MISSING DATA, OUTLIER IDENTIFICATION AND HANDLING IN MEDICARE ESRD PUBLIC USE FILES

UPDATED AUGUST 2013

PRIMA Health Analytics

49 BALD EAGLE ROAD
SOUTH WEYMOUTH, MA 02190 U.S.A.
JMARKSTEPHENS@GMAIL.COM
TEL: 1.781.588.0248

Dialysis provider data made available through public use files by Medicare is notoriously "dirty", containing transcription errors, missing data, and inconsistencies. Analysis using these data is rife with uncertainty and prone to poor conclusions unless the data are edited and "cleaned" prior to use in business decision support.

Missing Data, Inconsistent Data and Outliers

Problem data can be classified into three broad categories. Each category calls for a different approach to identification and handling. There is often a logical hierarchy to identification and handling of bad data based on the category.

Missing Data. Missing data is straightforward to identify. It is more difficult to determine whether missing data is bad data or not. For this, one needs to know where a value is required, versus where one is optional.

Inconsistent Data. Data may be internally inconsistent or externally inconsistent. Internal inconsistency occurs when two or more variables from the same source are in conflict. For example, a report shows total annual treatments of 5000, but a patient census of only 10 patients. An average of 500 treatments per patient is not possible, so one or the other (or both) of the values must be incorrect. (This would also be an example of an "impossible" value, which is a subtype of inconsistent data. Impossible values can often be identified without reference to other data. For example, a negative number of patients or treatments would be an impossible value.) External inconsistencies appear when data are joined from multiple sources and compared. For example, one source shows 10 patients and the other source shows 50 patients. Or one source shows 15 incident patients and the second source shows 10 total patients. Cross-validation should be used to find inconsistent data between like fields in the Cost Reports, Dialysis Facility Reports and Dialysis Facility Compare. Thresholds can be set based on % variance between values. Other types of inconsistent data identification can include analysis methods based on business knowledge. For example, EPO rebates are supposed to be reported as a negative number in the cost reports, but a rather large percentage of facilities report then as a positive number. We could therefore globally define all positive values for EPO rebates as errors and handle them by switching the sign (multiply by -1).

Outliers. Outliers are those data values that are deemed to be "out-of-range"; that is, when compared to other values for the same measure, they are considered implausible. (For variables with defined ranges, out of range values would be classified as impossible values, versus outliers.) Outlier values may be the hardest category of bad data to identify. There are many techniques that can be used to define outliers, both

parametric techniques (which pre-suppose the distributional characteristics of the data) and non-parametric techniques (which do not pre-suppose an underlying distribution.)

Outlier Identification Techniques

Identifying outliers is a science in and unto itself. There is no such thing as a simple test. However, there are many ways to look at a distribution of numerical values to see if certain points seem out of line with the majority of the data. Expert knowledge of what values data can have is probably the best solution.

1. **Statistical Techniques for Normally Distributed Data.** For normally distributed data, we know that 99.8% of the data should be within +/- 3 standard deviations from the mean. Anything beyond 3 standard deviations could be considered an outlier. This method is seldom appropriate for health care statistics, however.

2. **Techniques for "Normal-like" Data.** More often, health care statistics have distributions which have some normal-like characteristics, but may be skewed or look like the right side of a normal distribution, bounded on the lower end by zero. Outlier identification methods that can be considered for these types of data are:

    a. **Outer Fence Method.** Used by CMS for defining outliers for average costs in the development of the PPS Composite Rate[1]. The definition of the outlier "fence" was defined as the 75th percentile plus three times the interquartile range (IQR, which is the 75th percentile minus the 25th percentile); while the lower outer fence is the 25th percentile minus three times the IQR. This is a very liberal definition and will identify very few outliers. It may make sense to apply only an upper limit in some cases as lower limits of zero are often valid.

    b. Median Absolute Deviation. In this method, data values are compared to the median (50^{th} percentile) of all values. Absolute deviations from the median are calculated and then the median of the deviations themselves is used as the benchmark against which individual data value deviations are compared. The test statistic is calculated as the absolute deviation divided by the median of all absolute deviations, and is thus a ratio of the individual deviation to the average deviation. A threshold for outlier status is set based on the distribution of test statistic scores.

3. Distance-based Approaches. The concept of distance-based outliers relies on the notion of the neighborhood of a point, typically, the k-nearest neighbors. This method avoids the excessive computation that can be associated with fitting the observed distribution into some standard distribution and in selecting discordancy tests. Distance based methods suffer from detecting local outliers in a data set with diverse densities[2]. This is an approach that has practical value in this study, where, for instance, we might set the outlier thresholds at the 1st and the 99th percentiles of the data distribution, thereby defining 2% of the data values as outliers. Such an approach should not be arbitrary, however, and needs to be informed by the data under question.

4. Density-based Approaches. The density-based approach estimates the density distribution of the data and identifies outliers as those lying in low density regions. Density-based techniques have the advantage that they can detect outliers that would be missed by techniques with a single, global criterion. Parameter selection for upper bound and lower bound can be difficult[2].

Handling Techniques

Once we have defined our approach to identifying "bad data", a more critical decision is what to do about it. One needs to understand how the data will be used in order to decide what approach(es) should be taken to handling inconsistent data, outliers and missing values. In statistical analysis, like regression analysis, data transformations and multiple imputation techniques are commonly used to accommodate outliers and fill in the holes where there are missing data. For descriptive reporting, e.g. – means, standard deviations, sums, etc, this is not possible or appropriate. Case exclusion is another method often used in statistical analysis, where there are few outliers, but probably will not be desirable in this project because of the high rate of bad data. Accommodation should not be ruled out. That is, it may make sense to identify outliers but not "fix" them. One approach would be to flag outlier records in a metadata file.

So a major handling decision will be, do we change bad data, or set it to missing, or accommodate it? If we change a value do we store the original somewhere? Handling methods to be considered can include:

- Accommodation.
- Case exclusion.
- Set to Missing: Set inconsistent and outlier values to missing.
- Global replacement: Set inconsistent/outlier values to zero, mean, median, etc.
- Impute Values via Prediction (based on other similar data)
- Impute Values via Interpolation (e.g. based on ratios relative to # of patients, etc.)
- A Combination of the above.

Bivariate or multi-variate outliers (e.g. – cost per treatment, nurse-to-patient ratio) will require a step-wise handling approach because we may not know, *a priori*, which of the source variables are the culprits, as none of them may have been flagged as outliers in the univariate outlier analysis. A stepwise approach to handling bivariate or multi-variate outliers could be:

1. Cross-validation of one or more variables to find the "culprit"
2. Fix the culprit. Recompute.
3. Still an outlier? Suspect the other variable(s).
4. If underlying variables cannot be "fixed", consider replacement with a value based on a measure of central tendency (e.g. – mean, median, mode.)

References

[1] See UM-KECC Report Chapter VII Section C:
http://www.sph.umich.edu/kecc/assets/documents/UM-KECC%20ESRD%20Bundle%20Report.pdf

[2] From: V. Ilango. *A Five Step Procedure for Outlier Analysis in Data Mining*.
http://www.europeanjournalofscientificresearch.com/ISSUES/EJSR_75_3_02.pdf