Use of Severity of Illness to Evaluate Quality of Care

THOMAS A. MACKENZIE, ALISON GREENAWAY-COATES, MARINA S. DJURFELDT and WILMA M. HOPMAN

Case Mix Research Group, Department of Community Health and Epidemiology, Queen's University, Kingston, Ontario, Canada K7L 3N6

This study compared a severity of illness system (APACHE II) and a 10% random sample of charts in terms of their ability to identify cases with quality problems. Using condition-specific data bases of 337 pneumonia, 363 acute myocardial infarction and 266 hip fracture charts, severity of illness information was used to separate cases into those with a high and a low likelihood of a poor outcome. Cases with low admission severity of illness combined with subsequent death were flagged as potential quality problems. Physician evaluation was used as the gold standard to measure flag performance. Flags were tested against a 10% random sample drawn from within the three condition-specific data bases. Analyses focused on a combination of sensitivity and positive predictive value. The low severity plus death flag performed much better than a 10% random sample approach, suggesting that outcomes monitoring flags based on severity of illness could play an important role in screening cases for potential quality problems.

Subsequent research has focused on both process and outcomes issues, since good quality of care includes not only an appropriate or acceptable outcome, but also the activities that take place during the hospitalizations. This begins with pre-admission activities and moves through the spectrum of patient care, up to and including discharge planning, patient satisfaction and long-term outcomes beyond the hospital stay [2].

The 1984 change to a prospective payment system based on diagnosis related groups (DRGs) in the US Medicare program resulted in concern that the quality of care could decline due to incentives to reduce length and cost of stay [3]. Similar concerns regarding quality, cost and length of stay were apparent in Canadian settings [4]. These concerns led to research to measure and ensure the quality of care. One of the methods included the development of explicit and implicit criteria for use in chart review [5–7].

In the US, considerable research has also been focused on specific outcomes, beginning with an examination of hospital mortality data and attempts to explain the variation in mor-

Received 22 September 1995; accepted 2 January 1996.
Correspondence: T. A. MacKenzie, Case Mix Research Group, Department of Community Health and Epidemiology, Queen's University, Kingston, Ontario, Canada K7L 3N6. Tel. (613) 545-2548; Fax (613) 545-6686.
tality rates between hospitals [8–14]. There is a growing awareness of the short-comings of classification systems, such as DRGs in the US and case mix groups (CMGs) in Canada, in assessing outcomes. The importance of severity of illness as a method of adjusting hospital mortality data has led to an increasing interest in severity systems such as Acute Physiology and Chronic Health Evaluation (APACHE, APACHE II) [15,16], MedisGroups [17] and the refined DRG grouper, RDRG [18]. Use of these systems to adjust hospital data for severity of illness improved the ability of the models to predict outcomes such as mortality, cost and length of stay [19–28].

Much of the research into quality monitoring is from the US, and the examination of structures, process and outcomes of clinical services in a Canadian setting is still in its infancy [29]. However, efforts are being made to apply the research findings to the Canadian setting [30] as well as to apply the methodology to Canadian data [31–32].

OBJECTIVES

This project was designed to explore the development of an innovative and cost-effective approach to outcomes monitoring in health care facilities. The first objective was to develop a flag which screened individual cases for quality problems, so that charts selected for review would include a relatively high percentage of problem cases, and those not selected, a relatively low percentage of problem cases.

The flag involved the use of severity of illness information to separate cases into two categories, those with a high and those with a low likelihood of a poor outcome. Low-risk cases which subsequently had a poor outcome could then be flagged as potential quality problems. Random selection of charts for review does yield a percentage of cases with quality problems, but there is often a large percentage of cases without problems [7], which can therefore be both costly and frustrating for the reviewer.

The second objective was to use this information to generate risk-adjusted rates of problems, which could be used to standardize mortality rates at the hospitals [8–10,13].

Additional objectives included the assessment of the rate of true problems in the data base, and feedback to the hospitals regarding the resource implications for hospitals choosing to screen charts in this way.

METHOD

The study sample was drawn from 26 participating hospitals, 11 of which were teaching hospitals. Hospitals from all ten Canadian provinces were included in the study, varying in size from 100 beds to over 1000 beds. The sample was proportionally drawn from the participating hospitals so that the findings would be representative of the hospitals.

The condition-specific samples initially consisted of 500 acute myocardial infarction (AMI), 500 pneumonia and 500 hip fracture records. Owing to the time constraints placed on return of reviewed charts, the completion rate for the reviews done by the physicians was 64.4%, which resulted in data bases consisting of 363 AMI, 337 pneumonia and 266 hip fracture records. Since the physicians received their charts in random assignments of 12 at a time and there were no partially reviewed packages, it is unlikely that this resulted in any systematic bias. From within each of these three conditions, a 10% random sub-sample was drawn in order to compare the results of the flags to a 10% random sample approach. The cases were sampled from the fiscal year 1988–1989.

