Introduction

Recording of treatment time in physiotherapy is required for two reasons. Firstly, accurate recording of treatment time is required in clinical research. The time spent practising specific tasks is an important variable influencing motor learning (Schmidt and Lee 1999) and has been used as a measure of treatment, either alone or in combination with treatment description in several studies evaluating treatment effectiveness. Richards et al (1993) used therapists’ records of treatment time to define intervention and Sunderland (1992) used a combination of physiotherapist and observer records to gain records in single minutes. Malouin et al (1992) recorded treatment in 5min time units, adding a description of gait training, and Goldie et al (1996) required clinicians to record time spent in specific activities.

Several recent studies have shown a link between treatment time or intensity and outcome. A research synthesis by Kwakkel et al (1997) reported a small but significant relationship between treatment intensity and effect. Further support has come from studies by Kramer et al (1997), Feys et al (1998) and Kwakkel et al (1999). An important part of any further investigation into this link with outcome is the accurate measurement of treatment time.

Physiotherapists are routinely required to provide daily records of treatment time for administrative purposes. The allocation of resources in the health system is increasingly dependent on such statistics, particularly in times of economic constraint. The National Allied Health Casemix Committee is currently developing Version 2 of the Australian Allied Health Classification System, in which the time taken to complete a treatment procedure is one of the measures which will be used in the development of performance indicators for the cost effectiveness of allied health intervention (National Allied Health Casemix Committee 1997). So, whilst less detailed information may be required than for research, accuracy is still important.

A review of the literature revealed that, to date, no study has investigated the accuracy of data relating to physiotherapy treatment time in stroke rehabilitation. Since the recording of treatment time in stroke rehabilitation is important in both clinical research and daily practice, it is important to evaluate the accuracy of such data. Therefore, this study addressed this issue.

In this study, physiotherapists were asked to use the specific method for recording treatment time employed in a previous study (Goldie et al 1996). This method was designed to provide a more descriptive
A summary of how treatment time is spent in stroke rehabilitation is usually required for administrative purposes as part of clinical practice. Keith (1997) states that research into outcomes of rehabilitation is severely hampered by lack of specification of treatment. The method was developed in consultation with senior neurological clinicians in a number of rehabilitation centres in Melbourne. Clinicians were asked to provide a list of categories into which treatment of patients with stroke could be divided. These categories included upper limb, bed mobility, sitting, sit to stand, standing, early gait and advanced gait activities. The final version of the record form was agreed on after a further extensive process of collaboration involving clinical pilot studies.

Despite the increased complexity involved in both the use and evaluation of this recording method, its utilisation was warranted by the potential value of providing behavioural operational definitions for treatment rather than the more general methods usually employed. The method was based on a model of motor learning, and defined treatment in behavioural terms. Physiotherapists were asked to record time spent in supervised practice of motor tasks according to specific categories. The recording method was designed to accommodate any of the differing philosophies of treatment currently used in stroke rehabilitation. This method of recording was selected since the long term aim was to provide descriptive data about the activities undertaken in physiotherapy during stroke rehabilitation. The specific aim of this study was to investigate how accurately the physiotherapists recorded treatment time according to this method.

One key issue to be resolved was how to obtain the criterion measure of treatment time against which each physiotherapist’s recording could be judged. Following consideration of options such as the use of live observers, a decision was made to videotape the treatment session. This provided a permanent record allowing accurate measurement of treatment time with an electronic timer. Following the treatment session, the criterion measure of treatment time was obtained by two criterion raters and used as the standard against which to correlate the clinical physiotherapists’ self-reported estimate of treatment time.

In summary, this study used a specific method designed for stroke rehabilitation to investigate:

- the extent of error in clinicians’ recordings of time,
- whether clinicians systematically over- or underestimate actual time scores, and
- the relationship between clinician and criterion time scores.

**Method**

**Subjects** Twenty-six physiotherapists participated from four rehabilitation hospitals in Melbourne. The physiotherapists were working in stroke rehabilitation and satisfied the selection criteria of having participated in a training session about the use of the treatment record form, having at least 15 sessions’ experience with this form in daily practice (as part of a more extensive study of stroke rehabilitation), and agreeing to be involved in the study.

Two physiotherapists agreed to act as criterion raters. Both had been involved in the development of the treatment record form and each had at least five years’ experience in stroke rehabilitation.

**Apparatus** A Panasonic Portapak recorder and video camera were used to record the treatment sessions. Video recordings were viewed using a Panasonic videocassette recorder and criterion ratings of treatment time were measured using an electronic timer.

