
For more information on FAIR Health data products, please visit fairhealth.org.

Application of the MAD (Median Absolute Deviation) Methodology to Exclude Extreme Data Values in FAIR Health Products

Outliers are typically defined as the values in a distribution of data that are extreme and likely to be erroneous. FAIR Health employs a methodology to help identify and exclude such data from its products. This methodology is based upon a median absolute deviation (MAD) technique, which was recommended by the Upstate Health Research Network (UHRN), a consortium of academic experts at universities throughout the United States, and was approved by the FAIR Health Board of Directors.

The MAD algorithm, a recognized statistical methodology, is a robust analog to a more commonly used outlier technique, which uses standard deviation from the mean¹. MAD uses deviation from the median, which is less susceptible to distortion caused by outlying values.

The FAIR Health MAD technique starts with the calculation of the median for a distribution of charges for a particular procedure code in a specific geozip. The median is determined by arraying the data for a given procedure code/geozip combination from lowest to highest and identifying the value at which half of the data are higher and half are lower². Once the median value has been identified, the median absolute deviation or MAD is calculated.

MAD is determined by calculating the deviation from the median as follows:

- Subtracting each charge in the distribution from the median charge to determine each respective deviation from the median;
- The absolute values of the deviations from the median are arrayed in order from lowest to highest and the median of the absolute deviations is determined;
- The median of the absolute deviations (from the median) is multiplied by the constant of 1.48³;
- This product is defined as the MAD.

Charge thresholds are then established using the median value for each geozip/procedure code combination and its corresponding MAD. FAIR Health sets its charge thresholds using a relatively conservative formula of + 4 times the Median Absolute Deviation as calculated by the MAD methodology explained above (median charge + 4x MAD). Typically, charges that are more than ± 4 MAD from the median are excluded from FAIR Health data products.

FAIR Health statisticians and data analysts are continually reviewing the processes by which FAIR Health validates and arrays data in its products. While FAIR Health currently uses the aforementioned MAD algorithm to identify and exclude outliers, this process may be modified, augmented or amended as deemed necessary. Additionally, FAIR Health utilizes a number of business rules that are used to exclude various other types of invalid data. These business rules can sometimes impact the calculation of outliers. Thus, it is important to note that the MAD technique is only one step in a larger set of data quality procedures that FAIR Health uses to identify and exclude data points that are likely to be erroneous.

1. Tukey, John W., (1977), *Exploratory Data Analysis*, Addison-Wesley.
2. Or the mean of the middle two values, if there are an even number of data points.
3. 1.48 is a constant developed by statistical theoreticians that when multiplied by the median of the absolute deviations, produces a result that is asymptotically equivalent to a standard deviation method as applied to normal distributions. For more information on the statistical theory behind MAD, please see the following references:

- Ruppert, David, (2010), *Statistics and Data Analysis for Financial Engineering*, Springer.
- Iglewicz and Hoaglin (1993), *How to Detect and Handle Outliers*, ASCQ Quality Press.
- Barnett and Lewis, (1994), *Outliers in Statistical Data*, Wiley.
- Peter J. Rousseeuw and Christophe Croux (1993), *Alternative to the Median Absolute Deviation*, *Journal of the American Statistical Association*. Vol. 88. No. 424.