Pergamon

International Journal for Quality in Health Care, Vol. 9, No. 2, pp. 115–120, 1997

Printed in Great Britain

PII: S1353-4505(97)00001-X

1353-4505/97 $17.00 + 0.00

Comparing Results of Concurrent and Retrospective Designs in a Hospital Utilization Review

B. SANTOS-EGGIMANN,* M. SIDLER,† D. SCHOPFER‡ and T. BLANC†

* Department of Social and Preventive Medicine, University of Lausanne, Lausanne, Switzerland; † Public Health Department, State of Vaud, Lausanne, Switzerland; ‡ St-Loup Hospital, St-Loup/Orbe, Switzerland.

Hospital utilization reviews are performed on the basis of lists of explicit criteria, such as the Appropriateness Evaluation Protocol, both concurrently and retrospectively, in an increasing number of settings as part of efforts to improve the performance of hospitals and to reduce health care costs. Retrospective data collection has advantages in terms of expenses and ease of sampling, but relies on the quality of medical records. We report on a comparison between concurrent and retrospective data collection performed simultaneously and independently by two reviewers on the same hospital stays in the regional St-Loup Hospital. Results suggest that retrospective data collection produces higher rates of inappropriate hospital utilization, due to a limited number of criteria that are recorded concurrently, but are not found in the retrospective reading of medical records. These results should encourage a further investigation of the comparability between concurrent and retrospective designs in other settings. © **1997 Elsevier Science Ltd. All rights reserved.**

Key words: Concurrent design, retrospective design, hospital utilization.

INTRODUCTION

In the late 1970s, instruments for hospital utilization reviews based on explicit criteria were developed in the United States in order to help identify hospital admissions or days of stay that were medically appropriate or not [1]. From the first experiences aiming at controlling hospital costs generated by Medicare patients, several lists of explicit criteria were tested. One of them, the Appropriateness Evaluation Protocol (AEP), was developed by Gertman and Restuccia [2] and diffused overseas. Most of the hospital utilization reviews conducted in the 1980s and early 1990s in Europe were based on adaptations of the AEP required by the local environments [3]. Overviews of national experiences based on such lists of explicit criteria have shown that the rates of inappropriate days vary widely between and within countries, and rates of medical inappropriateness reaching 38% of the admission days and 66% of the later days of stay were reported in retrospective hospital utilization studies [4]. The diversity of methodological approaches is likely to explain part of this variation [5,6].

The concurrent (during the patient's stay) or retrospective (relying on medical records) design of data collection is a major component of the methodology in hospital utilization reviews based on explicit criteria. Indeed, a retrospective data collection aiming at verifying the fulfilment of explicit criteria requires the excellence of medical charts. In such circumstances, the presence or absence of each of the criteria that can justify a hospital day should be systematically reported in the patient's file. This condition is crucial to the use of an instrument that elicits a judgement of medical appropriateness based on the fulfilment of one single criterion. When the quality of medical charts is unsure, a retrospective utilization of the AEP might measure the quality of the information reported in medical charts as much as the appropriateness of hospital days under study, and the hypothesis can be made that the rate of inappropriate days is overestimated. On the other hand, an argument against concurrent hospital utilization reviews is the difficult access to medical charts during the patient's stay. The search for medical charts in hospital wards, the verification that recorded information has been updated, and the conduct of additional interviews with medical teams are resource consuming. The more intensive resource consumption in concurrent reviews might be justified by the fact that the expected changes in hospital practice induced by concurrent reviews will affect the current patients, whereas retrospective reviews will merely modify the management of future patients. Beyond the stronger impact on the behavior of medical teams expected from concurrent reviews, the superiority of concurrent over retrospective reviews to estimate the rate of inappropriate days has to be evaluated to justify the investment of such resources.

This study, based on data gathered at the St-Loup

Received 28 May 1996; accepted 5 December 1996.
Correspondence to: Brigitte Santos-Eggimann, Division des Services de santé, Institut universitaire de médecine sociale et préventive, 17, Rue du Bugnon, CH-1005 Lausanne, Switzerland. Tel. +41 21 314 72 52; Fax +41 21 314 73 73.