The record form required the clinicians to make daily recordings of the time in minutes for each of the defined treatment activities: upper limb, bed mobility, sitting, sit to stand, standing, early gait and advanced gait. Definitions were provided for each category on the recording form. Instructions on the reverse of the form emphasised that only practice time spent under the direct supervision of the clinician was to be recorded. No attempt was made to record independent practice.

**Procedure** Video recordings were made of one treatment session for each of the 26 clinicians. A number of measures were employed to attempt to minimise intrusion into daily clinical routines. A convenient time for videotaping the treatment session was negotiated with each participating clinician. During the taping procedure, the camera was
positioned as unobtrusively as possible. Physiotherapists were instructed to proceed with their treatment as usual, and to complete the treatment record form at the time they would normally assign for this task. At the time of this study, clinicians were recording treatment time on a daily basis during the first eight weeks of rehabilitation for selected patients.

The 26 videotaped sessions were later viewed separately by the two criterion raters, at their own pace and with as many repetitions as necessary to gain an accurate record of time spent in each category on the record form.

**Statistical analysis** Time scores were available for 26 treatment sessions for each of the two criterion raters and each individual clinician undertaking the treatment sessions. The K-S Lillefors test (Norusis/SPSS Inc. 1990) revealed that the distributions of each of the criterion raters’ time scores and the clinicians’ scores departed significantly from normal ($p < 0.05$) in all categories. Therefore the data were analysed using non-parametric methods.

**Criterion time scores** To give an estimate of “error” in the criterion times, the time scores of one criterion rater were subtracted from those of the other. The time scores for the two criterion raters were averaged for each category to derive the criterion time against which clinicians’ records were compared. The averaging process provided the best estimate of the “true” score. To assist in making judgments about the importance of the differences between criterion raters, the absolute values of these differences were then expressed as a proportion of derived criterion time. To investigate the reliability of the two criterion raters, their scores were correlated using the Spearman rank order correlation coefficient.

**Clinicians’ error scores** An error score was calculated for each clinician for each category by subtracting the criterion score from that of the clinician. Positive scores indicate that the clinician was overestimating time spent, and negative scores show an underestimation. Systematic differences between the clinicians’ scores and the criterion scores were investigated with the Wilcoxon signed rank test. To investigate the relationship between the two sets of scores, the Spearman rank order correlation coefficient was calculated. Inferential statistics were conducted with a Type I error of 0.05.

**Results**

Figure 1 shows boxplots of the distribution of differences between the two criterion raters’ time scores for each of the seven categories from the 26 videotapes. Some large individual differences are...
evident, especially in the advanced gait category, however the differences tend to cancel each other out resulting in medians close to zero (see Figure 1 and Table 1).

Figure 2 shows boxplots of the distributions of error scores (ie the difference between criterion and clinician scores) for the remaining six categories. There did not appear to be a systematic pattern to the direction of the error and the Wilcoxon signed rank test confirmed that these differences were not statistically significant ($p > 0.05$). Some clinicians overestimated time spent in various categories and some underestimated. These errors tended to cancel each other out, again resulting in medians close to zero (Table 2). A comparison between Tables 1 and 2 reveals that the absolute values of these errors are generally larger for clinician error than for differences between the criterion raters.

Table 2. Magnitude of clinicians’ errors.

The number of time recordings made for each category out of 26 videotapes is listed in brackets, and may differ from Table 1. To omit a video tape from a category there had to be agreement from both criterion raters and the clinician that no treatment had occurred. In some cases the clinician recorded in a category that both criterion raters did not, making the number of treatment sessions for Table 2 slightly larger. The “derived” criterion is the average of the two criterion raters’ time scores.

<table>
<thead>
<tr>
<th>Treatment Category</th>
<th>Upper Limb (21)</th>
<th>Bed Mobility (15)</th>
<th>Sitting (16)</th>
<th>Sit to Stand (19)</th>
<th>Standing (25)</th>
<th>Early Gait (22)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Median error scores (min)</td>
<td>0.9</td>
<td>-0.5</td>
<td>-0.4</td>
<td>-0.1</td>
<td>1.0</td>
<td>1.3</td>
</tr>
<tr>
<td>Median absolute error scores (min)</td>
<td>2.9</td>
<td>1.6</td>
<td>1.1</td>
<td>0.4</td>
<td>2.3</td>
<td>3.9</td>
</tr>
<tr>
<td>Median derived criterion (min)</td>
<td>7.4</td>
<td>6.4</td>
<td>1.2</td>
<td>1.2</td>
<td>7.2</td>
<td>5.9</td>
</tr>
<tr>
<td>Median error as a proportion of the derived criterion</td>
<td>39.6%</td>
<td>55.8%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>47.2%</td>
<td>56.5%</td>
</tr>
</tbody>
</table>

