International Journal for Quality in Health Care 1999; **Volume 11, Number 5**: pp. 419–424

Reliability study of the European appropriateness evaluation protocol

S. LORENZO[1], T. LANG[2], R. PASTOR[3], A. TAMPIERI[4], B. SANTOS-EGGIMANN[5], H. SMITH[6], A. LIBERATI[4] AND J. RESTUCCIA[7]

[1]Fundación Hospital Alcorcón, Madrid, Spain, [2]Medical Informatics Unit, Paris Hospital, France, [3]Department of Epidemiology and Biostatistics, National School of Public Health, Instituto de Salud Carlos III, Madrid, Spain, [4]Laboratory of Clinical Epidemiology, Istituto di Ricerche Farmacologiche Mario Negri, Milan, Italy, [5]Department of Social and Preventive Medicine, University of Lausanne, Switzerland, [6]Primary Medical Care, University of Southampton, UK, [7]Health Care Management Program and Operations Management Department, Boston University School of Management, USA

Abstract

Objective. To help to co-ordinate and harmonize research on utilization review in Europe, the US Appropriateness Evaluation Protocol (AEP) was adapted for use in the European setting. The aim of this paper is to assess the reliability of the European version of the AEP (EU-AEP).

Design. Nineteen English-language medical records were reviewed by a physician reviewer from each of six participating countries: Austria, France, Italy, Spain, Switzerland and the UK. Each of the six reviewers was asked to assess the appropriateness of the 19 admissions and 31 hospitalization days (19 admission days and 12 randomly selected days of hospital stay, excluding days of discharge) using the revised review instrument. To evaluate inter-rater reliability, the κ statistic was used to measure overall and pair-wise agreement for the assessment of appropriateness of admission and of day of care, respectively.

Results. For admission, the overall κ statistic among the six reviewers was 0.64, with κ values for each pair of reviewers in the range 0.46–0.86. For day of care, the κ was 0.59, with pair-wise κ coefficients in the range 0.25–0.95.

Conclusion. The observed agreement could be considered substantial, especially if the fact that medical records were hand-written in a language native to only one of the reviewers is considered. Besides all the study limitations, this finding provides at least preliminary support for the application of the EU-AEP as a reliable instrument in the European setting, including application in comparative studies involving two or more countries.

Keywords: Appropriateness Evaluation Protocol, appropriateness of hospital use; European Appropriateness Evaluation Protocol; reliability study; utilization review

Utilization review (UR) is defined as review of the patient's medical record through application of explicit criteria and/or expert opinion to assess the appropriateness of decision making related to the site, frequency and duration of patient care [1–3]. The resultant information is useful in selecting corrective actions by which to reduce inappropriate use of hospital resources, thereby containing costs while not jeopardising access to appropriate hospital use. Such actions include: (i) identifying the reasons for inappropriateness in order to guide changes in policies, procedures and operating systems inhibiting appropriate hospital use [4]; providing feedback about comparative rates of appropriateness to hospitals and physicians [5]; and changing payment methods to give providers and patients incentives to use health care resources appropriately [6].

In 1993, our research group, investigators in a European Union BIOMED study of appropriateness of hospital use, found that although some UR instruments had been shown to be reliable in studies within a single country [7–11], none had been shown to be so between countries [12–14]. This limitation precluded the use of UR to obtain comparative rates of appropriate use across countries as benchmarks and, through investigation of the reasons for the differences in rates, identify opportunities for corrective action intended to yield improvement in this important performance measure. To address this limitation, we decided to identify an existing

Address correspondence to Dr S. Lorenzo, Gabinete del Plan de Calidad, Fundación Hospital Alcorcón, Av. Budapest 1, 28922 Alcorcón (Madrid), Spain. E-mail: slorenzo@fhalcorcon.es

UR instrument, modify it for application in the European setting, and test it for reliability across countries. As the Appropriateness Evaluation Protocol (AEP) was the most commonly used UR instrument throughout Europe [9], we selected the Adult Medical–Surgical version of this US review instrument (US-AEP) for this purpose [15].