Hospital (Canton of Vaud, Switzerland) during a permanent concurrent hospital utilization review:

— investigates whether concurrent or retrospective data collection produces similar estimates of the rate of inappropriate hospital utilization; the hypothesis is that a higher rate of inappropriateness is estimated in a retrospective data collection;

— examines judgements based on concurrent and retrospective data collection and seeks to determine for which criteria the two methods produce different results; the underlying hypothesis is that most discordant judgements obtained by the two methods might be explained by days considered as appropriate in a concurrent review based on one single criterion that is not constantly detected retrospectively, based on a reading of medical charts;

— compares the agreement between two reviewers in concurrent and in retrospective designs. The hypothesis here is that agreement in retrospective reviews might be higher, since the information available to reviewers in retrospective reviews is essentially the same, written in medical charts, and is not defined by the active search of information through interviews that reviewers judged necessary or not, as required in a concurrent design.

MATERIAL AND METHODS

The St-Loup Hospital is a 114-bed acute care regional hospital; in addition, a special unit was recently created within the hospital to provide further care and rehabilitation services to patients discharged after an acute episode. This hospital pioneered in 1989 in the canton of Vaud a hospital utilization review performed concurrently with the specific aim to improve the hospital's performance and to reduce its mean length of stay. This initiative was suggested and supported by the Public Health Department of the State of Vaud, which subsidizes regional non-profit hospitals based on a global prospective budget. In this first experience with the AEP, the original instrument was adapted by senior physicians after a chart review process. Changes were then approved by the Department of Public Health and the resulting list of criteria was finally tested for acceptability in the same hospital. The St-Loup Hospital has continuously conducted a concurrent review of admissions and stays since 1992. At the end of 1993, a utilization review was performed independently by two reviewers to test the reliability of the instrument; the first reviewer (Reviewer 1) was a nurse hired by the hospital to conduct the permanent review underway since 1992 (DS), and the second (Reviewer 2) was a physician hired by the public health department (MS) who was involved in the conduct of reviews in other hospitals based on the modified AEP since 1991. In order to explore our hypotheses, the same reviewers were asked 1 year later to study the same admissions and stays, retrospectively and independently from each other. This re-review was designed to compare

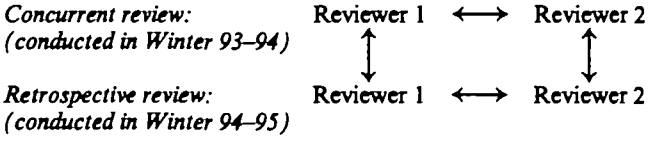

FIGURE 1. Comparison scheme. ↕Inter-reviewer comparisons; ↕ intra-reviewer compositions.

the results obtained by each reviewer in a concurrent and retrospective review of the same set of hospital days (Fig. 1, intra-reviewer comparisons). The level of agreement between the two reviewers was also studied in both circumstances (Fig. 1, inter-reviewer comparisons). Data were analysed by the Division of Health Services in the Department of Social and Preventive Medicine of the University of Lausanne.

Each second admission in the medicine, surgery, urology, orthopedics, gynecology, ear nose and throat, and ophthalmology departments from 6 December to 19 December 1993 and from 6 January to 31 January 1994 was included in the study sample; the break between the two periods was due to the sickness leave of one of the two reviewers. All days (admissions and days of stay) of the patients admitted in the sample were analysed, after exclusion of a limited number of days when a patient temporarily left the hospital.

The hospital utilization review was performed on the basis of the adapted AEP in Appendix. A first adaptation by the St-Loup Hospital in 1989 ended in a single list of 24 explicit criteria that applied both to admission days and to days of stay, and the override option was maintained [7]. The list was then reduced to 21 criteria used in the permanent concurrent review; reviewers could report the presence of up to 10 criteria. The final judgement integrates the override option described in the original AEP.

The process of concurrent review differed slightly between the two reviewers: both completed the review based on medical charts and additional interviews with the medical staff, but Reviewer 1 was on site continuously while Reviewer 2 collected the information on site once a week. The two reviewers were unaware of the retrospective data collection that would be performed later, and both reviewers performed the retrospective data collection on the same sample 1 year later.

