RESEARCH ARTICLE

Utilization of Hospital Laboratory Data for Establishing Normal Reference Interval of Quantitative Medical Parameters: Double Filtration Method

Abhaya Indrayan[1], Mohini Bhargava[2], Shubham Shukla[3]

Abstract

Background: India and other developing countries need their own reference intervals for various medical parameters because these populations differ from Western population for the genetic profile, anatomical structure, dietary habits, and lifestyle. Such reference intervals have not been worked out so far for most parameters because it is difficult to have a large database on healthy people in these countries.

Aims and objectives: Large hospitals generally have a huge database of laboratory values but a substantial proportion of them belong to sick subjects whose values cannot be included for establishing normal reference intervals. Thus, the database remains unutilized. We propose a simple method to utilize these data for establishing reference intervals.

Materials and methods: A simple double filtration method is used to exclude all outliers and abnormal values that could finally provide uncontaminated data on healthy values. This method is based on quartiles and interquartile range. The method is illustrated on a dataset from the laboratory of a large tertiary care hospital.

Results: The filtered values have been seen to follow a smooth distribution pattern and can be used to establish our reference intervals using the usual 2.5th and 97.5th percentiles. The method is illustrated for A/G ratio in the data from our hospital, and the reference interval obtained.

Conclusion: Double filtration method can be used on hospital laboratory data to establish reference intervals of medical parameters.

Keywords: Double filtration method, Laboratory data, Normal values, Reference interval.

Indian Journal of Medical Biochemistry (2020): 10.5005/jp-journals-10054-0130

Introduction

Normal reference interval of quantitative medical parameters is an integral part of medical practice. Values found in individual subjects are assessed against such references to establish the diagnosis, to calibrate the treatment, to assess the prognosis, and to monitor the progress of the disease. However, such references are rarely worked out for our population in India and many other developing countries. The only large scale study we could locate for quantitative medical parameters in this region is by Sairam et al.[1] In the absence of local normals, we have to make-do with Western values. Our values could be different because of our markedly different genetic profile, specific body structure, typical nutrition, and a distinct lifestyle. Using Western reference intervals on our population may be causing unknown errors of misdiagnosis and missed diagnosis. These errors can be minimized by preparing and using our own reference intervals. Time has come for us to stand on own feet and not depend on Western values, and avoid all errors, howsoever minor.

Establishing normal reference interval requires measurement of a large number of healthy subjects, called the reference population. Although it is difficult to define a healthy subject exactly, the basic problem is that the healthy subjects rarely submit themselves to such investigations because these investigations are generally ordered in our setup when complaints occur. However, most large hospitals in India carry out millions of investigations and unwittingly prepare a large database. These hospitals generally cater to the well-to-do sections of the society—their values are even more suitable for establishing reference as they are expected to be healthier. Most of these databases on laboratory investigations do not record the health condition of the person—thus are considered unsuitable to workout reference intervals. We present an ingenious but simple method to filter healthy values from among these and propose a procedure to obtain an uncontaminated set of values for working out reference intervals. This would allow fruitful utilization of massive data lying unused in laboratories of large hospitals. The method is illustrated on A/G ratio values recorded in our hospital.

[1,3]Department of Clinical Research, Max Healthcare Institute, New Delhi, India

[2]Department of Laboratory Medicine, Max Healthcare Institute, New Delhi, India

Corresponding Author: Abhaya Indrayan, Department of Clinical Research, Max Healthcare Institute, New Delhi, India, Phone: +91 9810315030, e-mail: a.indrayan@gmail.com

How to cite this article: Indrayan A, Bhargava M, Shukla S. Utilization of Hospital Laboratory Data for Establishing Normal Reference Interval of Quantitative Medical Parameters: Double Filtration Method. Indian J Med Biochem 2020;24(1):9–11.