Methods

Instrument development

Groups of experts, consisting of physicians and health services researchers experienced in UR, from seven European countries participated in the instrument development process: Austria, France, Italy, Portugal, Spain, Switzerland and the UK. Linguistic, conceptual and technical issues arose during the process of adaptation of the US instrument to the European setting. Consensus was needed between the research group members on the criteria that needed modification, given the existing differences among the participating countries. The main differences were not only cultural, but in the organization and financing of the health systems of the different countries. After several iterative rounds of modification of the US-AEP, our research group reached consensus on the European version of the instrument (EU-AEP) [12].

The European version of the AEP assesses the appropriateness of the timing and the level of care of adult medical and surgical patients [12]. It has two parts, one based on clinical criteria aimed at identifying inappropriate admissions or hospital days, and one used to classify the determinants of inappropriateness. Differences between the EU-AEP and the US-AEP are relatively minor. The general structure of the instrument and its independence in relation to the diagnosis are kept in the new review tool. The admission, the subsequent days of care and the patient's readiness for discharge are reviewed, and both the services provided and the patient's condition (i.e. severity of illness and stability) are considered in making the decisions. The relevant information reviewed comes from the medical record, and all documents included in the medical record are considered to be sources of information [16]. The decision rule is relatively simple. If one of the 15 admission criteria is met, the admission is deemed appropriate; if one of the 25 day-of-care criteria is met, that hospital day is considered appropriate. The reviewer or a consultant may override the decision in cases in which the criteria do not sufficiently capture the patient's situation. If the admission or a hospital day is deemed inappropriate, the instrument provides a list of reasons identifying the cause of this inappropriateness, and a list of alternative levels of care required by the patient. The list of reasons for inappropriateness has been the main modification made to the original US-AEP [12].

Study design

Each participating country, except Portugal, provided an expert reviewer to conduct a reliability study of the new instrument. All reviewers were physicians with experience in making UR assessments in their own countries, except the Austrian reviewer who had relatively little experience due to the lack of UR studies conducted in Austria at that time. The reviewers served as investigators in the international BIOMED project and they had participated in the research team that had developed the European review instrument. Thus, all reviewers were familiar with the use and interpretation of the EU-AEP.

Each reviewer assessed the same set of 19 English medical records, nine abstracted and typed from the US, and 10 hand written from the UK. The records consisted of a subset of records from a variety of clinical services that treat adult medical and surgical patients in several US and UK hospitals, both teaching and non-teaching, that had been used to test the reliability of new reviewers in Boston and Southampton. These records were chosen according to clear-cut decisions on appropriateness of hospitalizations. The selected records were copies of hospital medical records with the exception of the patient's identifying information, which was purged to maintain confidentiality. These records were mailed to the reviewers in each participating country, who conducted their reviews independently. Reviewers were asked to keep confidential the information in the medical records to be reviewed and to destroy them once the review process had finished. The reviewers used the AEP user's manual [17], besides that there was no other standardization of the review process. The appropriateness of admission was assessed from the 19 medical records. To evaluate the appropriateness of day of care, 19 admission days (i.e. the first hospitalization day of the 19 medical records) and 12 randomly selected days of hospital stay (excluding days of discharge) were reviewed using the new instrument. Thus, each reviewer was asked to review the 19 admissions to assess their appropriateness, and 31 hospitalization days to assess the appropriateness of day of care. Using this sample size, the statistical power of our reliability study to find an underlying agreement greater than 0.5 ('moderate' to 'substantial' agreement) was 0.70 for the appropriateness of admission and 0.86 for the appropriateness of day of care.

Statistical methods

To evaluate the inter-rater reliability of the EU-AEP, the overall and pair-wise κ coefficients were calculated separately for the assessment of inappropriateness of admission and of day of care. κ coefficients were computed using Schouten's modification to allow for missing data [18]. The SEs of κ coefficients were calculated using jack-knife estimates [19]. The statistical analysis was carried out using the SAS statistical package (version 6.10) [20].