We did not compute statistical tests of significance for the rate of inappropriateness, since the consecutive days in a stay cannot be considered as independent observations, and the appropriateness of one day is likely to be linked to the appropriateness of the previous and the next day. Rationales for analysing all the days of stay rather than one single day randomly selected in each stay have been discussed elsewhere [8,9]. Results shown in this paper should be considered as exploratory. The levels of agreement between reviewers or between methods of data collection were described by Cohen's Kappa statistic [10]. Tests of statistical significance are reported for the Kappa statistic according to Fleiss [11], since we hypothesized

TABLE 1. Rate of inappropriate admission days ($n = 155$) and days of stay ($n = 880$), by reviewer and by method of data collection

Method	Reviewer 1		Reviewer 2	
	Number inappropriate	Percentage inappropriate	Number inappropriate	Percentage inappropriate
Admission day				
Concurrent	7	4.5	14	9.0
Retrospective	14	9.0	24	15.5
Day of stay				
Concurrent	91	10.3	187	21.3
Retrospective	177	20.1	265	30.1

TABLE 2. Number of criteria recorded, by reviewer and by method of data collection

Number of criteria recorded	Reviewer 1				Reviewer 2			
	Concurrent data collection		Retrospective data collection		Concurrent data collection		Retrospective data collection	
	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
0	102	9.9	193	18.6	198	19.1	280	27.1
1	252	24.3	304	29.4	436	42.1	336	32.5
2	385	37.2	198	19.1	181	17.5	169	16.3
3	104	10.0	133	12.9	138	13.3	128	12.4
4	73	7.1	97	9.4	31	3.0	64	6.2
5	35	3.4	65	6.3	19	1.8	32	3.1
6–10	84	8.1	45	4.3	32	3.0	26	2.5
Total	1035	100.0	1035	100.0	1035	100.0	1035	100.0

that the relationship expected between consecutive days of a single stay in terms of their appropriateness does not necessarily imply that the assumption of independence is violated as far as reviewer's assessments are concerned.

RESULTS

The sample included 162 patients admitted over the enrollment period, totaling 1098 hospital days (mean length of stay: 6.7 days). The medical chart was not found for the retrospective review of seven stays (57 hospital days), which were excluded from analyses, and 6 days in three stays were further excluded, when patients were temporarily discharged and the hospital reserved the bed. Consequently, analyses were performed on 155 acute care stays and 1035 days, admission days included.

Intra-reviewer concurrent vs retrospective comparisons (cf Fig. 1)

Level of inappropriate hospital utilization. For both reviewers, the estimates of the rates of inappropriate admissions and of inappropriate days of stay were lower in the concurrent than in the retrospective approach (Table 1). The judgement in concurrent and retrospective designs was concordant in 908 days for Reviewer 1 (agreement 87.7%; $\kappa 0.50$, $P < 0.0001$), and in 877 days for Reviewer 2 (agreement 84.7%; $\kappa 0.58$, $P < .0001$). In both cases, the level of agreement beyond chance can be considered as fair to good, according to Landis and Koch guidelines for interpretation of Kappa values between 0.40 and 0.75 [12]. Over the 1035 days under study, the mean number of recorded criteria was higher for Reviewer 1 who registered more criteria in the concurrent (2.41) than in the retrospective data collection (2.09). Reviewer 2 recorded a mean of 1.60 criteria in both circumstances.

Study of criteria explaining the disagreement between concurrent and retrospective designs. Table 2 shows that 60–74% of the days judged appropriate were justified by one or two criteria only. The proportion of days justified by a single criterion in the concurrent data collection was 26.9% ($n = 252$ days) for Reviewer 1 and 51.9% for Reviewer 2 ($n = 436$ days, of which 433 were considered appropriate after use of the override option). Among the 252 days characterized by a single criterion in his concurrent review, Reviewer 1 found no criterion retrospectively in 82 days (32.5%). Overall, in 7.9% of the 1035 days studied by Reviewer 1, one single criterion justified the days in a concurrent approach, and no criterion was found retrospectively. Of the 436 days in which Reviewer 2 found one single criterion concurrently, 101 (23.2%) were characterized by the absence of a criterion in his retrospective review (9.7% of the 1035 days under study).

