The objectives-based logbook
Raghoebar-Krieger, Helga Maria Josette

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2001

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
The reliability of logbook data of medical students: an estimation of interobserver agreement, sensitivity and specificity

H.M.J. Raghoeb-Krieger, D.Th. Sleijfer, W. Bender, R.E. Stewart, R. Popping

4Centre for Medical Education, Faculty of Medical Sciences, University of Groningen, Groningen, The Netherlands
5Department of Internal Medicine, University Hospital Groningen, Groningen, The Netherlands
6Northern Centre for Healthcare Research, Faculty of Medical Sciences, University of Groningen, Groningen, The Netherlands
7Department of Sociology, Faculty of Psychology, Education and Sociology, University of Groningen, Groningen, The Netherlands

Medical Education 2001;35:624-631.
Abstract

**Objective** Logbooks are widely used in medical schools as an evaluation tool to assess students’ progress towards objectives. To estimate whether students fill in their logbooks reliably, we measured interobserver agreement by comparing doctors’ data and students’ data.

**Method** Completed logbooks were collected at two subdivisions of the department of Internal Medicine at the University Hospital Groningen. The logbook contains 231 preprinted diseases. Doctors and students recorded the diseases they had encountered. Interobserver agreement, expressed by the coefficient of Jaccard (J), was calculated for the complete set of diseases and for a subset of core diseases. To assess the kind of errors which students made, sensitivity and specificity were determined.

**Results** Logbook data of doctors and students are not fully consistent (mean J for the complete set of diseases .23 and for the core diseases .36). The quality of the logbook data is high in the sense that students do not record many false identifications (mean specificity for the complete set of diseases and for the core diseases is .96 and .93, respectively); the quality is poor in the sense that students do not record all the diseases available at the department (mean sensitivity for the complete set of diseases is .36 and for the core diseases .51).

**Conclusion** This study shows inconsistencies in recording diseases in a logbook by students compared with doctors. In particular, the diseases which are present at a department are underreported by students. Supervision and feedback are important mechanisms to optimize the students’ use of (1) all diseases which could be encountered and (2) the logbook.
Introduction

A concern in medical education is whether students encounter during the internships the clinical experiences required by the teaching programme. A log system can shed light on this aspect as an evaluation tool for the student and the teacher as well as the medical school. A log system is used to monitor the clinical content of medical students’ experiences. Therefore, it is an instrument which provides feedback to students’ about their progress, allowing them to correct weaknesses, and assuring them about what has been accomplished. A logbook is also used to evaluate the adequacy of the instructional programme of the medical school. Although log systems have been used widely, few studies have been performed to determine the accuracy of the documentation of students’ experiences in logbooks. It seems difficult to create log systems which are simple to use, and which yield reliable and valid data.

At the Faculty of Medical Sciences of the University of Groningen, in the Netherlands, a logbook has been developed based on national requirements described in the Blueprint 1994. The Blueprint was developed by the eight Medical Faculties in the Netherlands. In this document objectives are set which medical students should achieve by the time of graduation from medical school. In Groningen, the objectives of the Blueprint specific to the several medical disciplines were adapted and turned into a logbook for students to be applied during the internships. The logbook describes the diseases that students must have seen, and students complete the logbook by recording the diseases which they have encountered. In this way, the logbook serves both as a guideline and as an evaluation tool for internships.

An appropriate logbook evaluation system implies several requirements such as feasibility, reliability and acceptability of logbooks. The reliability of logbook data is important because they are used to evaluate students’ experiences, and inconsistent scores lead to inaccurate results and therefore to incorrect notions about the learning progress of students, and about the teaching programme. Hence, the success of the logbook as an evaluation instrument depends on the production of reliable data by the students. A logical step therefore, is to investigate the reliability of the data gathered from students’
logbooks. The central question in our study was: ‘To what extent do students fill in their logbooks reliably?’ To answer this question, we measured interobserver agreement by determining whether different observers, using the same logbook at the same department at the same time, record the same data (diseases). The observers were experienced medical doctors on the one hand and students on the other hand. The logbooks completed by the doctors were used as a standard. Moreover, if there was disagreement between doctor and student we determine what kind of errors caused the disagreements. The study took place at the department of Internal Medicine at the University Hospital Groningen.

