International Journal for Quality in Health Care 2002; **Volume 14, Number 6**: pp. 483–492

Metric properties of the appropriateness evaluation protocol and predictors of inappropriate hospital use in Germany: an approach using longitudinal patient data

OLIVER SANGHA[1,2,3]†, SEBASTIAN SCHNEEWEISS[4,5], MANFRED WILDNER[2,3], E. FRANCIS COOK[4,6], TROYEN A. BRENNAN[1,6], JENS WITTE[7] AND MATTHEW H. LIANG[1,6]

[1]Department of Health Policy and Management, [4]Department of Epidemiology, Harvard School of Public Health, Boston, MA, [2]Robert B. Brigham Arthritis & Musculoskeletal Diseases Clinical Research Center, [6]Division of General Internal Medicine, [5]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, MA, USA, [3]Department of Medical Informatics, Biometrics, and Epidemiology, Ludwig-Maximilians-University, School of Medicine, Munich, [7]Department of General Surgery, Central Hospital, Augsburg, Germany

Abstract

Background. The German health care system, renowned for its unrestricted access, high quality care, and comprehensive coverage, is challenged by increasing health care costs. This has been attributed partly to inefficiencies in the in-patient sector, but has been studied little. Attempts at quality improvement need to relate costs to outcomes. Until now, there has been no standardized methodology to evaluate the appropriateness of hospital care.

Objective. To develop and evaluate the metric properties of a method to assess inappropriate hospital care in Germany based on a widely used measure, the Appropriateness Evaluation Protocol (AEP).

Methods. The original AEP was translated and adapted to reflect differences in the provision of health care in Germany. Psychometric testing was performed in a stratified sample of all patients admitted to the Departments of Medicine and Surgery of a 400-bed teaching hospital during 1 year. Three board-certified physicians participated in each department to evaluate intra-rater reliability, while two additional independent physicians judged inter-rater reliability.

Results. Inter-rater agreement for the evaluation of hospital days among surgical patients was 84% (80–87%), with an average kappa value of 0.58 (0.48–0.68). Corresponding figures for patients in medicine were 76% (73–80%) with a κ value of 0.42 (0.34–0.42). Inter-rater agreement for hospital admissions and κ was 74% (62–86%) and 0.44 (0.21–0.67) in surgery, and 92% (85–100%) and 0.31 (0–0.80) in medicine, respectively. Thirty-three per cent of all admissions and 28% of consecutive hospital days were judged inappropriate in surgery; among medicine patients, reviewers found 6% of admissions and 33% of hospital days inappropriate. Time since admission was the strongest predictor of inappropriate hospital use adjusted for length of stay, comorbidity, age, and gender.

Keywords: Appropriateness Evaluation Protocol, appropriateness of care, Germany, health services research, quality improvement, utilization review

National health expenditures in developed countries have increased enormously as a result of the explosion of expensive medical technology and its dissemination, which has increased public expectations, and an aging population. Even the wealthiest nations are compelled to cut back on costs and re-examine their assumptions about the coverage of health

Address reprint requests to S. Schneeweiss, Brigham and Women's Hospital and Harvard Medical School, Division of Pharmacoepidemiology and Pharmacoeconomics, 221 Longwood Ave (BLI-341), Boston, MA 02115, USA. E-mail: schneeweiss@post.harvard.edu

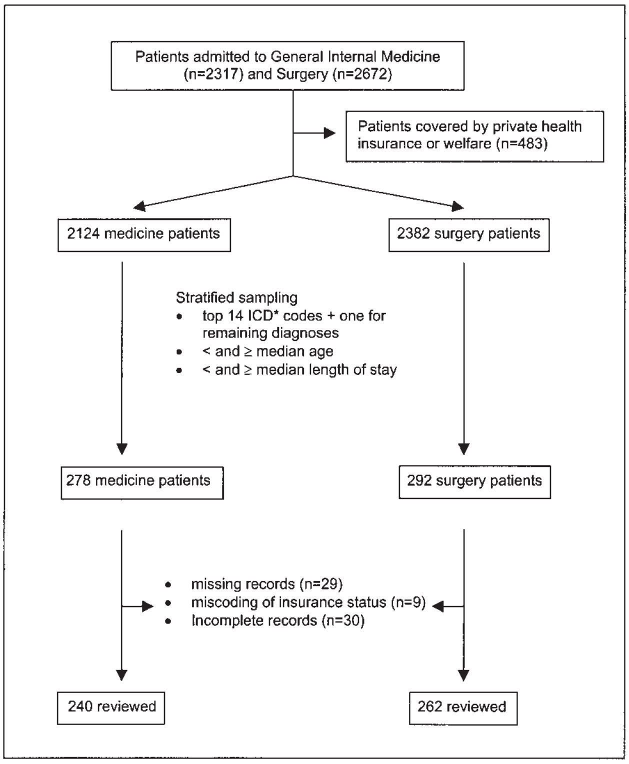
† Deceased.

services. Since the single largest budget item for medical care until recently was hospital care, efforts have been made to reduce costs in this sector, particularly through avoiding unnecessary and prolonged hospitalizations. Eliminating unnecessary hospitalizations may also improve the quality of care and increase access to care for patients on waiting lists in some health care systems [1].

Substantial rates of inappropriate hospital care have been reported in several countries [2–13], but few centers have been able to reduce inappropriate care using the results of such studies [12,14,15]. Avoiding inappropriate hospital care is difficult, in part because its determinants are multifactorial and include to various degrees practice patterns, patient characteristics, the organization of in-hospital care, the coordination between hospital care and other providers within the health care system, and financial incentives.

