Dekompression und partielle Rekonstruktion der Rotatorenmanschette des Schultergelenkes. 2x seit-zu-seit-Naht (Fiberwire 2?) SSP-/ISP-Teilverschluss (R) Leukozyten 14,2 Tsd/µl, watch and wait Lc 14 Tsd/µl, Hb und Tc i.N., watch and wait Lc 16,65 Tsd/µl, Hb 14 g/dl, Tc 238 Tsd/µl, sehr guter AZ, watch and wait watch and wait Lc 14,28 Tsd/µl, Hb 13,8 g/dl, Tc 211 Tsd/µl, guter AZ, watch and wait Lc 11 Tsd/µl, Hb 13,3 g/dl, Tc 220 Tsd/µl, guter AZ, watch and wait Lc 10,3 Tsd/µl, Hb 12,7 g/dl, Tc 199 Tsd/µl, guter AZ, watch and wait

Aktuelle Anamnese: Der Patient berichtet von einem weiterhin sehr guten Allgemeinzustand, er fühle sich körperlich belastbar, sei aktiv. Es bestehe keine Infektneigung, keine Auffälligkeiten des Stuhlgangs und der Miktion. Bis auf eine kurze Nachtschweiß-Episode vor ca. 2 Wochen bestünden keine B-Symptome.

Bisherige Medikation: Blutdruckmedikation

Untersuchungsbefund: Patient in gutem Allgemeinzustand und leicht übergewichtigen Ernährungsstatus. Enoral blande. Haut: guter Turgor, keine Effloreszenzen, keine Blutungszeichen. Herz und Lunge auskultatorisch ohne pathologische Befunde. Abdomen weich, keine Resistenzen, keine



417

76m

Sehr geehrte Frau Kollegin, sehr geehrter Herr Kollege,

wir berichten Ihnen nachfolgend über o.g. Patienten, der sich am 19.01.2017 in unserer Ambulanz vorstellte.

**Diagnosen:**

1. **Chronische lymphatische Leukämie (B-CLL)** ED 04/08  
initial Stadium: RAI I  
Phänotyp: CD19+, CD20+, CD5+, CD11c+, CD23+, CD25+, CXCR4+, CD10-, CD38-, CD103-, FMC7-  
Risikofaktoren: ZAP-70 neg., Normalkaryotyp (SNP-Array); IgHV mutiert (2-5)
2. **Diabetes mellitus Typ II** ED 1986
3. **Akute vestibulocochleäre Störung** DD Endolymphhydrops re 05/14 Schlafapnoe-Syndrom
4. **Depression** ED 1960, Lithiumtherapie
5. **Arterielle Hypertonie**
6. **Z. n. multiplen Hörstörungen**

**Studienteilnahme:** Keine

**Verlaufsparameter:** Blutbild, Lymphknoten

**Aktueller Remissionsstatus:** SD

**Aktueller Karnofsky-Index/Gewicht/Größe:** 90%/72kg/170cm

**Jetzt:** Ambulante Verlaufskontrolle

**Verlauf und Therapie**

- 04/08 Zufallsbefund Leukozytose
- 05/08 Lc 10 Tsd/µl, Hb 12,3 g/dl, Tc 222 Tsd/µl, kein Therapiebedarf
- 09/08 CT-Abdomen: inguinal re LK 1,2 cm, ansonsten keine Lymphadenopathie.
- 03/09 Lc 10 Tsd/µl, Hb und Tc i.N.
- 09/09 Lc 9,8 Tsd/µl, Hb 12,7 g/dl, Tc 216 Tsd/µl
- 03/10 Lc 10,8 Tsd/µl, Hb und Tc im Normbereich
- 03/11 Lc 10,4 Tsd/µl, Hb, Tc i.N.: **SD**
- 09/11 Lc 11,16 Tsd/µl, Hb, Tc i.N.: **SD**
- 03/12 Lc 10,7 Tsd/µl, Hb, Tc, LDH im Normbereich: **SD**
- 09/12 BB: **SD**
- 04/13 BB: **SD**
- 09/13 Lc 10,6 Tsd/µl, Hb, Tc, LDH, Hapto i.N.: **SD**
- 04/14 Lc 11,8 Tsd/µl, Hb, Tc, LDH, Hapto i.N.: **SD**
- 10/14 Lc 12,4 Tsd/µl, Hb, Tc, LDH, Hapto i.N.: **SD**, Sono: o.p.B.
- 10/15 Lc 11,44 Tsd/µl, Tc 226 Tsd/µl; Hb 14,5 g/dl, LDH 258 U/l; **SD**, Sono: o.p.B.
- 04/16 Lc 13,32 Tsd/µl, Tc 254 Tsd/µl; Hb 13,6 g/dl, **SD**
- 01/17 Lc Tsd/µl; Tc Tsd/µl; Hb g/dl, **SD**

**Aktuelle Anamnese:** Körperlich Wohlbefinden, insbesondere keine B-Symptome, keine Lymph-adenopathie. Im Oktober respiratorischer Infekt. Seit November erneut schwere depressive Episode, wie zumeist in den Wintermonaten, psychiatrische Vorstellung Ende Januar geplant.