Method

Instrument
The logbook used is an adaptation of the objectives of the Blueprint, related to the discipline Internal Medicine, as has been described previously. In short, it indicates among other things which diseases students should learn about during their internship at the department of Internal Medicine. The diseases are grouped on the basis of symptomatology, pathogenesis, and location, and are hierarchically structured in clusters of illness (e.g. haematological disorders). There are 13 clusters in total and each cluster is composed of several subclusters (e.g. anaemia). Each subcluster consists of a number of diseases (e.g. iron-deficiency anemia). The diseases within a cluster are based on data with respect to incidence, prevalence and representativeness. The complete set of diseases in the logbook numbers 231 of which 37 are labelled as core diseases which it is essential that students should learn about. An example of part of the logbook is given in Table 1. Students register in the logbook the number and types of diseases they encounter during the internship.

Procedure
A common index for estimating the reliability of data collected in observational studies is the interobserver agreement. To determine interobserver agreement, logbook data entered by doctors and students were compared. This was done at two subdivisions of the Department of Internal Medicine, namely those of
Nephrology and Respiratory Medicine, where internships take place over a four-week period. Data collection took place in the period from October 1998 till February 1999 (Nephrology) and from March 1999 till May 1999 (Respiratory Medicine).

All students who started their internship at one of these departments were requested by the internship coordinator to register daily in their logbook the number and type of diseases they had encountered. The coordinator gave verbal instructions to write down the initial diagnosis of the patient, including all concomitant diseases. For instance, when a patient has been admitted because of iron-deficiency anemia, but also had hypertension and diabetes mellitus, all three diseases were registered. This explanation is also found on the first page of the logbook. The student is expected to interview all available patients and to do a physical examination and make a medical record. Supervision and feedback can be given on the students’ request. Moreover, feedback is available during the daily ward rounds as well as during the weekly discussion with the staff members. Furthermore, the written medical history made by the doctor, is also available for the student. Once a week the logbooks were collected just before a formal teaching hour, copied and returned to the students.
Two experienced medical doctors participated in this study. Both doctors were working full-time, one at the Nephrology subdivision and the other at the Respiratory Medicine subdivision. Their instructions with respect to the recording into the logbook were identical to the students’ instructions.

The Nephrology subdivision numbers 32 beds, and that of Respiratory Medicine 27 beds. During one week the doctor sees all patients who are admitted into his or her subdivision, and although the student has other educational responsibilities outside the subdivision, he or she is obliged to also observe all patients during the daily morning rounds. This means that the doctor and the student see the same patients during one week. Only these patients’ diseases are recorded and are included in the analysis.

The data of students were matched with data of doctors when both observers collected the data in the same week and in the same subdivision. Because some of the data were incomplete (because of illness, vacations, conference visits and so on), we could not match all the data gathered. After removal of incomplete data, we were able to use the data of 12 students. Because students may stay for more than one week in the same subdivision, 20 complete week sets were available to determine interobserver agreement. Moreover, we could also compare the datasets of four pairs of students, who completed their logs during the same week on the same subdivision.

**Analysis**

We analysed agreement about the diseases observed by the doctor and the student, and considered the observations of the doctor as the standard. Within the diseases a distinction is made between the complete set of diseases and a part of this set, the core diseases. We compared these two groups of diseases over pairs of columns in a data matrix where the outcome can be summarized in a 2 x 2 frequency table such as that shown in Table 2.

The left upper quadrant of the table (cell A), contains the number of diseases coded 1 (observed) by the doctor as well as by the student (correct positive, \( c_p \)). Two quadrants register the number of diseases for which the doctor and the student disagree, being coded 1 by the doctor and 0 (not observed) by the student (cell C, false negative, \( f_n \)), or the converse, coded 1 by the student and 0 by the doctor (cell B, false positive, \( f_p \)). Cell D contains the number of the diseases which both doctor and student did not register. We
consider these diseases as correct negative ($c_n$). We assume that these diseases are not observed. The marginal totals are the sums of these frequencies ($s_p$, $s_n$, $d_p$, $d_n$), and equal the number of the complete set of diseases (231), or the core diseases (37).

Table 2  Frequencies of judgements on the total number of diseases, according to observations by doctor and student

<table>
<thead>
<tr>
<th></th>
<th>according to doctor (gold standard)</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>disease observed 1</td>
<td>disease not observed 0</td>
<td>sum</td>
</tr>
<tr>
<td>disease observed 1</td>
<td>cell A correct positive $c_p$</td>
<td>cell B false positive $f_p$</td>
<td>$s_p$</td>
</tr>
<tr>
<td>disease not observed 0</td>
<td>cell C false negative $f_n$</td>
<td>cell D correct negative $c_n$</td>
<td>$s_n$</td>
</tr>
<tr>
<td>sum</td>
<td>$d_p$</td>
<td>$d_n$</td>
<td>$t_n$</td>
</tr>
</tbody>
</table>

$c_p$: number of diseases observed according to both student and doctor.