Results

Appropriateness of admission

Table 1 presents the assessment of appropriateness of admission for the 19 selected medical records, by country of

Table 1 Appropriateness of admission for the 19 selected medical records, by country of reviewer

	Austria	France	Italy	Spain	Switzerland	UK
Appropriate[a]	14	14	14	12	15	14
Intensity of service	13 (93%)	12 (86%)	13 (93%)	10 (83%)	9 (60%)	13 (93%)
Severity of illness	3 (21%)	4 (29%)	4 (29%)	3 (25%)	3 (20%)	3 (21%)
Unspecified[b]	0 (0%)	0 (0%)	0 (0%)	0 (0%)	3 (20%)	0 (0%)
Inappropriate	5	5	5	7	4	5

[a] Because of appropriate admission criteria are not mutually exclusive, the percentages sum >100%.
[b] Appropriate admissions without a specific criterion.

Table 2 Pair-wise κ coefficients of concordance for appropriateness of admission, by country of reviewer

	Austria	France	Italy	Spain	Switzerland	UK
Austria						
France	0.46					
Italy	0.46	0.46				
Spain	0.52	0.76	0.83			
Switzerland	0.56	0.56	0.80	0.63		
UK	0.46	0.46	0.86	0.76	0.56	

reviewer. The proportion of inappropriate admission ranged from 21% (four of 19 cases) as assessed by the Swiss reviewer, to 37% (seven of 19 cases) for the Spanish reviewer. Among all reviewers, intensity of service was a more common justification for appropriate admissions (from 60% for the Swiss reviewer to 93% for those from Austria, Italy and UK) than patient condition (from 20% for the Swiss reviewer to 29% for the French and Italian reviewers). The most frequent reasons used to classify inappropriate admissions were lack of expert opinion or investigation (35%), admission required by specialist (29%) and conservative practice (19%).

There was total agreement among all reviewers in 13 admissions (68%) and partial agreement in the remaining six admissions (32%). The overall κ coefficient for classifying admissions as appropriate or inappropriate among the six reviewers was 0.64 (SE = 0.12). This means that the observed disagreement was only 36% of the disagreement that would be expected if classifications had been made at random. Using the Landis and Koch's guidelines for interpreting κ values [21], this level of agreement is considered 'substantial'. The κ coefficients for each pair of reviewers ranged from 0.46, indicating a 'moderate' agreement, to 0.86, indicating an 'almost perfect' agreement (Table 2).

Appropriateness of day of care

The assessment of the appropriateness of day of care by country of reviewer is given in Table 3. The proportion of days assessed inappropriate was similar among the reviewers from France, Italy, Spain and the UK (35–45%), but it was significantly different for the Swiss (23%) and Austrian reviewers (69%, *P* for heterogeneity among reviewers <0.01). The most frequent criterion used to justify the appropriateness of day of care was the necessity of nursing services, followed by the necessity of medical services. On the other hand, by far the most common reason for an inappropriate day of care was the non-necessity of health services (from 71% for the Swiss reviewer to 100% for the Spanish and UK reviewers). The EU-AEP asks the reviewer to determine the necessity of a service, in contrast to the US-AEP, which usually asks the reviewer only to determine if the service was ordered by a physician and provided to the patient (with the exception of a few services such as I&O).

There was total agreement among reviewers from all countries in 16 days of care (52%) and some degree of disagreement in the remaining 15 (48%). The overall κ coefficient for classifying days of care as appropriate or inappropriate was 0.59 (SE = 0.08), a level approaching the 60% threshold for 'substantial' agreement. The pair-wise κ coefficients for agreement among reviewers from the six countries showed a wide range of variation. The agreement ranged from a high value of 0.95 ('almost perfect' agreement) between the Italian and Spanish reviewers to a low of 0.25 ('fair' agreement) between the Austrian and Swiss reviewers (Table 4, below the diagonal).