The reliability with which the two criterion raters judged treatment times was high for all but the gait categories (upper limb rho = 0.88, bed mobility rho = 0.81, sitting rho = 0.82, sit to stand rho = 0.85, standing rho = 0.93). For early gait, the correlation was significant but of only moderate strength (rho = 0.5). For advanced gait, the correlation did not reach a level of significance. Therefore this criterion category was not used in further analysis.

Despite the lack of perfect agreement between the two criterion raters, the average of the two sets of scores was calculated to gain the derived criterion (Table 1). The averaging process tends to cancel out random error and reduce systematic error so as to provide the best estimate of the true scores.

In attempting to make judgments about the importance of the differences between the criterion raters, we expressed these differences as a percentage of the derived criterion time for each category (Table 1). Absolute values of the differences were used, as we were interested in their magnitude regardless of whether they were positive or negative. For two categories, the median absolute difference between the two criterion raters accounted for a large proportion of median derived criterion time. For the category of sitting, the relatively small median difference (0.7 minutes) amounts to 74.5% or three-quarters of the median derived criterion time spent practising this activity. For sit to stand, the proportion is even larger.
When the absolute values of these error scores were expressed as proportions of derived criterion times (Table 2), these proportions were all at least 40% of median treatment time in each category. Again, for the two categories with treatment time of briefest duration (sitting and sit to stand), error accounted for a larger proportion. In fact, for these categories, error is the same (ie 100%) as the average time spent practising this activity.

The correlation between criterion and clinician scores was high for the upper limb (0.79), sit to stand (0.83) and standing (0.8) categories and moderate for bed mobility (0.72) and sitting (0.69). A moderate relationship is suggested for the early gait category (0.49).

**Discussion**

The results of this study have shown that, relative to the derived criterion, the clinicians’ recording of treatment time was of moderate to high accuracy for all except the early gait category. In addition, the clinicians did not systematically overestimate or underestimate treatment time when compared with the derived criterion measure. The average error scores were of the order of only one minute or less for all except early gait. However, they did constitute an average of 40% or more of the criterion time for each treatment category. There were also some very large individual errors. It is most likely that these large differences were due to category disagreements rather than difficulty in judging the precise start and finish of an activity.

It is important to consider the factors that may influence the accuracy of the clinicians’ scores. One factor influencing the correlation between the clinicians’ scores and the criterion measure may be due to error in the criterion measure itself. An effort was made to obtain a criterion measure of the true score of treatment time by electronically timing videotapes of treatment sessions under ideal conditions. Nevertheless, the agreement between the two raters was not perfect, especially for the category of advanced gait, where poor agreement precluded its further use as a criterion measure. Since the criterion ratings were scored under “ideal” conditions, it would be unreasonable to expect clinicians to do better. Both
criterion raters noted some difficulty in assigning treatment to the specific categories required by the form. There were also some minor technical problems with the videotaping procedure, affecting sound quality on two tapes. Averaging of a larger number of criterion raters’ scores is likely to have resulted in a better estimate of the true score. In establishing criterion-related validity, further studies will need to consider this issue. The best criterion for this study was obtained within the limits of resources available. The differences between criterion scores were mostly less than one minute and the criterion raters could generally rank the time spent in each category consistently. However, even though the average differences were not large, there was some difficulty in agreeing on the exact amount of time spent, and these differences were substantial proportions in categories of brief duration.

Human error on the part of the clinician was another important factor. It is likely that error was made, due not only to the difficulty of judging precisely how much time was spent in each category, but also to problems in recalling this information accurately. Human error may be greater when treatment sessions are interrupted. The complexity of the treatment session may also influence the accuracy of recordings. It may be possible to reduce error due to human judgment by asking clinicians to record immediately following the treatment session. However, this constraint may be unrealistic in the typical clinical situation when time data are likely to be recorded at the end of the day, or even later.

The design of the record form itself may also have contributed to error in recording the data accurately. Clinicians may have accurately recorded time spent but used the wrong category. Although the categories were defined, experience with the form has revealed that further clarification may be needed for some. For example, clearer definitions may be required to distinguish early and advanced gait activities, as the recordings for these categories showed large variations for the criterion raters, who scored under ideal conditions.