$f_p$: number of diseases observed according to student, but not observed according to doctor.

$s_p$: number of diseases observed according to student.

$f_n$: number of diseases not observed according to student, but observed according to doctor.

$c_n$: number of diseases not observed according to both student and doctor.

$s_n$: number of diseases not observed according to student.

$d_p$: number of diseases observed according to doctor.

$d_n$: number of diseases not observed according to doctor.

$t_n$: total number of diseases.

The Jaccard coefficient is used to measure interobserver agreement.\textsuperscript{13-15} This coefficient is a measure of reliability at the item level; it varies between 0 and 1. In this coefficient the number of items (diseases) observed by both doctor and student (correct positives) is divided by the sum of the number of items observed by both doctor and student, and the items observed by the doctor and not by the student (false negatives) and the items observed by the student and not by the doctor (false positives). Hence, items that were not observed at all are not taken into account. In this study this ratio was used because it was expected that considering the character of the logbook with 231
items, a relatively large number of items could not be observed at all in one week. Moreover as the logbook yields information about the students’ learning experiences with the diseases during the internship, it is more interesting to know about the reliability of experiences with diseases that are present, than with diseases that are not present. These arguments lead to the choice of the J coefficient instead of the more widely known Kappa coefficient of agreement at the item level\textsuperscript{16} which considers items not observed by two observers also as an agreement. Additionally, the J statistic does not take into account the proportion of cases in which agreement will take place by chance alone, while the Kappa coefficient does.

Sensitivity and specificity scores were calculated to define, respectively, the missed events (false negatives) and the false identifications (false positives). The proportion of false negatives equals 1 minus the sensitivity; the proportion of false positives equals 1 minus the specificity. Sensitivity is the proportion of diseases observed by the doctor and registered as such by the student; it is calculated by dividing the correct positive ($c_p$) by the sum of correct positives ($c_p$) and false negatives ($f_n$) (see Table 2). Specificity refers to the proportion of diseases not observed by the doctor and registered as such by the student. Specificity is calculated by dividing the correct negatives ($c_n$) by the sum of correct negatives ($c_n$) and false positives ($f_p$).

Based on all the diseases that both doctor and student had observed during 20 week sets (sum of correct positives), an analysis was made to evaluate whether doctor and student differed in the frequencies they had observed the diseases.

To gain insight in the differences in logbook data between students, a coefficient of agreement was also calculated on four pairs of students who were in the same subdivision during the same week.

**Results**

Table 3 shows the coefficient of agreement ($\textit{J}$), the sensitivity ($c_p/d_p$), and the specificity ($c_n/d_n$) for 20 pairs of one doctor and one student on the complete set of diseases and the core diseases. The \textit{J} values suggest low agreement between doctor and student: for the complete set of diseases the range is .08 to .43,
mean: .23. For the core diseases the range is .02 to 1.00, mean: .36. Interobserver agreement was greater for the core diseases than for the complete set of diseases in 13 pairs. In five pairs, the agreement was higher for the complete set of diseases than for the core diseases, and in two pairs the agreement was equal.

Table 3  The reliability of logbook data, expressed in the coefficient of agreement (J), sensitivity (c/p/dp) and specificity (c/n/dn) for 20 pairs of doctor and student by week