When the appropriateness of day of care was classified in three mutually exclusive categories (appropriate, inappropriate because the patient does not need a health service, or inappropriate because the patient needs another health service), the overall κ coefficient was 0.55 (SE = 0.06), showing a slightly lower agreement than in the dichotomous classification. Similarly, the κ coefficients among each pair of reviewers were slightly lower when appropriateness of day of care was classified in three categories (Table 4, above the diagonal).

Table 3 Assessment of day of care for the 31 randomly selected hospitalization days, by country of reviewer

	Austria[a]	France	Italy	Spain	Switzerland[a]	UK
Appropriate[b]	9	17	19	20	23	18
Medical services	6 (67%)	10 (67%)	11 (58%)	10 (50%)	12 (52%)	4 (22%)
Nursing services	7 (78%)	15 (88%)	17 (89%)	16 (80%)	15 (65%)	17 (94%)
Patient condition	0 (0%)	0 (0%)	1 (5%)	1 (5%)	2 (9%)	1 (6%)
Unspecified[c]	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (4%)	1 (6%)
Inappropriate	20	14	12	11	7	13
Other service needed	5 (25%)	1 (7%)	1 (8%)	0 (0%)	2 (29%)	0 (0%)
No service needed	15 (75%)	13 (93%)	11 (92%)	11 (100%)	5 (71%)	13 (100%)

[a] In Austria and Switzerland there are 2 and 1 unclassified days of care, respectively.
[b] Because appropriate day of care criteria are not mutually exclusive, the percentages sum >100%.
[c] Appropriate days of care without a specific criterion.

Discussion

Overall, the between-reviewer reliability of the EU-AEP was substantial. The level of overall agreement achieved in this study was 0.64 for admission (with agreements for each pair of reviewers in the range 0.46–0.86) and 0.59 for day of care (with pair-wise agreements in the range 0.25–0.95).

The research team assumed that there were no cultural or country specific practices which might have been indicative of an admission or day of care being appropriate in one country and not in another. This assumption was a consequence of the lack of disagreement on that point during the instrument development. Judgement about the need for admission is based on the information available in medical records until the end of the day of admission, whereas the day of stay is assessed according to the information available up to the day of review. The process of reviewing photocopied manuscript medical records was slower than expected due to language difficulties, particularly interpretation of abbreviations. In our study, the agreement was lower for days of care than for admissions, possibly because it might be more difficult to link data in the case notes to the selected day. The fact that there was no standardization of the review process could have influence as well.

Although the overall agreements attained for both admission and day of care can be considered substantial according to standard criteria, these measurements of concordance could be underestimated because the levels of pair-wise agreement that include the Austrian reviewer are consistently lower than those observed among the remaining reviewers. Thus, excluding the Austrian reviewer, the overall agreement among the other five reviewers was 0.70 for admissions and 0.68 for days of care. These differences are probably due to this reviewer's lack of UR experience, given that UR had been introduced recently in Austria at the time of this study. As has been shown within an individual country [22], the availability of expert utilization reviewers (which typically requires training and reliability testing) will probably be essential to ensuring reliable results.

Although the results presented in this paper provide preliminary support for the reliable application of the EU-AEP, the study design presents several limitations. Firstly, given that there is only one reviewer per country, the observed variability between each pair of reviewers can not be interpreted as the systematic difference between the corresponding countries. For comparisons among countries, further research including more reviewers representing each participating country needs to be conducted. Secondly, although the number of medical records reviewed was determined in advance to detect an overall underlying agreement greater than 0.5 ('moderate' to 'substantial' agreement), a larger number of records would have been needed to increase the precision of pair-wise agreements. We limited the sample size of our study because of difficulties in obtaining, photocopying and distributing medical records throughout several countries. Finally, in this reliability study, the reviewers participated in the EU-AEP development, and the records were selected according to clear-cut decisions on appropriateness of hospitalizations. Therefore, further research is needed to confirm our results when reviewers not involved in the development of the instrument assess a random sample of standard European medical records.