The German health care system, internationally renowned for unrestricted access, high quality care and comprehensive coverage, is particularly challenged, since the number of hospitalizations per capita and average length of stay (LOS) is substantially higher than in most industrialized countries. Hence, there are increasing efforts to reduce hospital admissions and inappropriate in-patient care. Beginning in 1995, the sickness funds commissioned statewide review organizations (MDK, comparable to the US Physician Review Organizations) to conduct systematic reviews of selected hospital departments. However, the review methodology was developed ad hoc and has not been tested for reliability and validity.

The aim of this study was to develop and evaluate the metric properties of a method to assess the appropriateness of hospital care in Germany based on a widely used measure, the Appropriateness Evaluation Protocol (AEP).


Patients admitted to General Internal Medicine (n=2317) and Surgery (n=2672)
Patients covered by private health insurance or welfare (n=483)
2124 medicine patients
2382 surgery patients
Stratified sampling
• top 14 ICD* codes + one for remaining diagnoses
• < and ≥ median age
• < and ≥ median length of stay
278 medicine patients
292 surgery patients
• missing records (n=29)
• miscoding of insurance status (n=9)
• Incomplete records (n=30)
240 reviewed
262 reviewed

* International Classification of Diseases 9th revision

Figure 1 Patient sampling.

Methods

Modification of the AEP

The methodology to assess the appropriateness of in-patient care in Germany was based on the AEP, developed by Gertman and Restuccia in 1981 [2], and later modified [16] and validated as a technique to assess unnecessary days of hospital care. The instrument has been implemented in many hospitals and has proven useful [6,8,12,16–18]. Methodological studies of the AEP have focused on its reliability and validity [2,8,19–23], adaptation of the original protocol to different kinds of hospitals, units within hospitals, or diagnostic groups [3,19,24], or the denominators used to obtain inappropriateness rates [6].

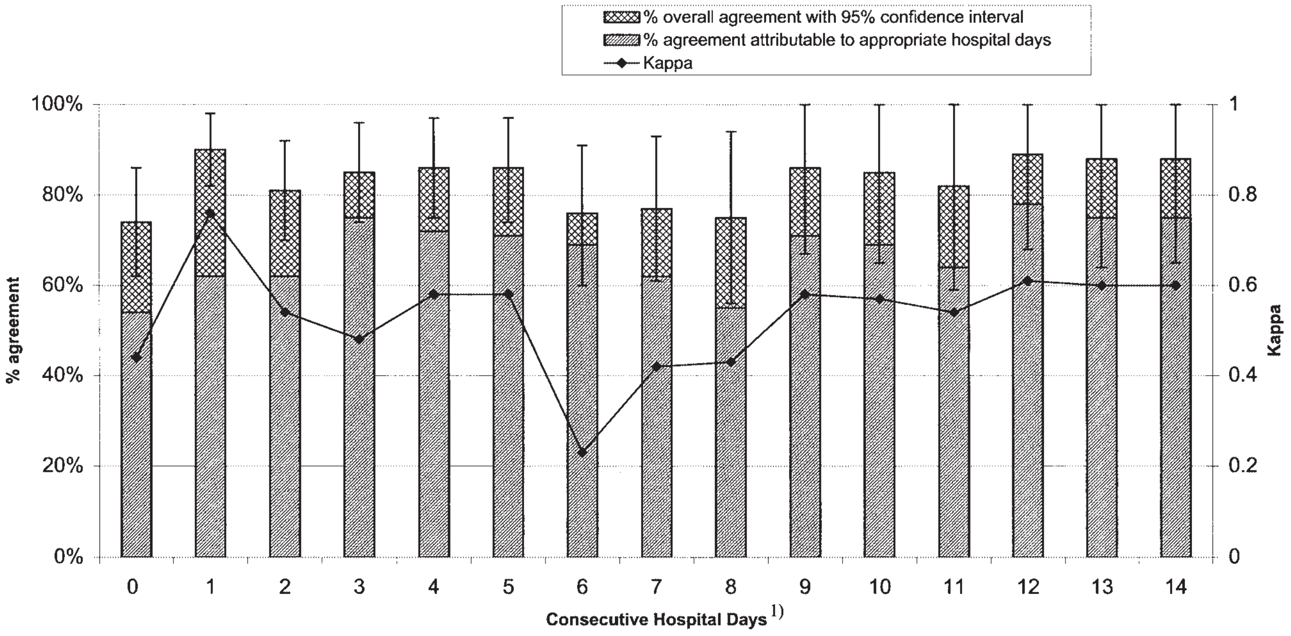
The AEP uses 27 criteria to assess the appropriateness of each hospital day (11 relate to medical services/procedures, seven to nursing/life support services, and nine to clinical characteristics of the patient necessitating close observation). Once a day has been identified as medically unnecessary (i.e. no information in the medical record corresponding to any of the 27 explicit criteria), the AEP also allows the description of factors potentially responsible for such medically unnecessary episodes of care using a complementary list of reasons and alternatives (‘reasons list’).

The AEP also rates the appropriateness of hospital admissions using 17 criteria, which pertain to clinical stability of the patient, necessity of medical interventions, and planned surgical procedures within 24 hours. An admission is deemed appropriate if one or more of these criteria are satisfied. The AEP allows an ‘override’ option, enabling the reviewer to evaluate an admission or particular hospital day as ‘appropriate’, despite the absence of one criterion. Conversely, it may be classified as ‘inappropriate’, even if one or more of the criteria are met.

The original instrument was translated into German by a native English speaker who was fluent in German, and by a German health care professional who was fluent in English.