<table>
<thead>
<tr>
<th></th>
<th>complete set of diseases</th>
<th></th>
<th>core diseases</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>J</td>
<td>c/p/dp</td>
<td>c/n/dn</td>
<td>J</td>
</tr>
<tr>
<td>doctor – student 1</td>
<td>week 0</td>
<td>.36</td>
<td>.57</td>
<td>.94</td>
</tr>
<tr>
<td>doctor – student 1</td>
<td>week 1</td>
<td>.11</td>
<td>.33</td>
<td>.97</td>
</tr>
<tr>
<td>doctor – student 2</td>
<td>week 4</td>
<td>.08</td>
<td>.10</td>
<td>.98</td>
</tr>
<tr>
<td>doctor – student 2</td>
<td>week 5</td>
<td>.16</td>
<td>.19</td>
<td>.97</td>
</tr>
<tr>
<td>doctor – student 3</td>
<td>week 8</td>
<td>.32</td>
<td>.38</td>
<td>.98</td>
</tr>
<tr>
<td>doctor – student 4</td>
<td>week 14</td>
<td>.43</td>
<td>.50</td>
<td>.98</td>
</tr>
<tr>
<td>doctor – student 4</td>
<td>week 15</td>
<td>.37</td>
<td>.40</td>
<td>.99</td>
</tr>
<tr>
<td>doctor – student 5</td>
<td>week 18</td>
<td>.29</td>
<td>.32</td>
<td>.98</td>
</tr>
<tr>
<td>doctor – student 5</td>
<td>week 19</td>
<td>.20</td>
<td>.23</td>
<td>.99</td>
</tr>
<tr>
<td>doctor – student 5</td>
<td>week 20</td>
<td>.28</td>
<td>.37</td>
<td>.97</td>
</tr>
<tr>
<td>doctor – student 6</td>
<td>week 19</td>
<td>.31</td>
<td>.36</td>
<td>.98</td>
</tr>
<tr>
<td>doctor – student 7</td>
<td>week 23</td>
<td>.25</td>
<td>.25</td>
<td>1.00</td>
</tr>
<tr>
<td>doctor – student 8</td>
<td>week 26</td>
<td>.29</td>
<td>.60</td>
<td>.93</td>
</tr>
<tr>
<td>doctor – student 9</td>
<td>week 25</td>
<td>.27</td>
<td>.30</td>
<td>.99</td>
</tr>
<tr>
<td>doctor – student 9</td>
<td>week 26</td>
<td>.16</td>
<td>.33</td>
<td>.93</td>
</tr>
<tr>
<td>doctor – student 9</td>
<td>week 27</td>
<td>.13</td>
<td>.27</td>
<td>.95</td>
</tr>
<tr>
<td>doctor – student 11</td>
<td>week 30</td>
<td>.23</td>
<td>.83</td>
<td>.86</td>
</tr>
<tr>
<td>doctor – student 11</td>
<td>week 31</td>
<td>.15</td>
<td>.54</td>
<td>.85</td>
</tr>
<tr>
<td>doctor – student 12</td>
<td>week 31</td>
<td>.09</td>
<td>.15</td>
<td>.96</td>
</tr>
<tr>
<td>doctor – student 12</td>
<td>week 32</td>
<td>.14</td>
<td>.27</td>
<td>.95</td>
</tr>
</tbody>
</table>

Table 3 also shows that there is no improvement in the agreement between one doctor - student pair when it was measured for two or three consecutive weeks.

Sensitivity scores vary between the pairs: for the complete set of diseases the range is .10 to .83, mean: .36. For the core diseases the range is .14 to 1.00, mean: .51. With 15 pairs the sensitivity scores were higher for the core diseases than for the complete set of diseases. The specificity also differs between the pairs: for the complete set of diseases the range is: .85 to 1.00, mean: .96. For the core diseases the range is .66 to 1.00, mean: .93.
Table 4 shows whether doctor and student differ in the frequencies with which they had observed the diseases.

**Table 4** Discordancies between doctor and student in the frequencies in which diseases were observed, out of the total number of correct positives from 20 week sets

<table>
<thead>
<tr>
<th></th>
<th>total number of correct positives</th>
<th>number (percentage) observed in different frequencies</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>doctor=student&lt;sup&gt;1&lt;/sup&gt;</td>
<td>doctor&gt;student&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>‘complete set of diseases’</td>
<td>125</td>
<td>65 (52%)</td>
<td>39 (31%)</td>
</tr>
<tr>
<td>‘core diseases’</td>
<td>74</td>
<td>34 (46%)</td>
<td>27 (37%)</td>
</tr>
</tbody>
</table>

<sup>1</sup> doctor and student observed the disease with the same frequency  
<sup>2</sup> doctor observed the disease with a higher frequency than the student  
<sup>3</sup> doctor observed the disease with a lower frequency than the student

Out of the complete set of diseases and the core diseases that doctor and student had observed (sum of correct positives of 20 doctor – student pairs), it is shown that in most cases (52% and 46%) doctor and student observed the disease in the same frequency; in 31% and 37% they are observed more often by the doctor than by the student, and in 17% and 18% less often by the doctor than by the student.

