Choice of Statistical Tools for Outlier Removal Causes Substantial Changes in Analyte Reference Intervals in Healthy Populations

Hickman, Peter E; Koerbin, Gus; Potter, Julia M; Glasgow, Nicholas; Cavanaugh, Juleen A; Abhayaratna, Walter P; West, Nic P; Glasziou, Paul

Published in:
Clinical Chemistry

DOI:
10.1093/clinchem/hvaa208

Clinical Chemistry 66:12
1558–1561 (2020)

Choice of Statistical Tools for Outlier Removal Causes Substantial Changes in Analyte Reference Intervals in Healthy Populations

Peter E. Hickman,[a,b] Gus Koerbin,[c] Julia M. Potter,[a,b] Nicholas Glasgow,[a] Juleen A. Cavanaugh,[a] Walter P. Abhayaratna,[c] Nic P. West,[d] and Paul Glasziou[e,*]

BACKGROUND: Reference intervals are an important aid in medical practice as they provide clinicians a guide as to whether a patient is healthy or diseased.

Outlier results in population studies are removed by any of a variety of statistical measures. We have compared several methods of outlier removal and applied them to a large body of analytes from a large population of healthy persons.

METHODS: We used the outlier exclusion criteria of Reed-Dixon and Tukey and calculated reference intervals using nonparametric and Harrell-Davis statistical methods and applied them to a total of 36 different analytes.

RESULTS: Nine of 36 analytes had a greater than 20% difference in the upper reference limit, and for some the difference was 100% or more.

CONCLUSIONS: For some analytes, great importance is attached to the reference interval. We have shown that different statistical methods for outlier removal can cause large changes to reported reference intervals. So that population studies can be readily compared, common statistical methods should be used for outlier removal.

Comparison is an important element of clinical medicine, with clinicians using quantitative benchmarks for normality to assess how a particular case may be different and abnormal. In the clinical laboratory, reference intervals for analytes act as this benchmark because they provide an objective measure as to whether homeostasis is normal.

Typically, the central 95th percentile of a healthy population is used to define reference intervals, although there are important exceptions such as the 99th percentile for troponin *(1)* and the use of decision limits rather than reference intervals for some analytes such as cholesterol and hemoglobin A1c *(2)*.

Even when assessing an objectively healthy population, there are invariably a few outlier results, and use of a variety of empirical procedures to remove these outliers has become accepted practice *(3–5)*.

A recent article by Eggers showed that the 99th percentile for troponin was markedly different, depending upon which method for outlier removal was chosen *(6)*. We have replicated their findings using data from our large Canberra Heart Study database *(7)*.

The problem of outlier removal is not unique to troponin and the 99th percentile *(8)*. We have explored our large Aussie Normals database which we used to establish reference intervals for a large number of general chemistry analytes *(9)* and thyroid hormones *(10)* and report on the effect of different statistical methods for outlier removal on derived reference intervals for all these analytes.

Materials and Methods

Different parts of this study were approved by the Australian Capital Territory (ACT) Health Human Research Ethics Committee, the Australian National University Human Research Ethics Committee, or the Australian Institute of Sport Research Ethics Committee.

POPULATION STUDIED

Details of the recruitment into, and running of the Aussie Normals study have been published in detail elsewhere *(9)*. Briefly, a total of 1856 persons were recruited largely by invitation of persons selected from the local electoral roles as typical of the ACT demographic, and by advertising in local media. We excluded persons who were pregnant or had diabetes mellitus, asthma requiring oral steroids, any history of malignancy and other conditions involving systemic disease known to bias

[a]Australian National University Medical School, Garran, ACT, Australia; [b]ACT Pathology, Canberra Hospital, Garran, ACT, Australia; [c]College of Medicine Biology and Environment, Australian National University, Garran, ACT, Australia; [d]Griffith University, Brisbane, QLD, Australia; [e]Bond University, Robina, QLD, Australia.
*Address correspondence to this author at: Institute for Evidence-Based Healthcare, Bond University, Robina, Queensland 4226, Australia. E-mail pglaszio@bond.edu.au.
Received June 12, 2020; accepted August 14, 2020.
DOI: 10.1093/clinchem/hvaa208

biochemical concentrations. We accepted persons taking statins, the oral contraceptive pill, or hormone replacement therapy, and analytes affected by these medications were excluded from analysis.

