Statistical Modeling and Outliers in a Medical Registry:
Online UI Input Feedback and Offline Candidates for Review

Ronald P. Loui
Department of Computer Science, University of Illinois Springfield, IL

***Abstract*—This is a report on a few lessons learned while doing outlier detection on a large clinical outcomes registry. The most important lesson is that the statistical modeling is most useful for real-time feedback in the GUI at the data ingestion stage. That is when the quantification of the deviation is most useful, and it is when the error can be fixed. Practical experience reports such as this should help data entry pipelines use EMR data analysis to improve data quality.**

I. Introduction

It is an old and classic idea to find deviant values in a database as candidates for errors [1,2,3]. Not all errors will appear as statistical outliers (and not all outliers are errors). If the goal is to audit all of the data for correctness, finding the easy cases does not necessarily bring the audit closer to its goal of comprehensive review. Still, it is very useful, removing the obvious noise so other irregularities can receive more attention. Automated error detection has become prevalent whether based on statistics, grammar, syntax, or just spelling. When humans are busy, automation greatly improves data quality review.

Based on an effort to model elements in a clinical reporting registry, the main conclusion is that *statistical modeling is most useful for real-time feedback and flagging of outlier values during user input*. By the time the error is in the database, it may be too late to correct it.

Data elements included mundane quantities such as weight and smoking pack years, not just lab values such as BUN, BNP, and HDL, observations such as Proximal LAD and LVEF, and procedure values such as intubation time. Area code and zip code also succumbed to outlier detection because typographical errors that drop a digit or add a digit can produce statistical anomalies. Deviant values 20+ sigma and higher occurred.

II. Fitting Gamma Distributions

We fit gamma distributions to each data element with a maximum likelihood estimate for θ and a gradient fit for k. Gammas have wide variation: normal, semi-normal, exponential, and semi-exponential, all of which were observed in the data. Bimodal and multi-modal distributions should be assessed specially.

III. First Lessons

Quantified deviation is unnecessary for finding errors post hoc (because an ordered list suffices for human attention). Outliers should be removed before fit (because they skew). Ratios should also be modeled (because they often reveal anomalies more clearly). One height/weight ratio was 35 sigmas above the mean because digits had been dropped, but the weight itself was not extreme (because infants also occurred in the database).

IV. Real-Time Feedback at Input

Quantified deviation is most useful when deciding to alarm the user during input of values. The alarm could be determined based on percentile, but number of standard deviations seems more intuitive. Percentile reports make use of the fit, but do not carry the same visual punch: .0001 is not so alarming vs. .001, but most can learn quickly that +10 sigma is noteworthy.

Web forms give red-letter warnings for some errors, such as min, max, or missing input. The inclusion of outlier warnings during input is a modest but natural step further, either on the client side, with a summary statistic and distribution parameters, or with a round-trip to the server for rank statistic (nth highest).

V. Conclusion

These ideas are consequences of newly available data constantly subject to data analytics. Data cleaning has appeared before in medical applications [4,5,6], but using the fit to inform data entry is a new user-interface benefit (see also [7]). Outlier feedback should be implemented as a best practice in EMRs.

Bibliography

[1] Hodge, V. J., and J. Austin. "A survey of outlier detection methodologies," *Artificial Intelligence Review 22*, 2004.

[2] Miller, R. C. and B. A. Myers. "Outlier finding: focusing user attention on possible errors," *Symp. on User Interface Software and Tech*, 2001.

[3] Hellerstein, J. M. "Quantitative data cleaning for large databases," *United Nations Economic Commission for Europe* (UNECE), 2008.

[4] Laurikkala, J., et al. "Informal identification of outliers in medical data." In *Intelligent Data Analysis in Medicine and Pharmacology*, 2000.

[5] Roberts, S. J. "Extreme value statistics for novelty detection in biomedical data processing." *IEE Proc. Science, Measurement and Technology 147*, 2000.

[6] Van den Broeck, J., et al. "Data cleaning: detecting, diagnosing, and editing data abnormalities." *PLoS Med 2*, 2005.

[7] Hauskrecht, M., et al. "Outlier detection for patient monitoring and alerting." *Journal of Biomedical Informatics 46*, 2013.