A coefficient of agreement was also calculated for the four pairs of students who were at the same subdivisions during the same week. Table 5 shows the J values. For the complete set of diseases the range is .01 to .25, mean: .13. For the core diseases the range is .03 to .42, mean: .23.

**Table 5** The coefficient of agreement (J) for 4 pairs of students

<table>
<thead>
<tr>
<th></th>
<th>complete set of diseases</th>
<th>core diseases</th>
</tr>
</thead>
<tbody>
<tr>
<td>student 5 - student 6</td>
<td>week 19</td>
<td>.25</td>
</tr>
<tr>
<td>student 8 - student 9</td>
<td>week 26</td>
<td>.01</td>
</tr>
<tr>
<td>student 10- student 11</td>
<td>week 29</td>
<td>.18</td>
</tr>
<tr>
<td>student 11- student 12</td>
<td>week 31</td>
<td>.08</td>
</tr>
</tbody>
</table>
Discussion

The results of this study show that the logbook data are not fully consistent between doctor (standard) and student. Differences between doctor and student show up for the types as well as for the number of diseases that are observed. The mean coefficient of agreement (J) between doctor and student is .23 for the complete set of diseases, and .36 for the core diseases. First of all, in interpreting this coefficient as a high or low estimate of the reliability of logbook data, several factors should be discussed.

The coefficient chosen to measure interobserver agreement should always be taken into account when interpreting reliability coefficients. As has already been stated, the Jaccard coefficient is a stringent measure of agreement at the item level as it takes into account only items (diseases) that are observed. Items (diseases) that are not observed, are not counted as agreements. In an instrument like our logbook, where students and doctors only record a relatively small number of diseases, this is a relevant consideration as many ‘double zero’ items might (artificially) enhance the reliability assessment. However, the procedure that we followed in choosing the Jaccard coefficient implies that this statistic yields a lower limit of the reliability.

Another important factor to take into account when interpreting the reliability of data is the sensitivity (indicating the false negative errors or missed events) and the specificity (indicating the false positive errors or false identifications). These parameters are highly useful descriptors of the quality of data. Kaye stated that sensitivity and specificity are more useful than coefficients reflecting strength of interobserver agreement for evaluation of the quality of observations. In this study the mean specificity for the complete set of diseases is .96 and for the core diseases .93; the mean sensitivity is respectively .36 and .51. This means that there were more missed events than false identifications among students. The missed events indicate that the student recognized fewer diseases than the doctor. In other words, the student does not observe all the diseases present at the department. The false identifications indicate that the student recognized diseases that the doctor had not observed. Thus, the students record diseases which are not present on the department. As is pointed out by Carey and Gottesman, false identifications are more serious
than missed events, and therefore specificity is the most relevant quality index of observational data.\textsuperscript{18} In our study the specificity is high, which may indicate that the quality of the logbook data is high in the sense that students do not record many false identifications. On the other hand, as can be seen in Table 4, the doctors register more diseases than students in 31% and 37% of the cases. This may imply that doctors registered more (concomitant) diseases than students, due to which students indeed do not report many false identifications, so sensitivity and specificity should be interpreted with caution. Moreover, students make errors, they do not record all the diseases present in the department and therefore the data in the logbook are partially incomplete. In this sense, the quality of the data of the logbook is poor.

The instrument itself, the logbook, which is used to gather data is also a factor which should be taken into account when interpreting reliability.\textsuperscript{3,19} Vanek et al.\textsuperscript{3} and Dent and Davis\textsuperscript{20} described several factors which can influence the reliability of logbook data, such as the practicability of the logbook, the instructions for use, the vocabulary, complexity and length of the logbook, and the participation and compliance of the students. As has been described previously during the development phase of our logbook, all these features were taken into account.\textsuperscript{6} Our logbook is a practical pocket-book, specifically devoted to the Internal Medicine internship. Besides to written instructions what and how to register, extensive verbal instructions were given. The medical vocabulary used is based on the vocabulary of the Dutch National Blueprint, and the complexity and length is in accordance with the level that could be expected during the internship. Moreover, during our study, the participation of the students was satisfactory. Therefore, the level of agreement measured between the data of the doctor and the student does not seem to be influenced by the instrument.