LABORATORY METHODS

Thirty-six (36) routine laboratory assays were assessed using proprietary methods and performed on either an Abbott Architect ci8200 or ci16200, Metrological traceability is described by the manufacturer in its information for users (IFU). Hemolysis (H), icterus (I), and lipemia (L) was assessed on all samples using on-board HIL automated procedures.

The performance characteristics for the assays showed coefficients of variation (CVs) ranging from 0.8% to 5.5% for kidney, liver, bone minerals, specific protein, and iron analysis. Homocysteine showed a CV of <8% with thyroid stimulating hormone (TSH) and free thyroxine (FT4) <4.9% and free triiodothyronine (FT3) <15%.

STATISTICAL METHODS

Evaluation of the 36 reference intervals from the Aussie Normals study was undertaken using the outlier exclusion criteria of Reed-Dixon *(5, 11, 12)* and Tukey *(12)* using 1.5 and 3.0 × IQR fence criteria. Reference intervals were calculated using nonparametric (Analyse-it Software, Ltd.) and Harrell Davis statistical methods *(13)*. The differences in upper and lower reference limits (URL, LRL) were then compared against biological variation CVi, as shown by Westgard using the 2012 updated Ricos data *(14)* and the allowable limits of performance used by the Royal College of Pathologists of Australasia Quality Assurance Program (RCPAQAP) *(15)*.

Results

Of the reference intervals calculated, 27/36 showed little difference when the 3 outlier exclusion criteria and statistical methods were compared. Those analytes were sodium, potassium, chloride, bicarbonate, urea, total and direct bilirubin, total protein, albumin, uric acid, creatinine, lactate dehydrogenase (LD), alkaline phosphatase (ALP), calcium, magnesium, phosphate, C3, C4, alpha-1 antitrypsin, immunoglobulin G (IgG), haptoglobin, iron, transferrin, ferritin (women), homocysteine (women), FT4, and FT3.

Using the Tukey 1.5 x IQR fence exclusion criteria as the reference method, lipase, IgA, IgM, alanine transaminase (ALT), creatine kinase (CK), homocysteine (men), ferritin (men), TSH, and amino terminal pro-B-type natriuretic peptide (NT-proBNP) showed differences, (higher URL results) when compared with Reed Dixon and the 3.0 × IQR fence exclusion method with both nonparametric Harrell Davis statistical analysis. Aspartate transaminase (AST) and gamma glutamyl transpeptidase (GGT) only showed statistical differences against the Reed Dixon criteria (higher URL). The numerical differences are shown in Table 1. The URL's for creatine kinase (CK) (women) and GGT (men and women) using the Tukey exclusion criteria showed differences between the nonparametric and Harrell Davis analysis when compared with the CVi and RCPAQAP allowable limits. No statistical differences in URLs were seen between the nonparametric and Harrell Davis analysis when using the Reed Dixon exclusion criteria.

No differences were seen in the LRLs between the nonparametric and Harrell Davis statistical methods using the Reed Dixon or Tukey 1.5 × IQR fence criteria except for creatinine using the Tukey 1.5 × IQR fence criteria with both male and female ranges showing results that were 20% lower using the Harrell Davis statistical method.

Discussion

Using a large population of healthy persons, we have shown that for several very important analytes, the choice of statistical method used for outlier removal causes a substantial change in either the upper or lower boundary of the reference interval. The Tukey method typically led to a narrower reference interval. A Tukey fence of 1.5 × IQR is approximately equivalent to 3 SDs or 1%, and hence will remove about 1% even from a truly healthy population. Hence this outlier exclusion means a 95% reference range is effectively a 94% reference range—with 6% being classed as abnormal (1% from the fence; 5% from the definition). Thus with Tukey we would be looking at a reference interval of 3.0%–97.0%. For a 99% reference interval, the Tukey fence implies a 98% equivalent—or 2% (1% from the fence; 1% from the definition of a 99% interval), so the reference interval would become 0.5–98.5% or <98.5%. Some statistical methods have been developed to adjust for this exclusion problem *(16)*.

Twenty-seven of 36 analytes showed little difference in derived reference intervals, regardless of which statistical method is used. Analytes with tight physiological control (e.g, sodium) have a much smaller spread compared with analytes such as transaminases. It is these latter analytes that are more likely to be affected by different statistical manipulations.

Defining a truly healthy population is problematic. Age is associated with a higher troponin 99th percentile, even in healthy persons who have gone through rigorous health checks *(17)*. Mild obesity is associated with higher ALT activity in apparently healthy men and

Table 1. Reference intervals for 9 of 36 analytes using Harrel-Davis analysis and Tukey or Reed-Dixon exclusion criteria for eliminating outliers where a difference of >20% in Upper Reference Limit (URL) was observed between exclusion methods.

