*International Journal for Quality in Health Care* 2000; **Volume 12, Number 4**: pp. 325–329

# Liability and validity of the Appropriateness Evaluation Protocol in Turkey

SIDIKA KAYA[1], GÜLŞEN VURAL[2], KAFİYE EROĞLU[2], GÜLAY SAİN[3], HAKAN MERSIN[4], MELİH KARABEYOĞLU[5], KEREM SEZER[6], BEKTAŞ TÜRKKANI[7] AND JOSEPH D. RESTUCCIA[8]

[1]Takemi Program in International Health, Harvard School of Public Health, Boston, MA, USA, [2]School of Health Administration and School of Nursing and [3]Infectious Diseases Department, Hacettepe University Faculty of Medicine Hospital, Ankara, Turkey, [4]General Surgery Department, Oncology Hospital, Ankara, Turkey, [5]General Surgery Department, [6]Internal Medicine Department and [7]Obstetrics and Gynecology Department, Numune Hospital, Ankara, Turkey and [8]Health Care Management Program and Operations Management Department, School of Management, Boston University, Boston, MA, USA

## Abstract

**Objective.** To assess the inter-rater reliability between nurses and the convergent validity of the Appropriateness Evaluation Protocol (AEP) in the Turkish context.

**Methods.** Two nurses applied the original AEP concurrently to a random subsample of 335 patient-days in internal medicine, general surgery, and gynaecology departments at a university hospital and a government teaching hospital, as a part of a larger study. Inter-rater reliability was tested by calculating overall agreement and specific agreements between nurse reviewers' AEP assessments. Validity was tested by comparing the assessments of the nurses based on the AEP with the implicit judgements of five expert physicians on a random subsample of 818 patient-days. Sensitivity, specificity, positive and negative predictive values of the AEP were calculated. Reliability and validity were also evaluated by the $\kappa$ statistic.

**Results.** In the reliability test, there was a high level of agreement between the two independent raters applying the AEP in the three departments studied: overall agreement = 90.7–97.6%; specific inappropriate agreement = 69.1–92.3%; specific appropriate agreement = 88.3–96.6%. In validity testing, the AEP had a sensitivity of 0.83–0.97, specificity of 0.62–0.80, and positive and negative predictive values of 0.84–0.88 and 0.73–0.95 respectively. Kappa coefficients in internal medicine and gynaecology indicated almost perfect agreement in reliability testing and moderate agreement in validity testing. In general surgery, the $\kappa$ coefficients showed substantial agreement in both tests.

**Conclusion.** These results indicate that the AEP is a reliable and valid instrument to assess appropriateness of patient-days in Turkey.

**Keywords:** Appropriateness Evaluation Protocol, reliability, Turkey, utilization review, validity

Health expenditures in Turkey constituted 4% of the gross national product in 1998 [1], and from 1992 to 1996 inpatient care expenditures increased from 25% of total health expenditures to 29% [2]. Even though expenditures on health care in Turkey are lower than those in many developed countries, concern about the rising costs and limited efficiency of hospitals has been growing. Efficient and cost-effective use of resources is equally important for countries such as Turkey where resources allocated to health are so limited. During the last reform studies, a 'National Health Policy' document was produced for presentation to the Turkish Grand National Assembly. In this document, a financing principle of health services was stated as 'A mechanism to control the costs of services and limit demand according to needs should be developed' [3]. Utilization management based on utilization review can be instrumental in providing such a mechanism.

Implementing utilization review programmes in Turkey may yield solutions to problems of cost and efficiency. It is, however, crucial that such implementation is based on a method that is both reliable and valid in the context in which it is applied. The Appropriateness Evaluation Protocol (AEP) has gained widespread acceptance in performance of utilization review in the USA [4], and more recently, in many

Address reprint requests to S. Kaya, Takemi Program in International Health, Harvard School of Public Health, 665 Huntington Avenue, Building 1–1104, Boston, MA 02115, USA. E-mail: skaya01@hotmail.com

 

European countries [5–7]. It was shown to be reliable and valid in the USA context [8–10]. The current study was part of a larger study of the usefulness of this American protocol in Turkey. The objective of this study was to assess the inter-rater reliability between nurses and the convergent validity of the AEP in the Turkish context.