**Bisherige Medikation:** Ramipril, Insulin, Lithium (aktuell Spiegel im Zielbereich)

**Labor vom 19.01.17** Leukozyten 11,93 Tsd/µl; Thrombozyten 229 Tsd/µl; Erythrozyten 4,18 Mio/µl; Hämoglobin 13,1 g/dl; Hämokrit 38,7 %; MCV 92,6 fl; MCH (HbE) 31,3 pg; MCHC 33,9 g/dl; Hämolyse-Index (Serum) 8;

Natrium 140 mmol/l; Kalium 4,2 mmol/l; Haptoglobin 117 mg/dl; Harnstoff 27 mg/dl; Serum-Kreatinin 1,06 mg/dl; GFR-Abschätzung(MDRD) 68 ml/min/1,73qm; CKD-EPI GFR geschätzt 68 ml/min/1,73qm; Harnsäure 5,3 mg/dl; Glukose im Serum 163 mg/dl; LDH 172 U/l; GOT (AST) 23 U/l; GPT (ALT) 27 U/l; Alk. Phosphatase 77 U/l; Gamma-GT 16 U/l; Bilirubin gesamt 0,7 mg/dl; Immunglobulin G 643 mg/dl; Lithium 0,94 mmol/l

**Epikrise:** Die ambulante Vorstellung des Patienten erfolgte zur Verlaufskontrolle bei bekannter CLL, zuletzt langfristig ohne Therapiebedarf. Auch aktuell zeigten sich klinisch und labor-chemisch stabile Befunde, so dass wir, einen komplikationslosen Verlauf vorausgesetzt, einen Wiedervorstellungstermin für den 18.01.18 um 08:00Uhr im Lymphomzentrum vereinbarten. Bei Rückfragen und Problemen stehen wir unter o. g. Telefonnummer gerne zur Verfügung.

Mit freundlichen kollegialen Grüßen



### **13.3 Declaration of Consent, Multiple Choice Test, Questionnaire**



## Ähnlichkeitsexperiment Patient Similarity



In der medizinischen Entscheidungsfindung nutzen Ärzte ihre klinische Erfahrung, um einen aktuellen Patienten mit bereits früheren, ähnlichen Patienten zu vergleichen. Aus dieser Idee wurde ein Computer-System entwickelt, das es dem Arzt ermöglicht einen ähnlichen Patienten mit Hilfe der zugehörigen Arztbriefe aus einer Datenbank vorzuschlagen. Das nachfolgende Experiment dient der Evaluation dieses Programms in dem Arztbriefe paarweise nach ihrer Ähnlichkeit bewertet werden sollen.

Der Aufbau gliedert sich in folgende Abschnitte:

1. Multiple Choice Test
2. Ähnlichkeitsexperiment Patienten Similarity am Laptop
3. Fragebogen

Bei Rückfragen wenden Sie sich bitte zu jeder Zeit an den Versuchsleiter.

Wir danken Ihnen herzlichst, dass Sie sich bereit erklärt haben uns zu unterstützen!

Nicolas Woitzik

Probandendaten: ..... (Name, Vorname)

..... (Email)

Hiermit erkläre ich mich damit einverstanden, an dem oben beschriebenen Ähnlichkeitsexperiment teilzunehmen. Meine Teilnahme an der Untersuchung ist freiwillig. Ich bin darüber aufgeklärt, dass ich das Experiment zu jedem Zeitpunkt und ohne Angabe von Gründen abbrechen kann, ohne dass mir daraus Nachteile entstehen. Ich weiß, dass ich mich jederzeit mit Fragen zum Forschungsvorhaben an den/die Versuchsleiter/in wenden kann.

### **Einverständniserklärung zur Datenverarbeitung und Datenveröffentlichung**

Ich bin einverstanden, dass die im Laufe des Experimentes gewonnenen Versuchsdaten in anonymisierter Form für wissenschaftliche Auswertungen und Veröffentlichungen verwendet werden. Teilnehmernamen und -daten werden anonymisiert und sind nur den Versuchsleitern bekannt, wodurch Rückschlüsse durch Dritte ausgeschlossen werden sollen.