TABLE 3. Frequency of report for each criterion,* by reviewer and by method of data collection ($n = 1035$ days)

	Reviewer 1				Reviewer 2			
	Concurrent data collection		Retrospective data collection		Concurrent data collection		Retrospective data collection	
Criterion*	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
1	139	13.4	134	12.9	138	13.3	143	13.8
2	16	1.5	5	0.5	10	1.0	28	2.7
3	4	0.4	8	0.8	4	0.4	6	0.6
4	70	6.8	26	2.5	41	4.0	31	3.0
5	74	7.1	35	3.4	61	5.9	25	2.4
6	76	7.3	76	7.3	54	5.2	65	6.3
7	121	11.7	125	12.1	118	11.4	112	10.8
8	242	23.4	289	27.9	166	16.0	212	205
9	0	0.0	0	0.0	0	0.0	2	0.2
10	24	2.3	18	1.7	8	0.8	6	0.6
11	484	46.8	453	43.8	439	42.4	376	36.3
12	132	12.8	64	6.2	88	8.5	22	2.1
13	876	84.6	640	61.8	481	46.5	505	48.8
14	6	0.6	9	0.9	3	0.3	4	0.4
15	15	1.4	16	1.5	1	0.1	0	0.0
16	76	7.3	156	15.1	7	0.7	84	8.1
17	46	4.4	49	4.7	13	1.3	30	2.9
18	13	1.3	0	0.0	2	0.2	0	0.0
19	1	0.1	8	0.8	0	0.0	0	0.0
20	71	6.9	55	5.3	24	2.3	9	0.9
21	10	1.0	0	0.0	0	0.0	0	0.0

*cf. Appendix.

Many hospital days were characterized by the fulfilment of criterion number 11: treatment of major surgical or trauma wound, including care of surgical site, and/or presence of drains or catheters (except permanent urinary catheters), or 13: patient who requires close clinical monitoring by a nurse at least three times per day, according to written medical prescription (Table 3). Criterion 11 was recorded in 36–46.8% of the days. By contrast, criterion 13 was mentioned by Reviewer 1 in 84.6% of the days in the concurrent review and in 61.8% of the days retrospectively, whereas Reviewer 2 reported the presence of this criterion in 46.5%, respectively 48.8% of the days. When, in the concurrent review, a day was judged as appropriate on the basis of one single criterion, the criterion reported by Reviewer 1 ($n = 252$ days characterized by one single criterion) was criterion 13 in 78.2% (Table 4). In the same situation, the day was considered as appropriate by Reviewer 2, based on the presence of criterion 11 in 49.0% or criterion 13 in 39.7% of 433 days. Finally (as summarized in Fig. 2), among the 82 days characterized, according to Reviewer 1, by the presence of one single criterion with a concurrent data collection and judged inappropriate with a retrospective approach, 80 were related to criterion 13, and two to criterion 1. Among the 100 similar situations recorded by Reviewer 2, 58 were attributable to criterion 13, and 42 to criterion 11. Overall, the 80 days justified in the concurrent data collection of Reviewer 1 by the report

TABLE 4. Days judged appropriate based on one single criterion: frequency of criteria reported

	Reviewer 1				Reviewer 2			
	Concurrent data collection ($n = 252$)		Retrospective data collection ($n = 303$)		Concurrent data collection ($n = 433$)		Retrospective data collection ($n = 327$)	
Criterion reported*	Number	Percentage	Number	Percentage	Number	Percentage	Number	Percentage
1	51	20.2	49	16.2	28	6.5	52	15.9
6	—	—	1	0.3	—	—	2	0.6
7	—	—	9	9.0	—	—	10	3.1
8	—	—	—	—	17	3.9	7	2.1
10	—	—	1	0.3	4	0.9	4	1.2
11	2	0.8	118	38.9	212	49.0	119	36.4
13	197	78.2	120	39.6	172	39.7	132	40.4
16	—	—	5	1.7	—	—	1	0.3

*cf. Appendix.

	Reviewer 1	*Reviewer 2*
Number of days reviewed	1035	1035
	↓	↓
Number of days appropriate in concurrent review	937	834
	↓	↓
... with one single criterion reported concurrently	252	436
	↓	↓
.... with no criterion reported retrospectively	82	101
	↓	↓
Criterion implied: 1	*2*	*1*
11	—	*42*
13	*80*	*58*

FIGURE 2. Contribution of days justified by one single criterion to the assessment of appropriateness.

of criterion 13 amounted to three-quarters (73.4%) of the 109 days judged appropriate concurrently and inappropriate retrospectively by this reviewer. Criterion 11 or 13 recorded by Reviewer 2 as a single criterion in the concurrent data collection explained 83% of the 121 days in which this reviewer found that criteria were present only in his concurrent data collection.

