

Estimation of Variance (2009 Data)

The Hospital Cost Transparency (HCT) project annually reports inpatient payments for APR-DRGs that, during the previous calendar year, had at least 150 observations or at least \$1 million in allowed payments and a minimum of 25 observations. Furthermore, the original governing workgroup decided the variance should be estimated after stratifying the data by APR-DRG and severity of illness category (minor, moderate, major, and extreme). This means separately estimating the variance in each cell of a 4xR matrix, with R equal to the number of eligible APR-DRGs and 4 referring to the severity of illness categories.

The difficulty with this approach is that it was known in advance that a substantial number of cells would not have a sufficient number of observations to generate reliable variance estimates. Since the first year's data were ultimately published after combining the severity of illness into two categories (minor/moderate and major/extreme), the 2009 estimated variances used the same combined severity of illness categories.

In the event of small cell sizes, contingencies provided that the analyst should attempt to estimate the variance as follows (in order):

1. Use the pooled variance from the 2009 data if Levene's test fails to reject the null hypothesis that the variance does not differ by severity of illness category
2. The 2009 data supplemented with all observations from the 2008 HCT data set (adjusted for inflation to 2009 dollars)
3. Use the pooled variance from the combined 2008-2009 data if Levene's test fails to reject the null hypothesis that the variance does not differ by severity of illness category
4. The combined 2008-2009 data set supplemented with all observations from the 2007 HCT data set (adjusted for inflation to 2009 dollars)
5. Use the pooled variance from the combined 2007-2009 data if Levene's test fails to reject the null hypothesis that the variance does not differ by severity of illness category
6. The combined 2007-2009 data set supplemented with all observations from the 2006 HCT data set (adjusted for inflation to 2009 dollars)
7. Use the pooled variance from the combined 2006-2009 data if Levene's test fails to reject the null hypothesis that the variance does not differ by severity of illness category
8. The combined 2006-2009 data set supplemented with all observations from the 2005 HCT data set (adjusted for inflation to 2009 dollars)
9. Use the pooled variance from the combined 2005-2009 data if Levene's test fails to reject the null hypothesis that the variance does not differ by severity of illness category
10. Use the variance estimate from 2008 (these methods are described in a separate document)

Preliminary Data Preparation

The 2009 HCT aggregate claims data were first assessed for their suitability for risk-adjustment with the 3M Core Grouping Software. This software assigns a severity of

illness category based on diagnoses, procedures, length of stay, patient age, and patient discharge disposition. The risk-adjustment software generally will not provide a severity of illness score if either age or discharge disposition are missing, or if the principal diagnosis is either missing or invalid.

The risk-adjustment can also be influenced by secondary diagnoses and procedures. The standard inpatient discharge record contains nine diagnoses and six procedures. Internal testing revealed that deleting several diagnoses and procedures could alter the assigned severity of illness. Some claims systems do not capture all of the diagnoses and procedures in the discharge record, making risk-adjustment of these records potentially unreliable.

Accordingly, records in the aggregate claims data were required to meet the following risk-adjustment standards:

- Age is not missing
 - This requires the date of birth field to be populated with a valid date
- Length of stay is not missing
 - This requires the admit date and discharge date fields to be populated with valid dates
- Principal diagnosis is not missing
- Claims system captures at least five diagnoses
- Claims system captures at least four procedures

Additional validations were performed to assure the quality of the data and eliminate duplicate records. Details about this extensive process are available in a separate document. The final 2009 HCT aggregate claims file contained 45,431 risk-adjusted and validated records.

Records that did not meet the risk-adjustment standards and records that failed the additional validations were dropped. A separate data file of dropped records was maintained to assess selection bias (see Table 1).

Table 1: Characteristics of Submitted Claims Data

	Risk-adjusted records	2009 dropped records	2008 dropped records
Percent routine discharge	91.2%	56.4%	85.6%
Percent female	62.1%	52.3%	58.7%
Median age	37	45	29
Mean length of stay	3.5	3.4	6.2
Mean allowed payments	\$14,594	\$16,221	\$12,884

In previous years the differences between risk-adjusted records and dropped records were modest, although not negligible. The patients from dropped records tended to be a few years younger, but had longer hospital stays and were less likely to be routinely discharged. Data from calendar year 2009 had several noteworthy problems:

- Over 1000 records had missing discharge status. This is reflected in relatively small proportion of dropped records with routine discharge.