	Tukey (1.5 IQR)		Tukey (3.0 IQR)		Reed Dixon exclusion	
	Non parametric	Harrell Davis	Non parametric	Harrell Davis	Non parametric	Harrell Davis
NT-proBNP (ng/L)#	6 - 246	6-249	6-341	6-342	6-1017	6-1035
ALT (M) (U/L)	10 - 41	11-40	11-49	10-48	11-53	10-48
ALT (F) (U/L)	9-33	8-31	8-39	8-40	8-44	8-45
AST (M) (U/L)	14-35	15-33	14-37	14-38	14-44	14-44
AST (F) (U/L)	13-32	13-31	13-35	13-35	13-39	13-39
GGT (M) (U/L)	12-71	10-57	12-79	12-80	12-101	12-108
GGT (F) (U/L)	9-56	9-43	9-63	9-64	9-88	9-90
Homocysteine (M) (μmol/L)	7.6-16.5	7.6-15.5	7.7-18.3	7.7-18.4	7.7-19.2	7.6-19.4
CK (M) (U/L)	49-252	49-266	50-314	49-321	50-360	49-376
CK (F) (U/L)	37-221	36-182	37-229	37-230	37-278	36-290
Ferritin (M) (μg/L)	20-318	20-312	20-425	21-432	21-432	21-434
TSH (M+F) (mU/L)	0.42-2.67	0.42-2.68	0.43-3.41	0.42-3.41	0.43-3.39	0.42-3.41
cTnI (M) (ng/L)	<6.0[a]	<6.6[a]	<9.1[a]	<9.7[a]	<27.3[a]	<22.8[a]
cTnI (F) (ng/L)	<5.7[a]	<5.9[a]	<9.7[a]	<9.7[a]	<13.1[a]	<15.0[a]

[a]cTnI 99th percentile.
#Abbreviations: NT-proBNP–amino terminal pro-B-type natriuretic peptide; ALT–alanine transaminase; AST–aspartate transaminase; GGT–gamma glutamyl transpeptidase; CK–creatine kinase; TSH–thyroid stimulating hormone; cTnI–cardiac troponin I.

women *(9)*. Even low levels of activity can cause increases in CK activity. As there is often a gradual transition from health to disease, it is very difficult to ensure that all persons included are truly healthy.

Considerable importance can be attached to reference intervals. Examples are NT-proBNP used for the defining of persons with heart failure *(18)*, ferritin for assessing persons with possible hemochromatosis *(19)*, and TSH is of considerable concern in looking at women with possible mild thyroid disease in pregnancy *(20)*. ALT reference intervals are of importance in assessing persons with mild hepatocellular disease, especially the effect of alcohol *(21)*. Currently, quoted ALT reference intervals for men can vary as widely as 10–68 *(22)* and 10–41 *(9)*. Similar discord is seen with reference intervals for women. The American College of Gastroenterologists have recommended an upper reference limit for men of 33 and for women of 25 *(21)*, using the rationale that no cases of pathology should be missed, but at the risk of gross overinvestigation of many persons without pathology. The substantial overlap in results between healthy persons and those with some degree of liver pathology means no quoted reference interval will satisfy all parties.

The Clinical and Laboratory Standards Institute (CLSI) has published a detailed monograph on establishing reference intervals *(5)*. They have concentrated on the choice and sampling of the healthy population used for establishing a reference interval and to a lesser extent with the statistical treatment of data. This article indicates that the statistical method can make a considerable difference, and hence is a matter of importance for all analytes. The CLSI should develop a position statement on the preferred statistical treatment for outlier removal so that population studies can be meaningfully compared.

Nonstandard Abbreviations: LRL, lower reference limit; URL, upper reference limit; CLSI, Clinical and Laboratory Standards Institute; RCPAQAP, Royal College of Pathologists of Australasia Quality Assurance Program

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 4 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; (c) final approval of the published article; and (d) agreement to be accountable for all aspects of the article thus ensuring that questions related to the accuracy or integrity of any part of the article are appropriately investigated and resolved.*

P.E. Hickman, statistical analysis, administrative support, provision of study material or patients; J. Cavanaugh, financial support, provision of study material or patients; W.P. Abhayaratna, provision of study material or patients; N. West, provision of study material or patients.

Authors' Disclosures or Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: No sponsor was declared.