All items were reviewed by an expert panel, including representatives from professional societies of surgery, geriatrics, and internal medicine, from different State MDKs and from the Department of Epidemiology of the University of Munich, who provided methodological expertise. Changes were considered whenever criteria were not German health care practices, or health services not provided in Germany. New criteria were added if newer developments in practicing medicine were not considered in the original AEP.

The instrument was then converted to a computerized version, allowing reviewers online access and providing multiple pull-down menus to ease chart abstraction and to permit



1) Day "0" represents the hospital admission

Figure 2 Distribution of overall and specific inter-rater agreement in 54 surgical patients.

data entry in a database with automated plausibility checks. The instrument is available on the internet [25].

Study design and patient sample

The study was carried out at a 400-bed teaching hospital in Frankfurt, Germany, which volunteered its data. All patients admitted to general internal medicine and surgery were eligible for retrospective evaluation with the AEP. A stratified sample, with 60 strata based on age (less than, and greater than or equal to median age), hospital length of stay (less than, and greater than or equal to median LOS), 14 most prevalent diagnoses, and one stratum including the remaining diagnoses, was drawn from discharge lists. Prior to sampling, patients covered by private health insurance or welfare were removed from the list since review of medical records by the MDKs to assess appropriateness was only allowed for patients covered by national health insurance (~90% of the population). The sampling is displayed in Figure 1. Complete hospital stays of each sampled patient, including admission and all hospital days, were studied.

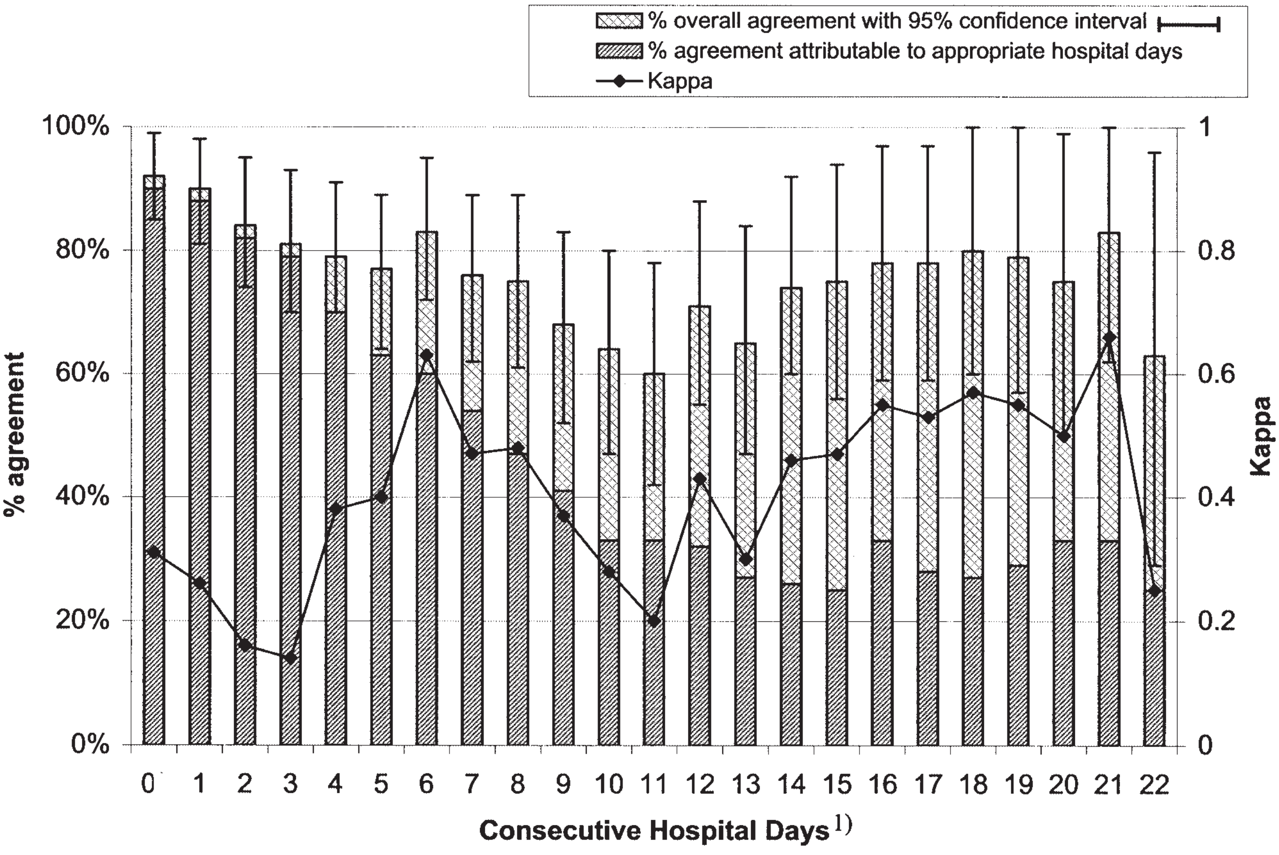
Three board-certified internists (reviewing medical patients) and three board-certified surgeons (reviewing surgical patients) who are employed by the MDK of the State of Hessen conducted the reviews. Reviewers were instructed to base their admission assessment only on the basis of the medical information relative to the day of admission and the following 24 hours. After 3 weeks, two internists and two surgeons were asked to review 25 medical records each to assess intra-rater reliability. None of the reviewers was aware of this reliability exercise at the time of the initial review. To examine inter-rater-reliability, four additional reviewers (two board-certified internists, two surgeons) from the MDK of the State of Bavaria reviewed 30 charts each, previously abstracted by physicians from the Hessen MDK.