- Exactly 250 records had missing gender. The impact of this appears relatively modest and the smaller proportion of dropped records with female gender in calendar year 2009 is not entirely explained by the missing values, even if all of the missing values are recoded to female.
- Exactly 34 records had incorrect date of birth or incorrect discharge date, resulting in negative values being populated for age and length of stay. The median age in these 34 records was -7989 years, so these records were excluded from age and length of stay comparisons. The median age was substantially higher for dropped records in calendar year 2009 than for dropped records in calendar year 2008. In addition, the median length of stay was also substantially higher for dropped records in calendar year 2009 than for dropped records in calendar year 2008. These differences are not fully explained by excluding the 34 records with negative values for age and length of stay.

The net result of this is patients from dropped records may be somewhat higher risk. The mean allowed payments for these patients were modestly different than for risk-adjusted records, but were substantially higher than dropped records from 2008. Selecting records for risk adjustment could introduce a small amount of bias favoring patients with slightly less risk, although this does not appear to result in substantial differences in allowed payments.

Inclusion Criteria

Inclusion criteria were then applied to 2009 HCT aggregate claims data:

- Inpatient claims
 - Excluding Med-Advantage claims
 - Excluding coordination of benefit claims
 - Excluding denied claims
 - Excluding workers compensation claims
- Patient resided in Oregon
- Patient was treated as an inpatient at an acute-care hospital located in Oregon
 - Excluding VA facilities, Shriners, and psychiatric hospitals
 - Excluding patients who discharged against medical advice
 - Excluding patients who expired

Estimating Variance

The 2009 HCT aggregate claims data were then stratified by APR-DRG and two combined severity of illness categories (minor/moderate and major/extreme) in order to generate cell sizes. A variance estimation flag was created to identify cells with at least 30 observations and cells where the variance was estimated as the pooled variance. This was repeated with the supplemented data sets: combined 2008-2009 HCT data, combined 2007-2009 HCT data, combined 2006-2009 HCT data, and combined 2005-2009 HCT data. The variance estimation flags were merged into the 2009 HCT aggregate claims data, the flags indicating the data set from which the variance was estimated. The variance estimate from 2008 was used if cells had fewer than 30 observations in the combined 2005-2009 HCT data and if the pooled variance could not be used (Levene's test rejected the null hypothesis that the variance did not differ by severity of illness category).

With the 2009 HCT data stratified by APR-DRG and two combined severity of illness categories, the standard deviation (SD) of the allowed payments was calculated for each cell with at least 30 observations. The data were aggregated by APR-DRG to calculate the grouped SD of allowed payments. The SD was similarly calculated in the three supplemented data sets. The SD variables from the three supplemented data sets were merged into the 2009 HCT aggregate claims file. Each record in the 2009 HCT data was assigned a SD variable based on the value in the variance estimation flag. Outliers were identified as observations with allowed payments outside +/- 4 SD from the mean. Outliers were removed from the 2009 HCT data and transferred to a data file of excluded records.

The 2009 HCT aggregate claims file was then aggregated by APR-DRG in order to determine the number of observations after removing outliers. Records were again flagged for inclusion based on frequency (APR-DRG had at least 150 observations) or cost (APR-DRG had at least \$1 million in allowed payments and at least 25 observations). Records not flagged were removed and transferred to the data file of excluded records. The data were then stratified by APR-DRG and two combined severity of illness categories (minor/moderate and major/extreme). The allowed payments were summed for each cell and the mean allowed payments, median allowed payments, mean length of stay, and the number of observations were calculated. The results were separately tabulated based on frequency and based on allowed payments.

The data were then further stratified by hospital. Again, the allowed payments were summed for each cell and the mean allowed payments, median allowed payments, and number of observations were calculated. The results were tabulated by APR-DRG for two combined severity of illness categories (minor/moderate and major/extreme) and each hospital. Cells with fewer than two observations were reported as "0 or 1."