Dieses Einverständnis kann von mir jederzeit widerrufen werden.

Freiburg, den ..... (Unterschrift Teilnehmer/in)

Freiburg, den ..... (Unterschrift Versuchsleiter/in)





7 In einem Routineblutbild fällt bei einem 72-jährigen Patienten eine Erhöhung der Leukozyten auf 23.000/µl auf. Sie möchten nun feststellen, ob bei dem Patienten eine CLL vorliegt. Auf welche Untersuchung können Sie hierbei zunächst verzichten?

- Immunphänotypisierung
- Anamnese
- Differenzialblutbild
- Knochenmarkpunktion
- Körperliche Untersuchung

7 Welche Untersuchungen sind zur Stadieneinteilung der CLL nach Binet notwendig?

- Anamnese und multiparametrische Immunphänotypisierung
- Körperliche Untersuchung und Blutbild
- Blutbild und multiparametrische Immunphänotypisierung
- Körperliche Untersuchung und multiparametrische Immunphänotypisierung
- Anamnese und Blutbild

7 Sie haben bei einem Patienten den Verdacht auf eine maligne hämatologische Erkrankung und ziehen dabei differenzialdiagnostisch sowohl eine CLL als auch ein Mantelzelllymphom in Betracht. Welche Untersuchungsmethode eignet sich am besten zur Differenzierung der beiden Erkrankungen?

- Differenzialblutbild
- Knochenmarkhistologie
- Immunphänotypisierung
- Knochenmarkzytologie
- Lymphknotenhistologie

7 Sie stellen bei einer 67-jährigen Patientin die Diagnose CLL. Dabei erheben Sie folgende Befunde: Leukozyten 31.500/µl, 51 % Lymphozyten, Hämoglobin 12,2 mg/dl, Thrombozyten 121.000/µl, zervikale und axilläre Lymphadenopathie, keine  $\beta$ -Symptomatik, keine weiteren körperlichen Symptome. Welches weitere therapeutische Vorgehen ist bei dieser Patientin angezeigt?

- Therapie mit Fludarabin, Cyclophosphamid und Rituximab
- Therapie mit Idelalisib und Rituximab
- „Watch and wait“
- Therapie mit Ibrutinib
- „Best supportive care“

7 Welcher der folgenden Befunde stellt eine Indikation für einen Therapiebeginn dar?

- Thrombozyten 106.000/µl
- Hämoglobin 11,2 mg/dl
- Gewichtsverlust von 6 % in 6 Monaten
- Lymphozytenverdopplungszeit von 3 Monaten
- Asymptomatische axilläre Lymphadenopathie

7 Sie möchten bei einem 63-jährigen Patienten mit Erstdiagnose CLL eine Therapie beginnen. Der Patient befindet sich im Stadium Binet C mit einer progredienten Thrombopenie von 60.000/µl. Außer einer medikamentös gut eingestellten arteriellen Hypertonie hat der Patient keine Komorbiditäten (CIRS-Score 2). Welche Therapie ist bei fehlenden Kontrain-

dikationen laut DGHO-Leitlinie erste Wahl?

- Therapie mit Fludarabin, Cyclophosphamid und Rituximab
- Therapie mit Idelalisib, Ibrutinib und Rituximab
- Therapie mit Bendamustin und Rituximab
- Therapie mit Ibrutinib
- Therapie mit Obinutuzumab und Chlorambucil

7 Welches dieser Medikamente stellt eine zielgerichtete Therapie gegen CLL dar und greift in den  $\beta$ -Zell-Rezeptor-Signalweg ein?

- Fludarabin
- Venetoclax
- Ibrutinib
- Rituximab
- Bendamustin

7 Welcher der folgenden Faktoren spricht bei diagnostizierter CLL für eine günstige Prognose?

- TP53-Mutation
- Hohes Alter des Patienten
- Mutierter IGHV-Status
- del(17p13)
- CIRS 8 Punkte

7 Bei welchem der folgenden Patienten mit diagnostizierter CLL kann eine allogene Stammzelltransplantation als Therapieoption nach DGHO-Leitlinie am ehesten in Betracht gezogen werden?

- 64-jähriger Patient, Erstdiagnose CLL, Stadium Binet C, CIRS 3
- 73-jährige Patientin, Erstdiagnose CLL, Stadium Binet B, CIRS 7
- 81-jährige Patientin, viertes Rezidiv einer CLL, Binet C, CIRS 16
- 63-jähriger Patient, Frührezidiv einer CLL, Binet C, CIRS 4
- 67-jähriger Patient, Spätrezidiv einer CLL, Binet C, CIRS 7

7 Sie behandeln folgenden Patienten mit Rezidiv einer CLL: 68 Jahre alt, CIRS 12, del(17p13), Erstlinientherapie mit Bendamustin und Rituximab. Welche Therapieoption sollte hier nach DGHO-Leitlinie am ehesten in Betracht gezogen werden?