Statistical analyses

Sample size estimates were based on the ability to detect a rate of inappropriate hospitalizations per bed days of 15% ± 3% [proportion ± standard error] with a statistical power of 90%.

Reliability between reviewers (inter-rater) and within one reviewer (intra-rater) was calculated for admission and each hospital day in three ways: (1) overall agreement by dividing the number of patients where both reviewers agreed on appropriateness of a day by all patients on that specific day (e.g. admission, day 1 and so forth); (2) specific agreement by dividing the number of patients where both reviewers agreed on appropriateness by the total number of patients; and (3) overall agreement between pairs of reviewers by Cohen's kappa statistic [26], which adjusts for the amount of agreement occurring by chance alone. It should be noted that the κ value may be paradoxically low as observed in these data when overall agreement is high, particularly when the prevalence of appropriate days is low, also as observed in these data [27].

Previous reports on the reliability of instruments measuring appropriateness of hospital care treated all hospital-days in the same way, disregarding whether they were the first, second or the last days a patient stayed in a hospital. No justification was provided for why inter-rater reliability should be the



1) Day "0" represents the hospital admission

Figure 3 Distribution of overall and specific inter-rater agreement in 49 general internal medicine patients.

Table 1 Proportions of inappropriate admissions and hospital days in Departments of General Internal Medicine and Surgery

	Patients[1]	Days	Proportion inappropriate (%)	95% confidence limits	
				Lower (%)	Upper (%)
Surgery					
Inappropriate admissions	269	N/A	33	30	36
Inappropriate hospital days	243	2277	28	24	32
Internal medicine					
Inappropriate admissions	245	N/A	6	0	12
Inappropriate hospital days	222	3200	33	29	34

[1]The number of patients eligible for hospital day evaluation has been reduced by the number of patients leaving on the same day.

same at the beginning, the middle, and the end of a stay. Therefore, we calculated reliability for individual days in the course of a hospitalization (i.e. for the admission day, day 1, day 2, . . .) [2,8,23]. Since there are increasingly fewer patients with longer hospital stays, we restricted the reliability evaluation to those days for which we could identify at least eight patients (days 1–22 in general internal medicine, days 1–14 in surgery). Overall and specific percentage agreement was also calculated for the sum of all hospital days. Overall, κ for hospital days is computed according to Fleiss [28], after testing whether the κ values are equal among days with Cochran's Q-test [29]. Landis and Koch suggested that κ

Table 2 Frequency of inappropriate hospital days by patient characteristics among surgical patients

	Patients, *n* (%)	Total hospital days	Proportion inappropriate days	95% confidence interval[1]
Total	243 (100)	2277 (100)	27.8	23.7–31.9
Sex				
Male	119 (49.0)	1087 (47.7)	30.2	23.9–36.5
Female	124 (51.0)	1190 (52.3)	25.5	20.0–31.0
Age (years)				
≤49	100 (41.2)	773 (33.9)	28.6	21.5–35.7
50–59	41 (16.9)	320 (14.1)	20.4	10.6–30.2
60–69	41 (16.9)	394 (17.3)	33.1	23.0–43.2
70–79	31 (12.8)	459 (20.2)	24.3	13.7–34.9
80–89	23 (9.5)	245 (10.8)	36.0	25.8–46.2
≥90	7 (2.9)	86 (3.8)	18.0	0.2–35.8
Appropriate admission				
Yes	170 (70.0)	1643 (72.2)	22.1	18.2–26.0
No	73 (30.0)	634 (27.8)	44.1	34.5–53.7[2]
Length of stay (days)				
1–7	103 (42.4)	397 (17.4)	32.5	25.0–40.0
8–14	91 (37.4)	775 (34.0)	25.0	19.0–31.0
15–21	35 (14.4)	586 (25.7)	23.5	14.6–32.4
22–28	7 (2.9)	158 (6.9)	27.1	7.8–46.4
≥29	7 (2.9)	361 (15.9)	19.2	5.9–32.5
Number of comorbid conditions				
0	187 (77.0)	1505 (66.1)	29.5	24.6–34.4
1	44 (18.1)	600 (26.4)	21.6	13.0–30.2
2	10 (4.1)	127 (5.6)	27.3	12.4–42.2
≥3	2 (0.8)	45 (2.0)	7.4	4.6–10.2[2]

[1] 95% confidence limits were calculated with the empirical standard deviation of the mean proportion and normal approximation [31].
[2] $P < 0.05$.

values of >0.75 may be taken to represent excellent agreement, while values <0.4 may represent poor agreement [30].

We explored further the association between age, gender, comorbidity, and hospital length of stay, and inappropriate hospital admission. Stratified analyses of inappropriate hospital days also included admission status (appropriate, inappropriate). The mean of patient-specific proportions of inappropriate hospital days over all days is an unbiased estimate of the overall proportion, taking the dependence of hospital days within patients into account; 95% confidence limits were calculated with the empirical standard deviation of the mean proportion and normal approximation [31]. To compute the overall proportion of inappropriate hospital admissions/stays, proportions were re-weighted according to the stratified sample weights. Predictors of inappropriate admissions were analyzed by fitting a multiple logistic regression model to the data [32]. Because hospital days for a given patient and their appropriateness are not independent, confidence limits were adjusted by generalized estimating equations [33]. We assumed a first-degree autoregressive covariance structure and estimated empirical standard errors.

All analyses were performed on a personal computer using the SAS statistical software package version 6.10 [34].