## Methods

In this study, the Turkish translation of the original AEP and adapted reasons list [11] were used. The study was conducted in two hospitals in Ankara, the capital city of Turkey. One of the hospitals was a large university hospital, and the other one was a large government teaching hospital. In the larger study, one-third of the patients hospitalized in internal medicine, general surgery and gynaecology departments on one randomly chosen day every month from March 1997 through February 1998 were reviewed concurrently. However, all gynaecological patients were reviewed in the government hospital because of the small number of patients. The unit of evaluation was a single hospitalization day of a patient who stayed in the hospital for at least 24 hours. The appropriateness of 2067 patient-days was evaluated by two nurses with PhDs. Before the reviews were conducted, the nurse reviewers were trained to apply the AEP by using the AEP reviewers' manual, and a baseline AEP competence was established.

Reliability and validity of the AEP were tested using two subsamples of the cases during the larger study. For assessing reliability of the AEP, a random subsample of 335 patient-days was reviewed by each nurse working alone. To assess validity of the AEP, one expert physician per department, except the gynaecology department at the university hospital, reviewed patient-days. The physician reviewers, who were selected by the chiefs of the departments, were all experienced clinicians committed to this study. Thus two internists, two general surgeons and a gynaecologist/obstetrician reviewed a random subsample of 818 patient-days within their own specialities according to their expert judgements, without using the AEP. The physicians were asked to judge whether each patient-day being studied was appropriate or inappropriate. They were blind to the AEP assessments of the nurses. All reviews were carried out concurrently.

Inter-rater reliability was tested by calculating the levels of overall agreement and specific agreement between nurse reviewer's assessments based on the AEP. Overall agreement is the proportion of judgements in which two reviewers agree. Specific inappropriate agreement for patient-days is defined as the proportion of patient-days (among those judged to be inappropriate by at least one of the two reviewers) that are rated as being inappropriate by both reviewers. Specific appropriate agreement is calculated in a similar way. In addition, overall agreement between nurses was evaluated by the kappa statistic, a measure of agreement that is corrected for chance agreement [12].

To test the convergent validity of the AEP (the extent to which decisions based on the instrument agree with those made by clinicians using expert judgement [13]) the assessments of the nurses based on the AEP were compared with those of the physicians. Sensitivity, specificity, positive and negative predictive values of the AEP were calculated. Physician assessments were used as the criterion standard in these analyses. Kappa coefficient was also calculated to evaluate the agreement between the assessments by the AEP and the physician's judgements. Landis and Koch's guidelines were used in interpreting $\kappa$ levels. According to these guidelines, $\kappa$ coefficients of between 0.41 and 0.60 are regarded as moderate, between 0.61 and 0.80 as substantial, and between 0.81 and 1.00 as almost perfect [14].

## Results

Overrides were used in 4.4% of assessments in the sample for reliability testing and in only 2.6% of assessments in the sample for validity testing. The reliability in internal medicine and general surgery, and the validity in general surgery and gynaecology were almost identical when using the override option or not using it. The reliability in gynaecology, however, was substantially lower when the override option was used ($\kappa$ without overrides = 0.94, $\kappa$ with overrides = 0.64; specific inappropriate agreement without overrides = 92.3%, specific inappropriate agreement with overrides = 57.1%). On the other hand, the specificity and $\kappa$ levels in internal medicine were increased somewhat with the utilization of overrides (specificity without overrides = 0.62, specificity with overrides = 0.72; $\kappa$ without overrides = 0.60, $\kappa$ with overrides = 0.65). Because overrides may be misused by inexperienced reviewers, may introduce the possibility of bias [15,16], and it could be argued that the instruments should be evaluated on their own without this 'subjective' reviewer influence [9], we refrained from using the override option. The assessments of appropriateness that are reported below reflect the judgements based strictly on the objective criteria alone.

The reliability results testing the level of agreement of the two nurse reviewers independently applying the AEP are shown in Table 1. In general, overall agreement on the assessments was very high (92.5%) and Cohen's kappa coefficient (0.80) indicated substantial agreement. The $\kappa$ value obtained was highly significant ($P < 0.0001$). Limiting comparison to only those patient-days assessed as inappropriate by at least one nurse (specific inappropriate agreement) gave a level of agreement of 74.5%. When the comparison was limited to only those patient-days assessed as appropriate by at least one nurse (specific appropriate agreement), level of agreement (90.5%) was found to be higher than specific inappropriate agreement.