Source of support: Nil

Conflict of interest: None

Materials and Methods

Lang et al.[2] have described four approaches to define reference intervals. These are (1) mean ± 2 SD range of values seen in healthy subjects when these have Gaussian (normal) distribution; (2) based on median and percentiles that account for skewed distribution; (3) based on outcome or prognosis; and (4) based on expert opinion. The first two methods have the limitation of leaving out

some (generally 5%) extreme values despite such values seen in absolutely healthy subjects.[3] The third approach throws formidable challenge of defining risk in the absence of cutoffs, and requires a long-term follow-up of healthy subjects with different baseline values to observe the thresholds at which the subjects begin to develop prognostically risky outcomes. Such a threshold in all likelihood would vary from person to person, and imputing judgment may be necessary. The last method has been used for recommending cutoffs, such as 130/85 mm Hg for blood pressure (BP) and 126 mm Hg for fasting plasma glucose (FPG) level. However, such examples of expert opinion-based thresholds are rare and it is an uphill task for medical experts to come to a consensus for most measurements. These cutoffs are not without problem either. For example, a person with BP 150/90 may never develop any hypertension-related problem in his or her entire life, and a person with FPG level of 140 mg/dL may never have diabetes-related issues. Thus, expert-based cutoffs too are not infallible. At best, these are the "safe" levels and not the reference intervals. Percentile-based interval is the most preferred as explained later.

Double Filtration Method

Laboratory data from a hospital will necessarily have a large number of values belonging to sick people. Those need to be filtered out as explained in this section. They may also be repeated investigations of the same person. Thus, the first step is to use the patient ID to delete the values obtained in repeat investigations so that such values do not contaminate our results.

Any method based on mean and standard deviation (SD) cannot be used to filter out outliers and abnormal values because these two statistical measures are greatly affected by extreme values. For example, if gamma-glutamyl-transferase (GGT) values of the five persons are (in U/L) 52, 8, 36, 983, and 21, the mean is 220 U/L, which is clearly not representative. This is an acknowledged limitation of the mean.[4] Since SD uses this mean, the value of SD is also distorted by extreme values. Outliers are very likely in hospital data because these data include values of sick people, sometimes seriously ill patients, whose values could be extremely high or extremely low. Thus, we should use quartiles for filtration. This statistical measurement is not much affected by extreme values.[5] We use double filtration using quartiles as follows to eliminate nearly all abnormal values.

- An established statistical method to identify outliers is to calculate interquartile range (IQR = $Q_3 - Q_1$, where Q_1 is the first quartile and Q_3 is the third quartile) and to consider values less than $Q_1 - 1.5 \times IQR$ (negative value is to be considered zero) and values more than $Q_3 + 1.5 \times IQR$ as outliers.[6] Although quartiles are largely unaffected by the extreme values but they may be marginally affected in the case of hospital data because that may have a large number of abnormal values. This filtering will exclude almost all outliers but some abnormal values that are not clear outliers may remain and need to be further excluded.
- To get completely uncontaminated data, we use the same filter again on the data available after the first filtering. Recalculate IQR, identify the abnormal values which are either less than $Q_1 - 1.5 \times IQR$ or more than $Q_3 + 1.5 \times IQR$ with new values of Q_1 and Q_3 and exclude them. This procedure yields uncontaminated healthy values that are suitable to obtain normal reference interval with no distortion. In case the reference values are required for various age–gender groups, this exercise is required separately for each group. Laboratory data of large hospitals are still likely to have at least 120 individuals of each age–gender group after such double filtration and these many are adequate to work out normal reference interval as recommended by CLSI.[7]

The two filtrations as above should leave uncontaminated healthy values. To confirm that this is really so, prepare a histogram of these values and check that the distribution is smooth. This distribution will be mostly symmetric and Gaussian but can also be skewed to the right[3] if the lower values are abnormal or to the left if higher values are abnormal. We carried out this exercise for various liver function tests and observed that the filtered uncontaminated values of globulin followed a symmetric distribution, but the distribution of alkaline phosphatase (ALP) was slightly right skewed and of albumin slightly left skewed in our subjects. This is on expected pattern for healthy values and validates our procedure. For example, the most common value of ALP in our dataset of healthy values was around 85 U/L and the values less than this were less common, whereas values more than this up to 175 U/L were quite common.