Results

The following changes were made to the criteria for admission of the original AEP after extensive review of the expert panel. Three criteria were omitted from the list of medical services ('thoracentesis or paracentesis that day', 'any test requiring strict dietary control'), one criterion from the list of nursing/life support services ('intramuscular and/or subcutaneous injections at least twice daily') and one criterion from the list of patient conditions ('coma for at least 1 hour') because they were considered as not requiring in-patient therapy, each being redundant or too infrequent to make up a single criterion. One criterion was modified to allow a post-operative day for any procedure with a high risk of developing clinical complications. Two items were omitted from the list for appropriate admissions ('wound dehiscence or evisceration', 'intramuscular antibiotics at least every 8 hours') because these are performed as outpatient services for most patients.

Sample characteristics

During 1997, 2317 patients were admitted to the Department of General Internal Medicine and 2672 to the Department

Table 3 Frequency of inappropriate hospital days by patient characteristics among general internal medicine patients

	General internal medicine patients, *n* (%)	Total hospital days	Proportion inappropriate days	95% confidence interval[1]
Total	222 (100)	3200 (100)	33.0	28.9–37.1
Sex				
Male	103 (46.4)	1540 (48.1)	25.3	19.9–30.7[2]
Female	119 (53.6)	1660 (51.9)	39.6	33.9–45.3
Age (years)				
≤49	38 (17.1)	360 (11.3)	37.2	26.1–48.3
50–59	30 (13.5)	475 (14.8)	33.1	21.6–44.6
60–69	44 (19.8)	608 (19.0)	40.0	30.0–50.0
70–79	40 (18.0)	543 (16.9)	25.4	18.0–32.8
80–89	54 (24.3)	857 (26.8)	30.0	22.0–38.0
≥90	16 (7.2)	357 (11.2)	33.1	18.9–47.3
Appropriate admission				
Yes	210 (94.6)	3075 (96.1)	31.3	27.4–35.2
No	12 (5.4)	125 (3.9)	60.4	35.5–85.3[2]
Length of stay (days)				
1–7	44 (19.8)	125 (3.9)	23.3	12.1–34.5
8–14	84 (37.8)	845 (26.4)	36.6	30.4–42.8
15–21	49 (22.1)	836 (26.1)	35.6	27.5–43.7
22–28	21 (9.5)	550 (17.2)	28.7	17.6–40.0
≥29	21 (9.5)	844 (26.4)	36.9	24.1–49.7
Number of comorbid conditions				
0	50 (22.5)	542 (16.9)	39.3	29.6–49.0
1	49 (22.1)	604 (18.9)	28.7	20.3–37.1
2	71 (32.0)	1155 (36.1)	33.5	26.3–40.7
≥3	52 (23.4)	899 (28.1)	30.0	22.7–37.3

[1] 95% confidence limits were calculated with the empirical standard deviation of the mean proportion and normal approximation [31].
[2] $P < 0.05$.

of Surgery. One hundred and ninety-three medicine patients and 290 surgery patients were covered by private insurance or welfare, and therefore not eligible for review by state MDKs. Sample-size calculations indicated 278 medicine patients and 292 surgery patients as the necessary target numbers to detect a rate of 15% (±3%) inappropriate hospital episodes. In internal medicine and surgery, 242 and 260 records were available for review, respectively. The difference between the numbers in the target samples and those that were reviewed was due to missing records ($n = 29$), miscoding of insurance status ($n = 9$), or incomplete medical records ($n = 30$).

Patients in medicine had an average age of 66.1 (±18.1) years, were 53.7% female, and had an average LOS of 14.5 (±11.5) days. Surgical patients were 54.1 (±20.9) years old, 58.9% female, and had an average LOS of 9.9 (±10.4) days.

Inter-rater reliability among surgeons

Fifty-four patients were available for the inter-rater agreement substudy among surgical patients. Overall and specific agreement and κ statistics between two different reviewers, including 95% confidence limits for individual hospital days, are shown in Figure 2. Day 0 refers to the judgement of whether an admission was appropriate or not. Taking all hospital days (excluding the admission day) into account, overall agreement is 84% (80–87%), with an average κ value of 0.58 (95% CI 0.48–0.68). Kappa values did not differ among days ($Q_K = 8.8$, $P = 0.79$). The corresponding values for the reliability of admission appropriateness are 74% (62–86%) overall agreement, and κ 0.44 (0.21–0.67).

Inter-rater reliability among internists

Inter-rater agreement was calculated among 49 medicine patients. Figure 3 shows overall and specific agreement as well as κ statistics for admissions and consecutive hospital days. Overall agreement between two reviewers was 76% (73–80%), with an average κ value of 0.42 (0.34–0.49) for all hospital days, and 92% (85–100%) with a κ value of 0.31 (0–0.80) for admissions. Kappa values did not differ among days ($Q_K = 15.2$, $P = 0.82$).