References

1. Jaffe AS, Ravkilde J, Roberts R, Naslund U, Apple FS, Galvani M, Katus H. It's time for a change to a troponin standard. Circulation 2000;102:1216–20.
2. Ozarda Y, Sikaris K, Streichert T, Macri J; IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). Distinguishing reference intervals and clinical decision limits–A review by the IFCC Committee on Reference Intervals and Decision Limits. Crit Rev Clin Lab Sci 2018; 55:420–31.
3. Ozarda Y. Reference intervals: current status, recent developments and future considerations. Biochem Med 2016;26:5–11.
4. Ceriotti F, Hinzmann R, Panteghini M. Reference Intervals: the way forward. Ann Clin Biochem 2009;46: 8–17.
5. Horowitz GL, Altaie A, Boyd JC, Ceriotti F, Garg U, Horn P. Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline. 3rd Ed. Wayne (PA): Clinical and Laboratory Standards Institute; 2008. p. 1–76.
6. Eggers KM, Apple FS, Lind L, Lindahl B. The applied statistical approach highly influences the 99th percentile of cardiac troponin I. Clin Biochem 2016;49:1109–12.
7. Hickman PE, Koerbin G, Potter JM, Abhayaratna WP. Statistical considerations for determining high-sensitivity cardiac troponin reference intervals. Clin Biochem 2017; 50:502–5.
8. Hickman PE, Koerbin G, Saenger AK, Kavsak PA. Statistical issues with the determination of the troponin 99th percentile–not just a problem for troponin? Clin Biochem 2016;49:1105–6.
9. Koerbin G, Cavanaugh JA, Potter JM, Abhayaratna WP, West NP, Glasgow N, et al. "Aussie Normals"–an a priori study to develop clinical chemistry reference intervals in a healthy Australian population. Pathology 2015;47: 138–44.
10. Hickman PE, Koerbin G, Simpson A, Potter JM, Hughes DG, Abhayaratna WP, et al. Using a thyroid disease-free population to define the reference interval for TSH and free T4 on the Abbott Architect analyser. Clin Endocrinol 2017;86:108–12.
11. Dixon WJ. Processing data for outliers. Biometrics 1953; 9:74–89.
12. Horn PS, Feng L, Li Y, Pesce AJ. Effect of outliers and non-healthy individuals on reference interval estimation. Clin Chem 2001;47:2137–45.
13. Wessa P. 2016. Harrell-Davis Quantile Estimator (v1.0.13) R code. Free Statistics Software, Office for Research Development and Education, v1.1.23-r7. http://www.wessa.net/ (Accessed September 2017).
14. Ricos C, Alvarez V, Cava F, Garcia-Lario JV, Hernandez A, Jimenez C. Current database on biologic variation: pros, cons and progress. Scan J Clin Lab Invest 1999;59: 491–500.
15. Royal College of Pathologists of Australasia Quality Assurance Program. http://www.rcpaqap.com.au/docs/2014/chempath/ALP.pdf (Accessed November 2019).
16. Beasley CM, Crowe B, Nilsson M, Wu L, Tabbey R, Hietpas RT, et al. Adaptation of the robust method to large distributions of reference values: program modifications and comparison of alternative computational methods. J Biopharm Stat 2019;29:516–28.
17. Hickman PE, Abhayaratna WP, Potter JM, Koerbin G. Age-related differences in hs-cTnI concentration in healthy adults. Clin Biochem 2019;69:26–9.
18. Alehagen U, Goetze JP, Dahlstrom U. Reference intervals and decision limits for B-type natriuretic peptide and its precursor NT-proBNP in the elderly. Clin Chim Acta 2007;382:8–14.
19. Malton K, Turnock D. A short report: reflective testing in the diagnosis of hereditary haemochromatosis: results of a short retrospective study. Ann Clin Biochem 2019;56: 408–10.
20. Medici M, Korevaar TI, Visser WE, Visser TJ, Peeters RP. Thyroid function in pregnancy: what is normal? Clin Chem 2015;61:704–13.
21. Kwo PY, Cohen SM, Lim JK. ACG clinical guideline: evaluation of abnormal liver chemistries. Am J Gastroenterol 2017;112:18–35.
22. Rustad P, Felding P, Franzson L, Kairisto V, Lahti A, Mårtensson A, et al. The Nordic Reference Interval Project 2000: recommended reference intervals for 25 common biochemical properties. Scand J Clin Lab Invest 2004;64:271–84.