The steps just outlined will produce a dataset of suitable healthy values that can be used to find the normal reference interval. The general procedure to establish and define normal reference interval of quantitative laboratory parameters is based on the mean and SD of the values found in healthy individuals as described in the Davidson's Medicine.[8] They define normal range as (mean – 2 SD, mean + 2 SD) that includes 95% of the healthy values if the distribution is Gaussian (normal). For many medical parameters, it is not so and, in those cases, similar limits for a skewed distribution are (2.5th percentile, 97.5th percentile). These limits also include 95% of healthy values. Note that we are concerned with the distribution of individual values and not means—thus a large sample does not help in this case in obtaining a Gaussian distribution. Fortunately, these percentile-based limits are the same as based on (mean, SD) in the case of a Gaussian distribution. Thus, the percentile-based limits have almost universal applicability—in Gaussian as well as in non-Gaussian distributions. We use these limits in this paper in place of (mean, SD)-based limits. Both exclude 5% extreme values—2.5% on either side.

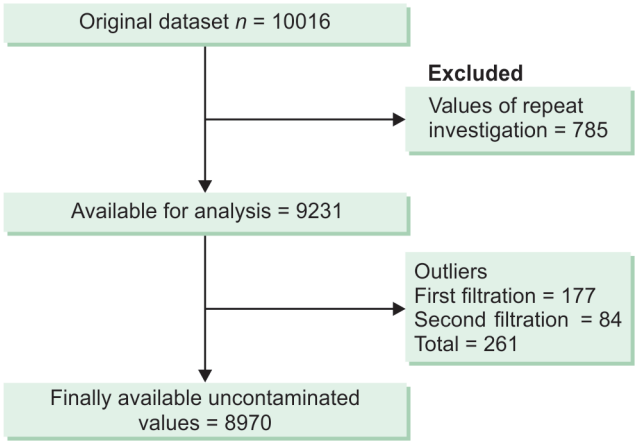
Results

We illustrate our method for A/G ratio for the values in the database of laboratory of our hospital. A/G ratio is among the most stable liver function parameter yet its reference interval is reported differently for various populations. For example, Furruqh et al.[9] report this to be (0.7–1.7) for Indian subjects and Myeloma[10] reports (0.8–2.0) for American subjects. Thus, different populations need their own reference interval for this parameter also.

We asked informatics technology (IT) section of our hospital to extract results of liver function tests, especially A/G ratio, for persons who were investigated in the month of July 2019. The A/G ratio was based on albumin level obtained by bromocresol green (BCG) method and globulin calculated by subtraction of albumin level from total protein where the total protein was obtained by Biuret endpoint method.

A total of 10,016 values were available. Of these, 785 values belonged to the persons who were investigated more than once. These were excluded right at the beginning and left 9231 values. The first filtration identified 177 outliers, and the second filtration further excluded 84 abnormal values. Thus, 8970 values remained for working out the reference interval (Flowchart 1). Since A/G ratio is a fairly stable parameter, it does not have as many outliers and

Flowchart 1: Available values of A/G ratio and analysis



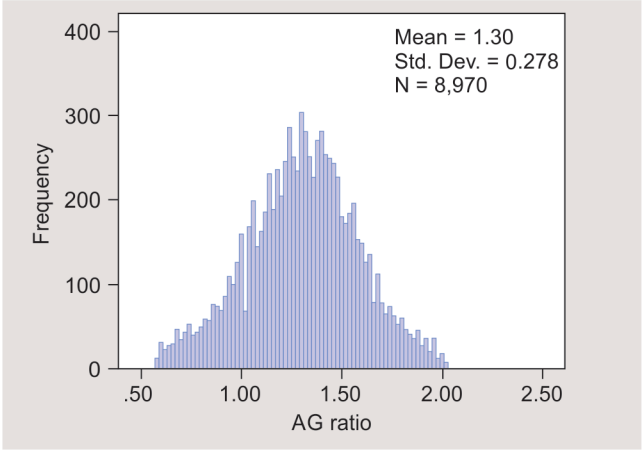


Fig. 1: Distribution of the values of A/G ratio after filtration

abnormal values as expected for other parameters in a hospital setup.