Inter-reviewer comparisons (cf. Fig. 1)

As presented in Table 1, Reviewer 1 constantly rated as inappropriate a lower proportion of both admission and hospital days than Reviewer 2. Whereas the overall agreement (percentage of cases on which both reviewers agree on an appropriateness decision) was similar in the concurrent (86.8%) and in the retrospective (86.3%) design, specific agreement (agreement rate on only those cases that were judged inappropriate by at least one reviewer) was higher in the retrospective data collection (54.3%, vs 37.2% in the concurrent review), leading to a higher Kappa statistic with the retrospective data collection (0.62, $P<0.0001$ vs 0.48, $P<0.0001$ in concurrent design).

DISCUSSION

The goal of this study was not to estimate the true rate of inappropriateness, but to explore the difference between concurrent and retrospective data collection. The rates that we found were low when compared with other published results [3, 4, 7,13–16], but differences in sampling frames make such comparisons difficult. The St-Loup Hospital was characterized by a low mean length of stay and had 5 years experience of hospital utilization review specifically intended to lower the rate of inappropriate hospital use. In addition, rates found in concurrent reviews performed in three other regional hospitals of the same state in 1991 did not exceed 15%, and other authors published low rates based on the original AEP [17,18]. We found higher rates of inappropriateness according to the reviewer hired by the Public Health Department than according to the hospital reviewer; this observation was also made in previous reliability studies [8]. A possible explanation is the more constant presence of the hospital reviewer in the wards, which could have facilitated the collection of relevant information about criteria indicating that a day is appropriate. The on-site reviewer might also have a more lenient interpretation of hospital records, and feel that his interpretation was legitimized by his knowledge of the way that events are reported in patients' charts.

Our findings were consistent with the hypotheses of higher estimates resulting from a retrospective data collection and with a higher level of agreement in retrospective reliability studies. Most of the differences between concurrent and retrospective data collections were associated with criteria 11 (treatment of major surgical or trauma wound, including care of surgical site, and/or presence of drains or catheters, except permanent urinary catheters) and 13 (patient who requires close clinical monitoring by a nurse at least three times per day, according to written medical prescription); these criteria were abstracted from the original AEP without change (criterion 13) or with minor adaptations (criterion 11). Our study has limitations: it was performed in a single setting, on days of stay that cannot be considered as independent observations since the sampling involved all the days in the selected stays, and with an adapted instrument. However, our results strongly suggest that further attention should be devoted to the comparison of estimates based on concurrent and retrospective data collections. The replication of similar findings could lead to improvements in selected items of an instrument that is spreading quickly, in a context of growing interest for quality assurance. At a local level, the comparison between concurrent and retrospective approaches is a valuable tool to detect the weaknesses of hospital charts, and could lead to necessary improvements before a retrospective data collection is preferred on legitimate cost considerations.

Acknowledgements: The authors are grateful to Prof. J. D. Restuccia for his helpful comments on the manuscript.