We hypothesize that a factor that can have influenced the agreement, is a lack of experience of the students in recognizing and in diagnosing diseases and disorders. It is described by Popping that the experience of the observers and the complexity of the object of observation (the diseases) are factors which can influence agreement.\textsuperscript{21} In diagnostic reliability studies it was reported that 80% of all diagnostic disagreements are caused by inadequate nosology.\textsuperscript{19} Therefore, it is obvious that students could not recognize all the diseases which the experienced doctors recognize, because their knowledge and experience with
diseases and disorders is still insufficient. Moreover, our students have just undergone the transition from the pre-clinical to the clinical years, which seems to be a difficult one.\textsuperscript{22} Internal medicine is the first internship after four pre-clinical years, and for most of the students it is their first confrontation with the hospital organization and with patients encounters. This may indicate that students are in a process of socialization and are therefore not able to concentrate fully on diagnosing diseases and making records of their experiences in the logbook. This may also explain why we do not find any increase in the agreement between doctor and student when it was measured for 2 or 3 weeks consecutively. It is expected that progress in learning would be reflected in a higher agreement in the second or third week of the internship.

Students seem to be better at recognizing the core diseases than the complete set of diseases. There was a higher agreement between doctor and student with the core diseases, and the sensitivity was also higher. Additionally, within the pairs of students agreement was higher for the core diseases than for the complete set of diseases. An explanation for this finding may lie in the importance of the core diseases in the teaching programme during the pre-clinical years. As a consequence, students are better prepared to recognize these core diseases. As our logbook is based on the national Blueprint, which requires that our students have experiences with core as well as `other’ diseases, it is important to focus on both during the internship. After all, if clinical problem solving is the key to learning, students should examine many patients with different kinds of diseases.\textsuperscript{23} This is also the reason why we focus on the diseases during the internship rather than on patient problems. Moreover, patient problems do not belong to a specific discipline and should therefore be worked on during the whole educational programme, from year 1 to year 6.

Another possible explanation for the disagreements is related to the use of learning potential at the clinic by students themselves. It has been shown that medical students do not make full use of the learning potential provided by the available patients\textsuperscript{24}, under-report their patient encounters\textsuperscript{9,10}, and differ in the reports they make.\textsuperscript{5,8,25} The low agreement between students found in this study also indicates that the available diseases are used differently by students. Students, at the same department during the same period do not register the same diseases. Thus, a better use of clinical facilities can be obtained by guiding and fostering students in achieving the goals of the internship.\textsuperscript{26,27}
However, lack of recognition of diagnoses during the internship should be corrected by feedback measures, because clinical experiences without feedback or evaluation lead to poor medical practice.\textsuperscript{28,29} Hence, an improvement in feedback with respect to the correct recognition and recording of the available diseases and disorders will probably lead to an improvement in the agreement coefficient.

This study has limitations. Although doctor and student are exposed to the same complete group of patients within a subdivision during one week, we can not prove with certainty that the recorded diseases belong to the same individual patients. Thus when the agreement between doctor and student for the observed diseases is calculated, it is likely but not certain that the data concern the disease(s) attributed to the same patient. This means, therefore, that the measured agreement is approximate.

In this study we assess reliability by using a standard (the doctor) and measuring the agreement between the student and the correct diagnosis according to the standard. Considering the results we have no reason to assume that the doctors did not record the correct diagnosis; nevertheless there is a lack of proof that the data of the doctors are objective and that we can talk about a standard. However, even if there was evidence that diagnosticians agree perfectly, this does not mean that they are right all the time.\textsuperscript{19}

We conclude that students’ logbooks represent objective data, non-objective data and missing data. The non-objective and missing data probably arise from the imperfect recording of students caused by: (1) insufficient knowledge and experience during this first internship, whereby students did not recognize the diseases for what they were; and (2) insufficient feedback of supervisors to correct students with regard to their omissions. Unreliable logbook data yield a non-representative account of actual learning experiences and therefore lead to wrong decisions about the learning progress of students and about the teaching programme. In order to use the logbook as an evaluation tool to determine how students are progressing towards the objectives of the internship, students should be supervised and receive feedback. This means that students should receive cues about relevant learning from the available diseases and about what should be the focus of their activities. This also implies that with regard to the diseases students had encountered, they should receive help
and direction in order to fill in their logbooks in accordance with their actual experiences.

References


5 Dolmans D, Schmidt A, Beek J van der, Beintema M, Gerver WJ. Does a student log provide a means to better structure clinical education? Med Educ 1999;33:89-94.


12 Mitchell SK. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychol Bull 1979;86:376-90.