- Venetoclax
- Ibrutinib
- Fludarabin, Cyclophosphamid und Rituximab
- Bendamustin und Rituximab
- Ofatumumab und Chlorambucil





**1** Welche Untersuchungen sind zur Diagnosestellung und Stadieneinteilung der chronischen lymphatischen Leukämie (CLL) ausreichend?

- Körperliche Untersuchung, kleines Blutbild, Blutaussstrich
- Körperliche Untersuchung, Blutbild inklusive Differenzialblutbild, Blutaussstrich und Immunphänotypisierung des peripheren Blutes
- Körperliche Untersuchung, Blutbild inklusive Differenzialblutbild, Blutaussstrich, Immunphänotypisierung des peripheren Blutes und Zytogenetik der CLL-Zellen
- Blutbild inklusive Differenzialblutbild, Blutaussstrich, Immunphänotypisierung des peripheren Blutes und Computertomografie (CT) von Hals, Thorax und Abdomen
- Blutbild inklusive Differenzialblutbild, Blutaussstrich, Knochenmarkpunktion und CT von Hals, Thorax und Abdomen

**1** Welche Prognosefaktoren sind dem CLL-IPI („International Prognostic Index for patients with CLL“) zufolge bedeutend und lassen eine frühzeitige und nähere Prognoseabschätzung zu?

- Alter, Geschlecht, Lymphozytenverdopplungszeit, genetische Aberrationen (del17p, del11q)
- Alter, Geschlecht, ECOG-Performancestatus, Serumthymidinkinase
- TP53-Status, IGHV-Status, Serum- $\beta_2$ -Mikroglobulin, klinisches Stadium, Alter
- Komorbidität, Serumthymidinkinase, Alter, ECOG-Performancestatus, genetische Aberrationen (del17p, del11q)
- Alter, Geschlecht, Lymphommasse, genetische Aberrationen (del17p, del11q), TP53-Status

**1** Welche Aussage zur Wahl der Rezidivtherapie bei CLL-Patienten stimmt? Bei der Wahl der Rezidivtherapie ...

- spielen Komorbiditäten keine Rolle.
- muss die Qualität und Dauer des Ansprechens auf die Erstlinientherapie berücksichtigt werden.
- spielen Zytogenetik und Molekulargenetik keine Rolle.
- muss die absolute Lymphozytenzahl berücksichtigt werden.
- kommen besonders bei Spätrezidiven (> 2 Jahre) bevorzugt neue, zielgerichtete Substanzen zum Einsatz.

**1** Welche Aussage zu Therapieoptionen bei CLL-Patienten in der Erstlinie ist falsch?

- BR (Bendamustin/Rituximab) kann bei älteren, fitten Patienten eingesetzt werden.
- FCR (Fludarabin, Cyclophosphamid, Rituximab) ist die Standardtherapie bei jüngeren Patienten ohne signifikante Komorbidität.
- Chlorambucil kann in der Erstlinientherapie mit einem von drei verschiedenen anti-CD20-Antikörpern kombiniert werden.
- Die Kombinationen Rituximab-Chlorambucil, Ofatumumab-Chlorambucil und Obinutuzumab-Chlorambucil sind gleich effektiv.
- Nur unter der Erstinfusion von Obinutuzumab treten gehäuft infusionsassoziierte Reaktionen auf.

**1** Welche Therapieoption ist für die Erstlinie bei CLL-Patienten mit 17p-Deletion oder TP53-Mutation (Hochrisikopatienten) primär empfohlen?

- Idelalisib + Rituximab
- Ibrutinib
- Bendamustin + Rituximab
- FCR gefolgt von allogener Stammzelltransplantation
- Alemtuzumab

**1** Im Falle eines späten CLL-Rezidivs (> 2 Jahre) ohne Nachweis einer 17p-Deletion oder TP53-Mutation steht welche Therapieoption primär zur Verfügung?

- Chemoimmuntherapie gefolgt von allogener Stammzelltransplantation
- „watch and wait“
- FCR
- Wiederholung der vorangegangenen Therapie
- BR

**1** Für körperlich fitte CLL-Patienten mit einem Frührezidiv (< 2 Jahre) und einer 17p-Deletion oder TP53-Mutation wird welche Therapieoption empfohlen?