The discharge abstract information, including diagnosis and procedure codes, length of stay, sex, age and discharge disposition, was available for all cases. In addition, information required by the APACHE II severity of illness system, such as temperature, blood pressure and lab values, was retrospectively abstracted from the medical record. The length of time required for this abstraction was monitored in order to evaluate the source implications of this methodology.

APACHE II uses 12 common physiologic indicators to stratify patients by risk of death, and assigns a score of 0–71. Higher scores are indicative of a more severe illness. The abstraction of the information necessary to operationalize the APACHE II severity system was completed by nurses and health records analysts who were trained in the data collection rules for the system. Abstraction was completed at Queen’s University and on-site at four partici-
Use of severity of illness

The abstractors completed 20 validation charts prior to participating in the study, and training was successfully completed when the raw scores of the abstractors achieved 80% agreement. The abstracted findings were used to generate the admission severity of illness score for APACHE II.

Evaluation of the charts by physicians was used as the reference standard to measure flag performance. The reviewing physicians were recruited through the senior administration of participating hospitals, who recommended 110 potential participants. A subset of this group was selected in such a way as to obtain a representative sample nationally. AMI cases were reviewed by eight cardiologists, hip fracture charts were reviewed by four orthopaedic specialists, and pneumonia charts were reviewed by five respirologists. Charts were randomly allocated, although physicians did not review charts from their own hospitals.

All physicians were trained in the implicit review tool developed by the RAND Corporation [7], which asked for an evaluation of the quality of care from a number of process perspectives, as well as an overall judgement as to whether the patient's outcome was better or worse than expected based on his or her condition on admission [7,33]. The review tool and the accompanying guidelines are designed for the assessment of the quality of medical care for a wide variety of medical and surgical diagnoses. Training and validation took place at the RAND Corporation in Santa Monica, California. Additional reliability and validity testing of the physicians was not done, since published statistics on these measures are available [33]. Physicians were not aware of the APACHE II scores at the time of their review. They were aware of whether or not the patients had died, however, which may have resulted in a more negative evaluation of patient care for those AMI patients who had died [34].

The scores assigned by the APACHE II program were initially analysed to determine which scores produced the optimal level of sensitivity and specificity. This was done by developing a program that evaluated all possible severity cutpoints. A subset of these analyses was calculated manually in order to check the reliability of the program. Based on these analyses, the optimal cutpoint for APACHE II was 19. Thus, the combination of an APACHE II score of less than 19 (low severity) with subsequent death, was used as the predictor of a potential quality problem.

Once the optimization was complete, analyses focused on sensitivity, which is the ability of the system to accurately flag true quality problems, and positive predictive value (PPV), which indicates the likelihood of a positive test being associated with a positive result. The combination of high sensitivity and high PPV maximizes efficiency by identifying quality problems with a minimal loss of effectiveness.

The flags were then tested against a 10% random sample approach, drawn from within each of the three data bases. The size of these 10% random samples differed, to match the size of the data bases from which they were drawn. For example, the final AMI data base consisted of 363 cases, and the 10% random sample drawn from within this data base therefore included 36 charts. A 10% level was chosen since this is a baseline that is currently used at many hospitals.

Significance testing was achieved with tests of differences in proportions (Z-tests) between each flag and the 10% random sample. All analyses were completed using the Statistical Analysis System [35], UNIX Version 6.07 on a Sparc1® SUN workstation.

RESULTS

Table 1 outlines the characteristics of the sample for each of the data bases. The mortality rate was the highest for the AMI patients, while

<table>
<thead>
<tr>
<th>Condition</th>
<th>Sample size</th>
<th>Mortality rate</th>
<th>Length of stay</th>
<th>Mean APACHE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acute myocardial infarction</td>
<td>363</td>
<td>15.4%</td>
<td>12.3 (22.4)</td>
<td>10.5 (5.5)</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>337</td>
<td>11.0%</td>
<td>10.9 (23.4)</td>
<td>11.5 (6.3)</td>
</tr>
<tr>
<td>Hip fracture</td>
<td>226</td>
<td>6.0%</td>
<td>21.7 (23.3)</td>
<td>8.3 (3.7)</td>
</tr>
</tbody>
</table>

*Standard deviations provided in parentheses.
TABLE 2. Rate of true problems in the data base and sensitivity and positive predictive values of low severity/death flag by condition

<table>
<thead>
<tr>
<th>Condition</th>
<th>Rate of true problems</th>
<th>Sensitivity</th>
<th>Positive predictive value</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMI</td>
<td>28 (7.7%)</td>
<td>64%*</td>
<td>53%*</td>
</tr>
<tr>
<td>Pneumonia</td>
<td>16 (4.7%)</td>
<td>44%*</td>
<td>29%*</td>
</tr>
<tr>
<td>Hip fracture</td>
<td>29 (10.9%)</td>
<td>31%*</td>
<td>56%*</td>
</tr>
</tbody>
</table>

*Indicates that the flag sensitivity or PPV is significantly greater than a 10% random sample approach at a level of $p < 0.05$.