Reliability testing by departments showed that there was a similar level of overall agreement between nurses in general surgery (90.7%), internal medicine (92.9%), and gynaecology (97.6%). Kappa coefficient indicated substantial agreement ($\kappa = 0.76$) in general surgery, and almost perfect agreement in internal medicine ($\kappa = 0.81$) and gynaecology ($\kappa = 0.94$). All $\kappa$ levels were statistically significant ($P < 0.0001$) indicating that agreement occurred more often than it would by chance

**Table 1** Inter-rater reliability of the AEP on patient-days by departments

| Reliability measure | Internal medicine[1] | General surgery[2] | Gynaecology[3] | All departments[4] |
|---|---|---|---|---|
| Overall agreement | 92.9% | 90.7% | 97.6% | 92.5% |
| Cohen's κ | 0.81[5] | 0.76[5] | 0.94[5] | 0.80[5] |
| 95% CI for κ | 0.70–0.92 | 0.63–0.88 | 0.83–1.00 | 0.73–0.88 |
| Specific agreement inappropriate | 74.4% | 69.1% | 92.3% | 74.5% |
| Specific agreement appropriate | 91.0% | 88.3% | 96.6% | 90.5% |

[1]$n = 154$. [2]$n = 140$. [3]$n = 41$. [4]$n = 335$. [5]$P < 0.0001$. CI, Confidence interval.

**Table 2** Validity of the AEP on patient-days when compared with the judgements of expert physicians by departments

| Validity measure | Internal medicine[1] | General surgery[2] | Gynaecology[3] | All departments[4] |
|---|---|---|---|---|
| Sensitivity | 0.93 | 0.97 | 0.83 | 0.93 |
| Specificity | 0.62 | 0.80 | 0.74 | 0.73 |
| Predictive value positive | 0.85 | 0.88 | 0.84 | 0.86 |
| Predictive value negative | 0.81 | 0.95 | 0.73 | 0.86 |
| Cohen's κ | 0.60[5] | 0.79[5] | 0.57[5] | 0.69[5] |
| 95% CI for κ | 0.50–0.69 | 0.73–0.86 | 0.42–0.71 | 0.63–0.74 |

[1]$n = 328$. [2]$n = 359$. [3]$n = 131$. [4]$n = 818$. [5]$P < 0.0001$. CI, Confidence interval.

alone. The specific inappropriate agreement level was higher in gynaecology (92.3%) than in general surgery (69.1%) and internal medicine (74.4%). The specific appropriate agreement levels were similar in all departments.

The validity of the AEP was tested by comparing the assessments of AEP reviewers with the 'gold standard' determinations of the expert physicians regarding appropriateness of patient-days (Table 2). When all departments were combined, the AEP had a sensitivity of 0.93, specificity of 0.73, and positive and negative predictive values of 0.86. Cohen's κ statistic (0.69) indicated substantial agreement.

For the three departments individually, the AEP had the highest sensitivity (0.97) and specificity (0.80) in general surgery, lowest sensitivity (0.83) in gynaecology, and lowest specificity (0.62) in internal medicine. Positive predictive values were similar in all departments (0.84–0.88). Negative predictive value was highest in general surgery (0.95) and lowest in gynaecology (0.73). Kappa coefficients in gynaecology (0.57) and internal medicine (0.60) showed moderate (borderline substantial) agreement while it showed substantial agreement (0.79) in general surgery.

## Discussion

The levels of overall agreement (92.5%) and specific inappropriate agreement (74.5%) found between nurse reviewers for all departments were remarkably similar to that reported by the developers of the AEP (94.3% and 79.3% respectively) [8]. The values of κ found in the reliability analysis (0.76–0.94) were higher than the values reported by previous investigators using the original AEP in the USA (0.59–0.73) [9,10], in Israel (0.59–0.63) [17] and in Spain (0.67) [18].