Intra-rater reliability among surgeons

Intra-rater reliability was assessed for 51 surgery and 49 medicine patients. Medical records were reviewed twice, 3

Table 4 Patient- and day-specific characteristics associated with inappropriateness of hospital days, in 243 patients admitted to a Department of Surgery, totaling 2633 days of hospitalization

	Odds ratio	95% confidence interval[1]	P trend
Sex			
Male	1	–	
Female	1.2	0.90–1.67	
Age (years)			
≤ 49	1	–	
50–59	0.78	0.43–1.4	
60–69	1.3	0.77–2.2	
70–79	1.2	0.63–2.3	
80–89	2.0	1.1–3.6[2]	
≥ 90	0.67	0.16–2.83	
Appropriateness of admission (inappropriate versus appropriate)	1.4	0.92–2.2	
Length of stay (days)			
1–7	1	–	<0.0001
8–14	0.58	0.36–0.92[2]	
15–21	0.3	0.15–0.60[2]	
22–28	0.27	0.11–0.71[2]	
≥ 29	0.13	0.06–0.33[2]	
Number of comorbid conditions			
0	1	–	
1	0.69	0.40–1.18	
2	1.1	0.55–2.39	
≥ 3	0.66	0.23–1.87	
Days since admission			
1–2	1	–	<0.0001
3–5	1.8	1.3–2.4[2]	
6–8	2.9	2.0–4.3[2]	
9–19	4.4	2.8–7.0[2]	
≥ 20	7.1	3.4–15[2]	

[1] 95% confidence limits were adjusted for dependence between hospital days within patients using generalized estimating equations [33].
[2] $P < 0.05$.

weeks apart, by two surgeons and two internists. Agreement was good for the two surgeons with 88% (79–97%) overall agreement for admissions and 88% (85–92%) for hospital days, and corresponding κ values of 0.60 (0.75–1) and 0.73 (0.65–0.81), respectively. One surgeon, however, had throughout worse reliability results, although the κ value failed to differ significantly ($Q_K = 20.5$, $P = 0.058$).

Intra-rater reliability among internists

Intra-rater reliability between the two internists was excellent with overall agreement of 96% (88–100%) for admissions and 93% (91–95%) for hospital days. The corresponding κ values were 0.65 (0.2–1) and 0.82 (0.77–0.88), respectively. Both internists had comparable results, although one internist did not judge any admissions among his sample ($n = 25$) inappropriate; hence it was not possible to calculate a corresponding κ statistic.

Predictors of inappropriate care

Thirty-three per cent of all admissions and 28% of consecutive hospital days were judged inappropriate among 269 surgical patients (Table 1). Age, gender, and hospital LOS was not associated with the proportion of inappropriate hospital days. In surgical patients, the highest rates of inappropriate care were among the very old (80–89 years), and among patients with either short (1–7 days) or longer (>3 weeks) hospital stays (Table 2). Patients with an inappropriate admission had significantly more unnecessary consecutive hospital days. When we restricted the analysis of proportions of inappropriate hospital days to patients with an appropriate admission, unnecessary hospital days were found only in 15.7% of patients with a length of stay <1 week.

Six per cent of all admissions and 33% of consecutive hospital days were judged inappropriate among 240 patients admitted to the Department of General Internal Medicine

Table 5 Patient- and day-specific characteristics associated with inappropriateness of hospital days, in 222 patients admitted to a Department of General Internal Medicine, totaling 3509 days of hospitalization

	Odds ratio	95% confidence interval[1]	*P* trend
Sex			
Male	1	–	
Female	1.5	1.01–2.3[2]	
Age (years)			
≤49	1	–	
50–59	0.88	0.44–1.8	
60–69	1.1	0.57–2.2	
70–79	0.95	0.49–1.9	
80–89	0.96	0.49–1.9	
≥90	1.4	0.48–3.9	
Appropriateness of admission (inappropriate versus appropriate)	0.61	0.20–1.9	
Length of stay (days)			
1–7	1	–	<0.0001
8–14	0.70	0.31–1.5	
15–21	0.35	0.15–0.80[2]	
22–28	0.15	0.06–0.40[2]	
≥29	0.16	0.06–0.42[2]	
Number of comorbid conditions			
0	1	–	
1	0.61	0.34–1.1	
2	0.56	0.32–0.98[2]	
≥3	0.63	0.34–1.2	
Days since admission			
1–2	1	–	<0.0001
3–5	8.1	4.3–15[2]	
6–8	25.1	13–50[2]	
9–19	74	38–142[2]	
≥20	150	76–293[2]	

[1] 95% confidence limits were adjusted for dependence between hospital days within patients using generalized estimating equations [33].
[2] $P < 0.05$.

(Table 3). Women had a higher proportion of inappropriate hospital days than men. Inappropriateness was equally distributed among age groups, with highest rates among the 60–69 year olds and among patients ≤49 years. Patients with an appropriate admission had a higher proportion of inappropriate consecutive hospital days; however, only 14 of the 240 admissions were considered unnecessary. Hospital length of stay was significantly associated with inappropriateness of care. This effect was pronounced when we restricted the analysis to patients who were appropriately admitted (length of stay 1–7 days = 12.3 inappropriate hospital days; 8–14 days = 32.2%; 15–21 days = 33.4%; 22–28 days = 24.9%; ≥29 days = 38.9).

When we simultaneously adjusted for gender, age, appropriateness of admission, length of stay, and number of comorbid conditions, there was a consistent increase in the probability of inappropriate hospital days with an increase in the time since admission. In the Department of Surgery, having an inappropriate admission or more than three comorbid conditions increased the likelihood of inappropriate days (Table 4). However, in medicine patients, length of hospital stay appears to be a protective factor and being female increases the risk of inappropriate days 1.5 times (Table 5).

Discussion

Internationally, the AEP has emerged as the most commonly used instrument in the assessment of hospital care. The strengths and limitations of the instrument are well documented in the United States, where it was developed and evaluated [2,4,16,18,24,35,36]. In Germany, identification of inappropriate hospital care has just emerged as a candidate for achieving substantial savings within health care. Regional review organizations (MDKs) have initiated systematic utilization reviews of hospital departments; however, the methodology and instruments applied were ad hoc in nature and were never formally tested for psychometric properties.