REFERENCES

1. Payne, S. M., Identifying and managing inappropriate hospital utilization: A policy synthesis. *Health Services Research* 1987; **22**: 709–769.
2. Gertman, P. M. and Restuccia, J. D., The Appropriateness Evaluation Protocol: A technique for assessing unnecessary days of hospital care. *Medical Care* 1981; **19**: 855–871.
3. Bentes, M., Gonsalves, M., Santos, M. and Pina, E., Design and development of an utilization review program in Portugal. *International Journal for Quality in Health Care* 1995; **7**: 201–212.
4. Liberati, A., Apolone, G., Lang, T. and Lorenzo, S., A European project assessing the appropriateness of hospital utilization: Background, objectives and preliminary results. *International Journal for Quality in Health Care* 1995; **7**: 187–199.
5. Ash, A., The design and analysis of hospital utilization studies. *International Journal for Quality in Health Care* 1995; **7**: 245–252.
6. Winterhalter, G., Blanc, T. and Kulczyki, E., Importance et causes de l'utilisation inappropriée identifiée à l'Hôpital de St-Loup. Service de la santé publique du Canton de Vaud (Cah Rech Doc IUMSP), Lausanne, 1991.
7. Felin, G., Apolone, G., Tampieri, A., Bevilacqua, L., Meregalli, G., Minella, C. and Liberati, A., Appropriateness of hospital use: An overview of Italian studies. *International Journal for Quality in Health Care* 1995; **7**: 219–225.
8. Santos-Eggimann, B., Hospital utilization reviews under field conditions: potential and improvements. *International Journal of Technology Assessment in Health Care* 1993; **9**(4): 514–521.
9. Peiro, S., Perez, S. and Portella, E., Independent observation in the review of sequential days clustered by stay [letter] and Santos-Eggimann B. Reply. *International Journal of Technology Assessment in Health Care* 1994; **10**: 720–722.
10. Cohen, J. A., A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 1960; **20**: 37–46.
11. Fleiss, J. L., The measurement of interrater agreement. In: *Statistical Methods for Rates and Proportions*, Fleiss, J. L. (Ed.), pp. 212–236. Wiley, New York, 1981.
12. Landis, J. R. and Koch, G. G., The measurement of observer agreement for categorical data. *Biometrics* 1977; **33**: 159–174.
13. Lorenzo, S. and Sunol, R., An overview of Spanish studies on appropriateness of hospital use. *International Journal for Quality in Health Care* 1995; **7**: 213–218.
14. Lang, T., Davido, A., Logerot., H. and Meyer, L., Appropriateness of admissions: The French experience. *International Journal for Quality in Health Care* 1995; **7**: 233–238.
15. Baré, M. L., Prat, A., Lledo, L., Asenjo, M. A. and Saileras, L. L., Appropriateness of admissions and hospitalization days in an acute-care teaching hospital. *Revue d'Epidémiólogie et de Santé Publique* 1995; **43**: 328–336.
16. Siu, A., Sonnenberg, F. A., Willard, G., Goldberg, G. A., Bloomfield, E. S., Newhouse, J. P. and Brook, R. H., Inappropriate use of hospitals in a randomized trial of health insurance plans. *New Engl Journal of Medicine* 1986; **315**: 1259–1266.
17. Santos-Eggimann, B., Paccaud, F. and Blanc, T., Medical appropriateness of hospital utilization: An overview of the Swiss experience. *International Journal for Quality in Health Care* 1995; **7**(3): 227–232.
18. Rishpon, S., Epstein, L. M. and Rennet, H., Unnecessary hospitalization days: Rates in two general hospitals in Israel. *Israeli Journal of Medical Science* 1989; **25**: 392–397.

APPENDIX

List of explicit criteria, adapted from the Appropriateness Evaluation Protocol by the St-Loup Hospital

Criteria linked to medical procedures

(1) Surgical procedure the same day or the next day.

(2) Any test requiring strict dietary control or realimentation underway.

(3) Any specialized investigation that could not be performed on an ambulatory basis.

(4) Treatment that requires frequent dose adjustments under direct medical supervision.

(5) Close medical monitoring by a doctor at least twice a day.

Criteria linked to paramedical services

(6) Patient admitted to intensive care unit.

(7) Respiratory care, administration of oxygen, bird, intensive respiratory physiotherapy.

(8) Parenteral therapy with any supplementation.

(9) Chemotherapy under direct medical supervision at least twice a day.

(10) Intramuscular and/or subcutaneous injections at least three times per day.

(11) Treatment of major surgical or trauma wound, including care of surgical site, and/or presence of drains or catheters (except permanent urinary catheters).

(12) Intake and output measurement.

(13) Patient who requires close clinical monitoring by a nurse at least three times per day.

Criteria linked to patient condition

(14) Cardiac frequency < 50 or > 140 per minute.

(15) Blood pressure: systolic > 90 or < 200 mmHg and/or diastolic < 60 or > 120 mmHg.

(16) Severe and/or symptomatic abnormality of a blood test, electrolytes or blood gases; symptomatic metabolic acute disorder; acute aggravation of a chronic metabolic disorder.

(17) Persistent fever with a minimum 38°C axillary temperature of at least 5 days' duration or appearing during hospitalization.

(18) Recent acute confusional state (not due to alcohol abuse).

(19) Other acute or recently aggravated neurological disorder (not due to alcohol withdrawal).

(20) New acute documented myocardial infarction.

(21) Cerebral ischemia.