For all departments, the level of overall agreement between the nurses was similar to the levels found between two physicians (95.9%) and between each one of them and a nurse (93.4%; 94.4%) in a previous study in Turkey. Specific inappropriate agreement level between nurses, however, was higher than the levels between nurse–physician pairs (61.8%; 65.6%) in the previous study [11]. These findings show that nurses can classify patient-days as appropriate or inappropriate based on the AEP in a reliable manner in Turkey.

The degree of sensitivity and specificity observed in this study was compared with that reported in other studies. The sensitivity of the AEP achieved in this study (0.83–0.97) was similar to the sensitivity reported by Tsang and Severs [19] for geriatric admissions evaluated by the admitting physician using the AEP and by one of the six consultants regardless of the AEP in the UK (0.97), and by Kemper *et al* [20] for days evaluated by three fellows in paediatrics and a paediatric nurse practitioner using the paediatric AEP and by three experienced paediatricians – whose majority of subjective judgements was used as a gold standard – in the USA (0.93). The specificity observed in this study (0.62–0.80) was also similar to that reported by Kemper *et al.* (0.78), and by Tsang and Severs (0.63). Positive and negative predictive values of the AEP (0.84–0.88 and 0.73–0.95 respectively) were comparable with those reported by Tsang and Severs (0.95 and 0.75 respectively).

The κ values found in the validity analysis ranged from 0.57 in gynaecology to 0.79 in general surgery. These values were higher than those reported by Strumwasser *et al.* (0.31; 0.47) for comparisons of the AEP day of care criteria with the majority judgements of fee-for-service and HMO physician panels in the USA [9]. For all three departments, the observed value of κ was 0.69, which was similar to the validity score of the original AEP for day of care ($\kappa = 0.7$) [21].

The reliability and validity measures observed in this study indicate that the AEP is a reliable and valid instrument to assess appropriateness of patient-days in Turkey.

The main reservation relative to the use of the AEP in Turkey is the fact that the only alternate facility to an acute care hospital is a chronic care hospital. There are no nursing homes, home health agencies or hospices in Turkey. Providing such alternatives may be an option to reduce inappropriate use in Turkish hospitals. However, building and staffing new facilities may be more expensive in many locations than tolerating use of a small percentage of hospital beds for patients who do not need an acute level of care. Thus, optimizing one objective of the medical care system, such as appropriateness of hospital use, may result in suboptimization of other objectives, such as appropriateness of use of all levels of care. Careful consideration of such trade-offs must be made by health care decision-makers [8].

To apply AEP successfully in hospitals to reveal opportunities for improved utilization of services and to monitor progress toward more efficient operations, physicians' involvement is clearly necessary. No matter how enthusiastic administrators, managers and health service researchers are, the appropriateness of health services utilization cannot be achieved without the participation of the physicians who actually decide the utilization of services and perform the procedures. In general, the physicians contacted by the authors did not object to utilization review. In fact, the chief medical officer of the university hospital was very enthusiastic about learning the amount of inappropriateness in clinics. It should also be mentioned that the survey in the government hospital was facilitated by an approval from the Ministry of Health. Since the Ministry of Health hospitals are very centralized, such an official approval is necessary for utilization review to be applied in these hospitals.

The limited use of override option (3.2% in the larger study, that includes this study, and 3.1% in the previous Turkish study [11]) indicates that nurses felt rather comfortable in applying the AEP. In 93.5% of the days in the larger study, the reviewing nurse used only written information from the patient's medical records. In the previous study, however, the amount was nearly 25%. This may reflect the variability of the quality of the medical records among the hospitals in Turkey. So, it may be desirable to avoid reliance on the quality of the medical records, and retrospective review depending exclusively on records in Turkey. Even though concurrent review is more labour-intensive and more time consuming, it seems that it is more applicable in Turkey.

The AEP can also be used for planning hospital beds. In Turkey, hospital beds have been planned according to the bed/population ratio of 26:10 000. The ratio was 25.5 in 1998. However, overall bed occupancy rate for hospitals was only 59% in the same year. Moreover, bed occupancy rate was higher in chronic care hospitals, which were also limited in number [1]. Therefore, the validity of this bed/population ratio is doubtful. If the inappropriate utilization rate in acute care hospitals is also high, the bed/population ratio could be reduced, new beds might be built for chronic rather than for acute care, some acute care beds might be converted to chronic care beds or long-term care beds, or some hospitals might have to be closed which might be politically difficult and publicly unacceptable.