The aim of this study was to adapt and validate the AEP to evaluate appropriateness of hospital care in Germany. Review of the original instrument and adaptation to the German health care system was done by representatives from the Societies of Internal Medicine, Surgery and Geriatrics, and from the regional review organizations, which are formally in charge of standardized utilization reviews on behalf of the German sickness funds. Although several items in the original AEP were omitted, in the German version the diagnosis-independence and the explicitness of the definitions was preserved. Adaptation of the original instrument for other countries (i.e. Italy [6], Portugal [37], the French-speaking part of Switzerland [38], and Turkey [39]) has been reported.

The German AEP retains good reliability comparable to the findings of the original AEP [4,8,24]. In contrast to the methods used in previous studies, we calculated reliability not only for hospital admissions and days of hospital stay in total, but also for each consecutive day of a hospital course. Overall agreement in judging the appropriateness of admissions and hospital days was high, particularly in surgical patients, with percentage agreement rates around or exceeding 80%. Agreement was consistently good for appropriate and inappropriate days, underscoring the utility of the AEP for standardized utilization review.

The proportion of inappropriate hospital days was high, both in general internal medicine and in surgery. These results are similar to those of other studies [2,4,8,40] and not surprising for the German health care system, where hospital care is basically free of charge and hospital reimbursement is predominantly per diem. German doctors are not required to justify the length of an individual hospital stay, and hence there is an incentive to prolong hospitalization when there are empty beds. Not surprisingly, occupancy rates in Germany exceed 79.8% where there are 6.7 acute care hospital beds per 1000 inhabitants (1997) [41]. Corresponding figures for the US are 63.0% and 3.3 per 1000, respectively. In Germany, substantial efforts are currently under way to optimize the delivery of hospital care, including the introduction of prospective payment in 2003.

On the other hand, our rates of inappropriate hospital care are similar to other countries that have used the AEP to evaluate the appropriateness of hospital care. Siu found 23% of all adult admissions in six US sites to be unnecessary [4]. Findings from European studies range from 2.7% and 4.2% in Switzerland [38] and 14.2% in Italy [42], to 19% [9] and 27.6% [43] in Spain. Corresponding figures for the proportion of inappropriate hospital days range between 11.2% and 15.3% in Switzerland [38], 35% in the US [4], 37.3% in Italy [42], and 31.7% [9] and 43.9% [43] in Spain. A study from Israel revealed 38.9% inappropriate hospital days in surgery and 54% in medicine [44]. It is difficult to generalize from these findings, because rates of inappropriate care are influenced by multiple factors such as differences in delivery and financing of health care, physicians' practice styles, patients' behavior, and expectations or study methodology.

We explored several predictors of inappropriate hospital days. However, as reported earlier [6,45], only hospital length of stay was clearly associated with inappropriate hospital use. This effect was even more pronounced when we restricted our analysis to patients who were appropriately admitted, since patients with short hospitalizations undergoing medical interventions that could be performed as outpatient procedures had substantially higher proportions of inappropriate care.

In our approach of evaluating all hospital days from each sampled patient, it is necessary to adjust for potential correlation among day-specific assessments performed in the same patient using standard statistical techniques. This requires a sample size that is slightly greater than would be necessary if each day-specific assessment were independent of all other days of that patient. Other studies randomly selected patient-days out of the pool of all hospital days of the entire patient population. The latter approach is statistically more efficient, but many more patient records must be identified. Our sampling is the most process-efficient approach, independent of the type of medical record (paper or electronic).

In conclusion, our study documented a substantial proportion of inappropriate hospital use under the current system of per diem reimbursement for hospital services. Our sampling approach may be the most process-efficient way to assess this proportion. With the forthcoming implementation of a prospective payment system in Germany based on diagnosis-related groups (DRG), the AEP will be used increasingly by supervisory agencies to assess the appropriateness of admissions in combination with audits of DRG coding. For the purpose of clinical quality management, a slightly modified AEP should be used to evaluate inappropriately early discharges after the change in payment system.

References

1. Brennan TA, Leape LL, Laird NM *et al.* Incidence of adverse events and negligence in hospitalized patients. Results of the Harvard Medical Practice Study I. *N Engl J Med* 1991; **324:** 370–376.

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

3. Winickoff RN, Restuccia JD, Fincke BJ. Concurrent application of the Appropriateness Evaluation Protocol to acute admissions in Department of Veteran Affairs Medical Centers. *Med Care* 1991; **29:** AS64–AS75.

4. 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.

5. Gloor JE, Kissoon N, Joubert GI. Appropriateness of hospitalization in a Canadian pediatric hospital. *Pediatrics* 1993; **91:** 70–74.

6. Apolone G, Alfieri V, Braga A *et al.* A survey of the necessity of the hospitalization day in an Italian teaching hospital. *Qual Assur Health Care* 1991; **3:** 1–9.

7. Bare ML, Prat A, Lledo L, Asenjo MA, Salleras L. Appropriateness of admissions and hospitalization days in an acute-care teaching hospital (in French). *Rev d'Epidem de Santé Publ* 1995; **43:** 328–336.

8. 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.

9. Alonso J, Munoz A, Anto JM. Using length of stay and inactive days in the hospital to assess appropriateness of utilisation in Barcelona, Spain. *J Epidemiol Commun Health* 1996; **50:** 196–201.

10. Henley L, Smit M, Roux P, Zwarenstein M. Bed use in the medical wards of Red Cross War Memorial Children's Hospital, Cape Town. *S Afr Med J* 1991; **80:** 487–490.

11. Hynes M, O'Herlihy BP, Laffoy M, Hayes C. Patients 21 days or more in an acute hospital bed: appropriateness of care. *Irish J Med Sci* 1991; **160:** 389–392.