Another potential problem resulting from the design of the form was that the categories were not always mutually exclusive. For example, confusion may...
The design of the form also required that only activity under the direct supervision of the physiotherapist be recorded. Clinicians may have included the time spent in verbal communication about the treatment whilst the patient was not actively practising. Recording on a simple form that only requires the total amount of time spent with the patient may give more accurate data, but does not provide the detail given in this form. The accuracy of the data in this study was reasonably high considering the complexity of the form.

The findings of this study have indicated that recording time spent practising as a measure of treatment may not be the best method in tasks which require only short amounts of time to complete, such as standing up from sitting. Given that the proportion of treatment time accounted for by error was found to be larger in such a task, counting the number of repetitions may be a better method of recording the amount of practice that has occurred. Nugent et al (1994) used physiotherapists’ records of the number of repetitions of a weight-bearing exercise aimed at improving the walking of patients following stroke as their measure of treatment. The methods employed for quantifying treatment will obviously need to reflect the reasons for collecting the data. Whilst the number of repetitions of a relatively rapid task may provide valuable information for a researcher wishing to quantify specific treatments, the actual amount of time taken to complete treatment may also be required. Keith (1997) proposes that a number of different ways of quantifying treatment delivery, such as measuring timing, intensity, duration, dosage and specificity, are necessary in attempting to understand why different treatments may produce different outcomes.

The method utilised may have biased favourably the results of this study. Whilst every effort was made to minimise the intrusion of the videotaping procedure, there is no doubt that clinicians were aware of the aim of the study. An alternative method would have been to videotape many treatment sessions for each clinician, so as to allow them to become accustomed to the procedure, randomly selecting only one of these sessions for analysis. It is unlikely that clinics would agree to such resources allocation. The use of discrete live observers may have placed less awareness on the part of the clinicians for accuracy. However, the criterion score obtained from the live observer is likely to have more error than the criterion score obtained from repeated viewing of a permanent record of the treatment session. The methodology chosen in this study, therefore, reflects the desire to evaluate the time spent in treatment as accurately as possible, again using realistically available resources.

It appears that there is a paucity of knowledge relating to this issue in the physiotherapy profession, making it difficult to draw comparisons with other studies. One study in the teaching profession (Sargent 1981) reported systematic overestimation of the amount of time spent giving direct instruction. In contrast, the results of our study showed that there was not a systematic overestimate or underestimate of time spent in physiotherapy treatment.

The moderate to high accuracy found in this study was obtained from 26 physiotherapists working in stroke rehabilitation. This is an adequate sample, allowing generalisation to a similar population. These clinicians had participated in a training session about the use of the form and had gained experience for at least 15 sessions in using the form on a daily basis, as part of a larger study of stroke rehabilitation. The clinicians were required to record treatment of selected patients only, not their total caseload. This factor may have influenced the results favourably. Accuracy may have been compromised if data were required for the total workload.

When the recording of treatment time is to be used for clinical research or administrative purposes, it is necessary to validate the accuracy of such data. This study has provided a model of how such a validation study can be conducted. Despite the potential for many factors to influence the accuracy of the data, the results showed that both criterion raters and clinicians are mostly able to agree on relative proportions of time devoted to different activities within a treatment session, using this relatively complex recording form. They are, however, less able to agree on exact times, especially with brief activities. Future use of this recording method will need to address these issues. Further studies are required to evaluate the accuracy of data using other recording methods. The challenge remains for the profession to provide further evidence about the meaningfulness of data that give a first stage description of what constitutes physiotherapy treatment.
Authors
JE Wittwer, School of Physiotherapy, La Trobe University, Bundoora, Victoria 3083. E-mail: j.wittwer@latrobe.edu.au (for correspondence). P Goldie, School of Physiotherapy, School of Physiotherapy, La Trobe University, Bundoora, Victoria 3083. TA Matyas, School of Psychological Science, School of Physiotherapy, La Trobe University, Bundoora, Victoria 3083. MP Galea, School of Physiotherapy, The University of Melbourne, 200 Berkeley Street, Carlton, Victoria 3053.

Acknowledgements
This study was supported financially by a Commonwealth Health Services Research and Development Grant. The commitment made by the physiotherapists from the following four hospitals is gratefully acknowledged: Royal Talbot Rehabilitation Centre, Caulfield General Medical Centre, Hampton Rehabilitation Centre and North West Hospital, Mount Royal Campus.

References