The validity and reliability of the AEP has been evaluated in the USA [22,23], Israel [24], Italy [11] and Spain [9,10]. The results of overall agreement of our study are underscored by the fact that there were five reviewers whose native language was not English. These results are comparable to those reported by Gertman [23], Rishpon [24] and Peiró [9] for reliability studies conducted by reviewers in the same language, usually using abstracted medical records, which are much easier to read and interpret than original medical records. The fact that our reviewers were not familiar with medical records from other countries, which differ in their presentation and the many abbreviations that can be difficult to understand, adds to the strength of our findings. We would expect that the reliability of findings in cross-country studies, where expert reviewers apply the EU-AEP to medical records from their native country, would be higher than those in the current study. A caveat to this expectation, however, is that the completeness of records does not vary so much between countries as to affect the information available on which to base appropriateness judgements. An association between

Table 4 Pair-wise κ coefficients of concordance for appropriateness of day of care, by country of reviewer[a]

	Austria	France	Italy	Spain	Switzerland	UK
Austria		0.43	0.33	0.33	0.12	0.37
France	0.47		0.56	0.68	0.34	0.62
Italy	0.43	0.60		0.93	0.70	0.77
Spain	0.38	0.67	0.95		0.48	0.73
Switzerland	0.25	0.44	0.74	0.61		0.38
UK	0.48	0.67	0.72	0.73	0.49	

[a] Below the diagonal, pair-wise κ coefficients when the day of care was classified as appropriate or inappropriate; above the diagonal, pair-wise κ coefficients when the day of care was classified in three mutually exclusive categories (appropriate, inappropriate because the patient does not need a health service, or inappropriate because the patient needs another health service).

low completeness of the medical record and greater frequency of inappropriate hospital days was found in a study in Spain [25]. A comparison between concurrent review of patients still in hospital and retrospective review, performed simultaneously and independently by two reviewers on the same hospital stays, concluded that retrospective data collection produces higher rates of inappropriateness. The authors concluded that their finding was due, in part, to information available from care providers during the concurrent review that is not available in the medical record [26]. Even though concurrent review may be more labour-intensive than retrospective review, it is nevertheless feasible and may be warranted to ensure reliable findings. In addition, concurrent review has the advantage of enabling reducing inappropriate utilization experienced by extant hospital patients, while only inappropriate utilization of future patients can be reduced by retrospective review [27].

Conclusion

Further study is needed to generalize our findings: the use of a larger number of standard European medical records, and more than one reviewer per country could confirm our findings and obtain representative results of all potential reviewers from each European country. Our findings provide at least support to the application of the EU-AEP as a reliable instrument in the European setting. Of special interest is the application of this instrument to identify the factors that promote and those that inhibit appropriate hospital use in particular countries [28], thus enabling transference of the former to and avoidance of the latter in other countries.

Acknowledgements

This work was supported by a European Union Grant Biomed (BMH1 CT93 1053). The authors thank J. Alonso, E. Guallar and S. Peiró for comments on an earlier draft of this paper, and to all the members of the BIOMED research group on appropriateness of hospital use for their support and collaboration through the study.