12. Payne SM, Ash A, Restuccia JD. The role of feedback in reducing medically unnecessary hospital use. *Med Care* 1991; **28:** AS91–AS106.

13. Chopard P, Perneger TV, Gaspoz J-M *et al.* Predictors of inappropriate hospital days in a department of internal medicine. *Int J Epidemiol* 1998; **27:** 513–519.

14. Restuccia JD. The effect of concurrent feedback in reducing inappropriate hospital utilization. *Med Care* 1982; **20:** 46–62.

15. Vardi A, Modan B, Blumstein Z, Lusky A, Schiff E, Barzilay Z. A controlled intervention in reduction of redundant hospital days. *Int J Epidemiol* 1996; **25:** 604–608.

16. Restuccia JD, Kreger BE, Payne SM, Gertman PM, Dayno SJ, Lenhart GM. Factors affecting appropriateness of hospital use in Massachusetts. *Health Care Fin Rev* 1986; **8:** 47–54.

17. Payne SM, Campbell D, Penzias BG, Socholitzky E. New methods for evaluating utilization management programs. QRB. *Qual Rev Bull* 1992; **18:** 340–347.

18. Booth BM, Ludke RL, Wakefield DS *et al.* Nonacute days of care within Department of Veterans Affairs medical centers. *Med Care* 1991; **29:** AS51–AS56.

19. Payne SM. Identifying and managing inappropriate hospital utilization: a policy synthesis. *Health Serv Res* 1987; **22:** 709–769.

20. Siu AL, Leibowitz A, Brook RH, Goldman NS, Lurie N, Newhouse JP. Use of the hospital in a randomized trial of prepaid care. *J Am Med Assoc* 1988; **259:** 1343–1346.

21. Davido A, Nicoulet I, Levy A, Lang T. Appropriateness of admission in an emergency department: reliability of assessment and causes of failure. *Qual Assur Health Care* 1991; **3:** 227–234.

22. Kemper KJ, Fink HD, McCarthy PL. The reliability and validity of the pediatric appropriateness evaluation protocol. QRB. *Qual Rev Bull* 1989; **15:** 77–80.

23. Strumwasser I, Paranjpe NV, Ronis DL, Share D, Sell LJ. Reliability and validity of utilization review criteria. Appropriateness Evaluation Protocol, Standardized Medreview Instrument, and Intensity–Severity–Discharge criteria. *Med Care* 1990; **28:** 95–111.

24. Kemper KJ. Medically inappropriate hospital use in a pediatric population. *N Engl J Med* 1988; **318:** 1033–1037.

25. Bavarian Public Health Research Center, University of Munich. *AEP—Appropriateness Evaluation Protocol* (German version): http://mfv.web.med.uni-muenchen.de/aep.html Accessed 7 December 2002.

26. Cohen J. A coefficient of agreement for nominal scales. *Educ Pschol Measure* 1960; **20:** 37–46.

27. Feinstein AR, Cichetti DV. High agreement but low kappa: 1. The problem of two paradoxes. *J Clin Epidemiol* 1990; **43:** 543–549.

28. Fleiss JL. *Statistical Methods for Rates and Proportions*. New York: Wiley, 1981.

29. Cochran WG. The comparison of percentages in matched samples. *Biometrica* 1950: 256–266.

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

31. Armitage P, Berry G. *Statistical Methods in Medical Research*. London: Blackwell, 1994.

32. Hosmer DW, Lemeshow S. *Applied Logistic Regression*. New York: Wiley, 1989.

33. Liang KY, Zeger SL. *Analysis of Longitudinal Data*. Oxford: Clarendon Press, 1994.

34. SAS Institute Inc. *SAS/STAT User's Guide*. Release 6.10 edition. Cary, NC: SAS Institute Inc., 1996.

35. Siu AL, Manning WG, Benjamin B. Patient, provider and hospital characteristics associated with inappropriate hospitalization. *Am J Publ Health* 1990; **80:** 1253–1256.

36. Restuccia JD, Gertman PM, Dayno SJ. A comparative analysis of appropriateness of hospital use. *Health Aff* 1984; **3:** 130–138.

37. Bentes M, Gonsalves ML, Santos M, Pina E. Design and development of a utilization review program in Portugal. *Int J Qual Health Care* 1995; **7:** 201–212.

38. 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.

39. Kaya S, Erdem Y, Dogrusoz S, Halici N. Reliability of a hospital utilization review method in Turkey. *Int J Qual Health Care* 1998; **10:** 53–58.

40. Mozes B, Halkin H, Katz A, Schiff E. Reduction of redundant hospital stay through controlled intervention. *Lancet* 1987; **25:** 968–969.

41. OECD. *OECD Health Data 1999—a Comparative Analysis of 29 Countries*. Paris: Credes, 1999. Http://www.oecd.org. Accessed 7 December 2002.

42. Angelillo IF, Ricciardi G, Nante N, Boccia A, Group AC. Appropriateness of hospital utilization in Italy. *Publ Health* 2000; **114:** 9–14.

43. De la Fuente OD, Peir S, Marchan C. Inappropriate hospitalization. *Eur J Publ Health* 1996; **6:** 126–132.

44. Mozes B, Schiff E, Modan B. Factors affecting inappropriate hospital stay. *Qual Assur Health Care* 1991; **3:** 211–217.

45. Paldi V, Porath A, Friedman L, Mozes B. Factors associated with inappropriate hospitalization in medical wards: a cross-sectional study in two university hospitals. *Int J Qual Health Care* 1995; **7:** 261–265.

Accepted for publication 29 July 2002