The statistical distribution of the values finally available for analysis was found to follow a smooth curve with slight skewness to the left (Fig. 1). We could have used mean ± 2 SD range as normal reference interval but we prefer (2.5th–97.5th percentiles) because of their almost universal applicability including for skewed distributions as stated earlier. This gives (0.7–1.8) as the normal reference interval of A/G ratio for our subjects. This interval is fairly close to those reported by Furruqh et al.[9] for Indian subjects, and provides indirect evidence of validity of our procedure.

Discussion

Our method of establishing normal reference interval of a quantitative medical parameter from hospital data assumes that the right analytical method has been followed to obtain the correct value of the analyte. With an automated system in place in almost all large hospitals, this would be fairly assured. In setups where this is not so, necessary precautions may have to be adopted to record correct values.

An exercise of the type recommended by us may not be enough by itself for establishing the reference interval. It is required to be validated using another database from the same milieu. If the validation set also gives nearly the same reference interval, the confidence rises manifold. In case different studies yield different but largely hemogeneous intervals, meta-analysis can be done to come to a consensus reference interval. This would entail, in this case, using random-effects model as discussed by Partlett and Riley.[11] A step further is harmonization of the laboratory results that strives to achieve the same result within clinically acceptable limits irrespective of the measurement procedure used, and when and where a measurement procedure is made.[12] Harmonization is not restricted to reference intervals but is highly desirable for all laboratory measurements.

Conclusion

Scientists in India and other developing countries have to make a beginning somewhere to delineate local reference intervals. In fact, these should be established for each laboratory.[7] We have tried to explain how this can be done with the data that at present is lying unused in large hospitals. We hope that the laboratory medicine professionals in India and other developing countries will come forward and establish normal reference intervals of medical parameters for their population using the method we have described.

References

1. Sairam S, Domalapalli S, Muthu S. Hematological and biochemical parameters in apparently healthy Indian population: defining reference intervals. Indian J Clin Biochem 2014;29(3):290–297. DOI: 10.1007/s12291-013-0365-5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4062657/.
2. Lang RM, Bierig M, Devereux RB, et al. Recommendations for chamber quantification. Eur J Echocardiogr 2006;7(2):79–108. DOI: 10.1016/j.euje.2005.12.014. http://www.pac4.ch/Pdf/AnnECardiol/Quantif%20Europ.pdf.
3. Indrayan A, Malhotra RK. Medical Biostatistics. 4th ed., Chapman & Hall/CRC Press; 2018. pp. 239–244.
4. Everitt BS, Palmer CR. Encyclopedic Companion to Medical Statistics. Wiley; 2011. p. 276.
5. Manikandan S. Measures of dispersion. J Pharmacol Pharmacother 2011;2(4):315–316. DOI: 10.4103/0976-500X.85931. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3198538/.
6. Indrayan A, Holt MP. Concise Encyclopedia of Biostatistics for Medical Professionals. Chapman & Hall/CRC Press; 2017. p. 427.
7. CLSI Clinical and Laboratory Standards Institute. EP 28-A3c: Defining, Establishing, and Verifying Reference Intervals as in the Clinical Laboratory – Approved Guidelines, Third Edition. CLSI 2010: p. 2. https://clsi.org/media/1421/ep28a3c_sample.pdf.
8. Walker BR, Colledge NR, Ralston SH, et al. Davidson's principles & practice of medicine. 22nd ed., Churchill Livingstone–Elsevier; 2014. p. 5.
9. Furruqh S, Anitha D, Venkatesh T. Estimation of reference values in liver function test in health plan individuals of an urban south Indian population. Ind J Clin Biochem 2004;19(2):72–79. DOI: 10.1007/BF02894260. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3454209.
10. Myeloma, Understanding your lab results. Myelomacentral.com https://www.myelomacentral.com/understanding-multiple-myeloma/understanding-your-lab-results/.
11. Partlett C, Riley RD. Random effects meta-analysis: coverage performance of 95% confidence and prediction intervals following REML estimation. Stat Med 2017;36(2):301–317. DOI: 10.1002/sim.7140. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5157768/.
12. Tate JR, Myers GL. Harmonization of clinical laboratory test results. E J Int Fed Clin Chem 2016;27:5–14. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975212/pdf/ejifcc-27-005.pdf.