There are two limitations of the study that should be mentioned. First, the AEP was validated against the judgements of only one physician per department. It was possible that different results would be obtained if different physicians were selected or the judgements of a physician panel were taken as the gold standard. The fact that each reviewing physician was staff of the hospital where s(he) reviewed days of patients staying in the same hospital could be considered as another limitation. However, it is notable that the AEP was found to be valid in comparison to the departmental judgements of five different physicians (internal medicine, general surgery, and gynaecology at the general hospital; internal medicine and general surgery at the university hospital). The consistency of this finding supports the validity of the AEP in Turkey.

The second limitation of the study was that inter-rater reliability was assessed between two nurses with PhDs. Future research should address the issue of agreement between nurse raters who have different levels of education, and the validity of the AEP assessed by panels of physicians. The reliability and validity of the AEP admission criteria should also be addressed.

## Acknowledgements

This study was part of a project supported by a grant from the Hacettepe University Research Fund, and by a Fulbright scholarship and a Takemi fellowship to S.K.

## References

1. Ministry of Health, General Directorate of Curative Services. *Statistical Yearbook of Hospitals 1998* (in Turkish). Publication no: 619, Ankara, 1999.
2. Ministry of Health, Republic of Turkey, Health Projects General Coordination Unit. *Health Expenditures and Financing in Turkey 1992–1996* (in Turkish). Ankara: Eksen.
3. Republic of Turkey Ministry of Health, Health Project Coordination Unit. *National Health Policy of Turkey.* Ankara, 1993.
4. Restuccia JD. The Evolution of Hospital Utilization Review Methods in the United States. *Int J Qual Health Care* 1995; **7:** 253–260.
5. 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.

6. 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.
7. Lorenzo S, Lang T, Pastor R *et al.* Reliability study of the European Appropriateness Evaluation Protocol. *Int J Qual Health Care* 1999; **11:** 419–424.
8. Gertman PM, Restuccia JD. The Appropriateness Evaluation Protocol: A technique for assessing unnecessary days of hospital care. *Med Care* 1981; **19:** 855–871.
9. Strumwasser I, Paranjpe NV, Ronis DL *et al.* Reliability and validity of utilization review criteria Appropriateness Evaluation Protocol, Standardized Medreview Instrument, and Intensity-Severity-Discharge Criteria. *Med Care* 1990; **28:** 95–111.
10. Siu AL, Sonnenberg FA, Manning WG *et al.* Inappropriate use of hospitals in a randomized trial of health insurance plans. *N Engl J Med* 1986; **315:** 1259–1266.
11. Kaya S, Erdem Y, Doğrusöz S, Halıcı N. Reliability of a hospital utilization review method in Turkey. *Int J Qual Health Care* 1998; **10:** 53–58.
12. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Measure* 1960; **20:** 37–46.
13. Payne SMC. Identifying and managing inappropriate hospital utilization: a policy synthesis. *Health Serv Res* 1987; **22:** 709–769.
14. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* 1977; **33:** 159–174.
15. Kreger BE, Restuccia JD. Assessing the need to hospitalize children: Pediatric Appropriateness Evaluation Protocol. *Pediatrics* 1989; **84:** 242–247.
16. Werneke U, MacFaul R. Evaluation of appropriateness of paediatric admission. *Arch Dis Child* 1996; **74:** 268–273.
17. 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.
18. Baré ML, Prat A, Lledo L *et al.* Appropriateness of admissions and hospitalization days in an acute-care teaching hospital. *Rev Epidémiol Santé Publique* 1995; **43:** 328–336.
19. Tsang P, Severs MP. A study of appropriateness of acute care geriatric admissions and an assessment of the Appropriateness Evaluation Protocol. *J Roy Coll Physicians Lond* 1995; **29:** 311–314.
20. Kemper KJ, Fink HD, McCarthy PL. The reliability and validity of the Pediatric Appropriateness Evaluation Protocol. *Qual Rev Bull* 1989; **15:** 77–80.
21. Werneke U, Smith H, Smith IJ *et al.* Validation of the Paediatric Appropriateness Evaluation Protocol in British practice. *Arch Dis Child* 1997; **77:** 294–298.

**Accepted for publication 21 March 2000**