TABLE 3. Comparison of flagging rates for AMI: APACHE II low severity/death versus random sample

<table>
<thead>
<tr>
<th></th>
<th>APACHE &lt; 19 + death</th>
<th>10% random sample</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cases flagged as non-problems</td>
<td>90.6%</td>
<td>90.0%</td>
</tr>
<tr>
<td>Cases flagged as problems</td>
<td>9.4%</td>
<td>10.0%</td>
</tr>
<tr>
<td>Physician review of cases flagged as non-problems</td>
<td>97.0%</td>
<td>93.0%</td>
</tr>
<tr>
<td>True negatives</td>
<td>97.0%</td>
<td>93.0%</td>
</tr>
<tr>
<td>False negatives (problems missed)</td>
<td>3.0%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Physician review of cases flagged as problems</td>
<td>52.0%</td>
<td>8.0%</td>
</tr>
<tr>
<td>True positives (yield)</td>
<td>48.0%</td>
<td>92.0%</td>
</tr>
<tr>
<td>False positives (excessive reviews)</td>
<td>48.0%</td>
<td>92.0%</td>
</tr>
</tbody>
</table>

hip fracture patients had the highest average length of stay. The mean APACHE II severity score was the highest for the pneumonia patients.

Table 2 outlines the sensitivity and PPV as compared to the 10% random sub-sample drawn from within each of the four data bases. The low admission severity and inpatient death using the APACHE II system had a significantly greater sensitivity and PPV than a 10% random sample approach in all condition-specific analyses.

In addition, Table 2 outlines the rate of actual quality problems in the data base, as identified by physician review. The physician definition of a chart with a quality problem was based on an outcome worse than expected, given the patient’s chronic and acute condition at admission, and was measured on a scale of 1-5. The lowest rate of problems was identified in the pneumonia data base, in which physicians identified 16 cases (4.7%) with quality problems. The rate was the highest for hip fracture charts, with 29 (10.9%) problem cases. For AMI, physicians identified 28 charts (7.7%) with quality problems.

Table 3 demonstrates the effectiveness of this approach for AMI by comparing the APACHE II low severity of illness plus death flag with the results of a 10% random sample screening approach. The yield of problem cases is comparable, at 9.4% and 10% respectively, but there is a substantial difference in the outcome of the review completed by the physicians. The reviewing physician agreed that there was a quality problem in 52% of the APACHE II plus death flagged cases, as compared to finding quality problems in 8% of the cases when random sampling was used. In other words, using quality monitoring flags as screening tools could mean that one in two charts going to review may have a problem, while less than one in ten may have a problem using a random sampling approach. The APACHE II low severity/death flag also missed fewer problems, with a false negative rate of 3% as compared to 7% for the 10% random sample approach.

It took a mean of 8.6 minutes to abstract the information required by APACHE II from the chart, with a standard deviation of 2.0 minutes.

DISCUSSION

The results of this project demonstrate that outcomes monitoring flags based on low admission severity of illness combined with death could play an important role in screening cases for potential quality problems.

The APACHE II combined with inpatient death flag was significantly better than random
Use of severity of illness

selection across all conditions. Generally, the flag was most effective for AMI, although the relative values for sensitivity and PPV varied across the three conditions. Thus, the relative importance assigned to sensitivity (finding positive cases) and PPV (minimizing number of negative cases) will determine the value of the flag for each condition.

There is a number of resource implications inherent in utilizing such a quality monitoring system. First there is the cost of abstracting the additional information to operationalize a severity of illness system. However, the average of 8.6 minutes it takes to complete an APACHE II review uses fewer resources than a physician review, which normally takes closer to 30 minutes and is considerably more expensive. There is also the necessity of educating and training hospital staff in the use of the quality flags. In addition, flagged cases require review to determine whether a quality problem exists. However, this is an ongoing process which can be integrated with other continuous quality improvement (COI) activities. Also, this is designed to replace rather than add to portions of the current audit process.

One limitation of these findings relates to differences in sensitivity and PPV across the three conditions. These differences mean that the value of this system will have to be tested in other conditions before its generalizability across case mix can be determined. A related limitation is associated with the establishment of the cutpoint for APACHE II. We used the cutpoint that maximized the true positive rate within our sample. We did not use a split-sample approach to validate the cutpoint, as that was not our focus. This means that in a different sample, it would be necessary to establish the applicability of this cutpoint or establish one more appropriate to the current sample.

Acknowledgements: The authors would like to acknowledge the guidance and financial support of the Canadian Council for Health Services Accreditation.

REFERENCES


Downloaded from https://academic.oup.com/intqhc/article-abstract/8/2/125/1832368 by guest on 17 January 2019