- Ibrutinib oder Idelalisib plus Rituximab
- „watch and wait“
- FCR
- Wiederholung der vorangegangenen Therapie
- BR

**1** Welche Kriterien stellen bei CLL-Patienten eine Behandlungsindikation dar?

- Jedes Stadium bei Vorliegen einer Deletion 17p oder einer TP53-Mutation
- Stadium Binet B mit grenzwertiger Thrombopenie ( $110.000/\mu\text{l}$ )
- Stadium Binet B ohne B-Symptome
- Fortgeschrittenes Stadium Binet C, Binet B und A mit B-Symptomen und/oder sehr kurzer Lymphozytenverdopplungszeit
- Rezidivierende Infektionen

**1** Welche Untersuchung muss bei CLL-Patienten vor jedem neuen Therapiebeginn erfolgen?

- Zyto- und Molekulargenetik
- Positronenemissionstomografie (PET)/CT
- CT-Hals, CT-Thorax, CT-Abdomen
- EKG
- Echokardiografie

**1** Für körperlich kompromittierte CLL-Patienten mit einem Frührezidiv (< 2 Jahre) und einer 17p-Deletion oder TP53-Mutation wird welche Therapieoption empfohlen?

- Wiederholung der Erstlinientherapie
- „watch and wait“
- FCR
- Ibrutinib oder Idelalisib plus Rituximab
- BR



Gerne würden wir Ihnen nun ein paar Fragen zum Experiment und zum System stellen.

**1. Angenommen Sie müssten sich nur zwischen den beiden Möglichkeiten entscheiden ein Brief sei **ÄHNLICH** oder **NICHT-ÄHNLICH**. Ab welchem Ähnlichkeitswert (1 = sehr unähnlich bis 7 = sehr ähnlich) würden Sie sagen, dass ein Brief **ÄHNLICH** ist.**

**2. Könnten Sie sich vorstellen mit diesem System praktisch zu arbeiten?**

Trifft zu  Trifft eher zu  neutral  trifft eher nicht zu  trifft nicht zu

In welcher Situation?

**3. Welche Kategorien haben Sie (verstärkt) benutzt, um die Ähnlichkeit der Patienten zu bewerten? Bitte listen Sie auf.**

**4. Für welche medizinische Fragestellung würden Sie ein solches Programm verwenden?**



**5. Glauben Sie ein derartiges Programm kann die medizinische Entscheidungsfindung verbessern?**

Trifft zu  Trifft eher zu  neutral  trifft eher nicht zu  trifft nicht zu

**6. Wie viele Briefe würden Sie sich anschauen, bis Sie einen passenden Patienten gefunden haben? Anders formuliert, nach wie vielen Briefen sollte ein „Treffer“ auftauchen?**

**7. Wo sehen Sie Verbesserungsmöglichkeiten? Welche zusätzlichen Optionen würden Sie sich für das Programm wünschen?**

Abschließend noch eine Frage zu Ihrer Person:

Für Ärzte:

Sind Sie Facharzt für Hämatologie/Onkologie? Ja  Nein

Seit wie vielen Jahren sind Sie praktisch tätig?

Für Studenten:

Seit wie vielen Semestern studieren Sie Medizin?

Haben Sie die M2 abgelegt? Ja  Nein

Wir bedanken uns sehr herzlich bei Ihnen für Ihre Teilnahme!



## **4 Eidstattliche Versicherung**



**Anlage 2**

Zum Antrag auf Zulassung zur Promotion

Zum Dr. med.  
(med. / med.dent.)

Woitzik, Nicolas Frank Philipp  
(Name) (Vorname)

**Eidesstattliche Versicherung**

gemäß § 8 Absatz 1 Nr. 3 der Promotionsordnung der Universität Freiburg für die Medizinische Fakultät

1. Bei der eingereichten Dissertation zu dem Thema

Machine Learning as an Adjunct to Medical Decision Making

handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht. Niemand hat von mir unmittelbar oder mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

3. Die Ordnung der Albert-Ludwigs-Universität zur Sicherung der Redlichkeit in der Wissenschaft habe ich zur Kenntnis genommen und akzeptiert

4. Die Dissertation oder Teile davon habe ich  
(Zutreffendes bitte ankreuzen)

bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

wie folgt an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt:

Titel der andernorts vorgelegten Arbeit:

\_\_\_\_\_

Name der betreffenden Hochschule:

\_\_\_\_\_

Jahr der Vorlage der Arbeit:

\_\_\_\_\_

Art der Prüfungs- oder Qualifikationsleistung:


\_\_\_\_\_

5. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

6. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

**Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.**

28.11.18  
Ort und Datum

  
Unterschrift