In order to assess selection bias due to applying the exclusion criteria, several variables (age, gender, length of stay, allowed payments, and discharge disposition) were compared between included and excluded records (note that dropped records are not part of this analysis). In addition, these variables were compared after filtering out records involving pregnancy and childbirth, since these records represent a substantial proportion of the data set and will tend to focus on narrow ranges of age, length of stay, and allowed payments.

Data were analyzed with SAS version 9.2, 3M[®] Core Grouping Software version 2010.2.1, and Microsoft Office[®] Excel 2003.

Results

Originally 132 unique APR-DRGs were identified for further analysis based on either frequency or cost. The resulting 2 x 132 matrix contained 264 cells, although 7 cells were null after generating cell sizes. Of the remaining 257 cells, the variance was directly estimated from the 2009 HCT aggregate claims data for 177 cells (68.9%) and was estimated as the pooled variance for 66 cells (25.7%). The variance was directly estimated from the combined 2008-2009 HCT data for 6 cells (2.3%), was directly estimated from the combined 2007-2009 HCT data for 3 cells (1.2%), and was directly

estimated from the combined 2005-2009 HCT data for 5 cells (1.9%). The pooled variance was not used with data from prior years as, for cells where the variance was not estimated from the 2009 HCT data, Levene's test rejected the null hypothesis with all of the combined HCT data sets. In addition, none of the variance estimates were generated from the combined 2006-2009 HCT data since this data set did not contribute a sufficient number of observations.

After removing outliers, 8 APR-DRGs no longer had at least 25 observations and at least \$1 million in allowed payments. The resulting 2x124 matrix had 248 cells, although once again 7 cells were null after generating cell sizes. The final data file included only the remaining 241 populated cells.

A total of 56 unique APR-DRGs had at least 150 observations. A total of 121 unique APR-DRGs had at least \$1 million in allowed payments and at least 25 observations. This represents a total of 38,091 observations and over \$518 million in allowed payments; 7340 records were excluded from further analysis. Diagnoses and procedures from six major clinical domains (obstetrics/gynecology, childbirth, digestive system, respiratory system, cardiovascular system, and musculoskeletal system) accounted for 82 unique APR-DRGs, over 83% of the observations (31,742) and about 81% of the total allowed payments (\$420 million).

APR-DRG 560 (vaginal delivery) had the highest total allowed payments (\$35,946,166) and APR-DRG 640 (normal newborn) ranks had the largest number of observations (6474). Diagnoses and procedures involving the digestive system accounted for 23 of the 124 APR-DRGs while diagnoses and procedures involving the cardiovascular system accounted for an additional 17 APR-DRGs. Diagnoses and procedures in the obstetrics/gynecology and childbirth domains accounted for 17,625 observations, or approximately 46% of the included observations.

After filtering out records involving pregnancy and childbirth, differences were generally modest when comparing included records to excluded records in the risk-adjusted aggregate claims file (see Table 2). Included records had a higher percentage of patients routinely discharged and a lower percentage with major/extreme severity of illness, indicating perhaps a slight bias toward lower risk patients. This did not result in substantially lower mean allowed payments; in fact, the mean allowed payments were higher.

Table 2: Characteristics of Risk-Adjusted Claims Data

	Included records	Excluded records
Percent routine discharge	87.5%	82.4%
Percent female	48.9%	51.4%
Median age	54	48
Mean length of stay (days)	3.5	4.6
Minor/moderate severity of illness	82.3%	77.6%
Mean allowed payments	\$19,251	\$18,008

Overall this indicates that, after filtering out records involving pregnancy and childbirth, the included APR-DRGs do not produce an egregiously biased subset of records from the risk-adjusted aggregate claims data. The magnitudes of the differences in median age, percent female, and mean allowed payments are substantially different if including pregnancy and childbirth records. It should be noted that relatively few pregnancy and childbirth records end up being excluded from the risk-adjusted aggregate claims data, so the risk of selection bias is minimal for pregnancy and childbirth records. However these records, since they are a very large subset of the aggregate claims data, can affect the assessment of selection bias in other records if not filtered out.