References

1. Payne SMC. Identifying and managing inappropriate hospital utilization. A policy synthesis. *Health Serv Rev* 1987; **22:** 709–769.
2. Donabedian A. *Explorations in quality assessment and monitoring: the criteria and standards of quality.* Vol. II. Ann Arbor, MI: Health Administration Press, 1982.
3. Demlo LK. Assuring quality in health care: An overview. *Evaluation Health Professions* 1983; **6:** 161–197.
4. Payne SMC, Restuccia JD, Ash A *et al.* Using utilization review information to improve hospital efficiency. *J Health Hospital Serv Admin* 1991; **36:** 4.
5. Payne SMC, Ash A, Restuccia JD. The role of feedback in reducing medically unnecessary hospital use. *Med Care* 1991; **29:** 8.
6. Feldstein PJ, Wickizer TM, Wheeler JR. Private cost containment: the effects of utilization review programs on health care use and expenditures. *N Engl J Med* 1988; **318:** 310–314.
7. Santos-Eggimann B, Paccaud F, Blanc T. Medical appropriateness of hospital utilization: an overview of the Swiss experience. *Int J Qual Health Care* 1995; **7:** 227–232.
8. Alonso J, Muñoz A, Antó JM. Using length of stay and inactive days in the hospital to assess appropriateness of utilization in Barcelona, Spain. *J Epidemiol Community Health* 1996; **50:** 196–201.
9. Peiró S, Portella E. Validity of the protocol to evaluate inappropriate hospital use (in Spanish). *Med Clin (Barc)* 1996; **107:** 124–129.
10. Bañeres J, Alonso J, Broquetas J, Antó JM. Inappropriate hospital admissions and days of stay in neoplasic and chronic pulmonary obstructive disease patients (in Spanish). *Med Clin (Barc)* 1993; **100:** 407.
11. Apolone G, Fellin G, Tampieri A *et al.* Appropriateness of hospital use. Report from an Italian study. *Eur J Public Health* 1997; **7:** 34–39.
12. Liberati A, Apolone G, Lang T, Lorenzo S. A European project assessing the appropriateness of hospital utilization: background, objectives and preliminary results. *Int J Qual Health Care* 1995; **7:** 187–199.

13. Hunt SM, Alonso J, Bucquet D *et al.* Cross-cultural adaptation of health measures. *Health Policy* 1991; **19:** 33–44.

14. Guillemin F, Bombardier C, Beaton D. Cross-cultural adaptation of health-related quality of life measures: literature review and proposed guidelines. *J Clin Epidemiol* 1993; **46:** 1417–1432.

15. Lang T. A European version of the appropriateness evaluation protocol: goals and presentation. *Int J Tech Assess* 1999; **15:** 185–197.

16. Lorenzo S. Utilization review methods: limitations (in Spanish). *Med Clin (Barc)* 1996; **107:** 22–25.

17. Restuccia JD. Appropriateness evaluation protocol user's manual.

18. Dunn G. *Design and Analysis of Reliability Studies*. London: Edward Arnold, 1989:136–161.

19. Stuard A, Ord K. *Kendall's Advanced Theory of Statistics*. Vol. 1. *Distribution Theory*. 6th ed. London: Edward Arnold, 1994: pp. 365–368.

20. SAS Institute Inc. *SAS/STAT User's Guide,* V6, 4th edn. Cary, NC: SAS Institute Inc., 1990.

21. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33:** 159–174.

22. Strumwasser I, Paranjpe NV, Ronis DL *et al.* Reliability and validity of utilization review criteria. Appropriateness evaluation protocol, standardized review instrument and intensity–severity-discharge criteria. *Med Care* 1990; **28:** 95–111.

23. Gertman PM, Restuccia JD. The appropriateness evaluation protocol: a technique for assessing unnecessary days of hospital care. *Med Care* 1981; **19:** 855–871.

24. Rishpon S, Lubacsh S, Epstein LM. Reliability of a method of determining the necessity for hospitalization days in Israel. *Med Care* 1986; **24:** 279–282.

25. Ramos-Cuadra A, Marión-Buen J, García-Martín M *et al.* The effect of completeness of medical records on the determination of appropriateness of hospital days. *Int J Qual Health Care* 1995; **7:** 267–276.

26. Santos-Eggimann B, Sidler M, Schopfer D, Blanc T. Comparing results of concurrent and retrospective designs in a hospital utilization review. *Int J Qual Health Care* 1997; **9:** 115–120.

27. Restuccia JD. The effect of concurrent feedback in reducing inappropriate hospital utilization. *Med Care* 1982; **20:** 1.

28. Lorenzo S, Beech R, Lang T, Santos-Eggimann B. An experience of Utilization Review in Europe: sequel to a BIOMED Project. *Int J Qual Health Care* 1999; **11:** 13–19.

Accepted for publication 16 June 1999