Widespread Incorrect Implementation of the Hoffmann Method, the Correct Approach, and Modern Alternatives

Daniel T. Holmes, MD,[1,2] and Kevin A. Buhr, PhD[3]

From the [1]Department of Pathology and Laboratory Medicine, St Paul's Hospital, Vancouver, Canada; [2]Department of Pathology and Laboratory Medicine, University of British Columbia, Vancouver, Canada; and [3]Biostatistics and Medical Informatics, University of Wisconsin, Madison.

Key Words: Bhattacharya; Estimation; Hoffmann; Indirect; Mixture model; Reference interval

ABSTRACT

Objectives: The Hoffmann method is a procedure for reference interval estimation using routine clinical results. Many authors incorrectly prepare Hoffmann plots on a linear rather than normal probability scale. We explore the consequences.

Methods: This was investigated algebraically, by random number simulations (45 simulations, n = 100,000 each) and using clinical data sets. Strategies compared were: Hoffmann's method as originally and incorrectly implemented, Bhattacharya's method, and maximum likelihood (ML). All R source code and data sets are provided.

Results: As the proportion of healthy individuals approaches 1, the incorrect approach generates reference interval estimates of approximately $\mu_H \pm 1.19\ \sigma_H$ delineating the central 77% of the healthy subpopulation, not the central 95%. Inappropriately narrow reference interval estimates were seen on random simulations and clinical data sets. ML methods performed best.

Conclusions: The erroneous variant Hoffmann method should not be used. ML methods outperform others and are not restricted by Gaussian assumptions.

In 1963, Hoffmann described a simple graphical method for estimating reference intervals from routine laboratory data.[1] He assumed that the distribution of results could be represented by a mixture of two underlying Gaussian distributions, corresponding to the healthy and diseased subpopulations, with the healthy subpopulation dominating the sample ❚Figure 1❚.

Hoffmann's method consisted of tallying the full set of results into a set of ordered categories representing measurement ranges ("bins"), calculating the cumulative frequencies of the categories and converting them to percentages, and using normal (Gaussian) probability paper to plot the cumulative percentages (on the y-axis with a Gaussian probability scale) against the measurement values corresponding to the category endpoints (on the x-axis with a linear scale).

Hoffmann demonstrated that under these assumptions, the result was a plot with two regions of approximate linearity corresponding to the healthy and diseased subpopulations. By extrapolating the linear region for the healthy individuals to the horizontal lines representing the 2.5th and 97.5th percentiles, the estimated reference interval corresponding to the central 95% of the healthy subpopulation could be read from the x-axis. ❚Figure 2❚ illustrates the procedure for a set of n = 100 simulated observations drawn from the theoretical density shown in Figure 1 where the procedure gives an estimated reference range of 3.0 to 7.3 compared to the actual reference range derived from simulation parameters of 3.0 to 7.0.

While "binning" the data made the traditional manual procedure more convenient, it is not a critical component of the approach. A correct modern approach involves sorting the n measurements, assigning each measurement its corresponding cumulative probability (ie, assigning the ith ordered measurement the cumulative probability i / n

or using a similar formula with a small continuity correction), and plotting those probabilities (y-axis) against the measurement values (x-axis), thereby forming an empirical cumulative distribution function (CDF) of the sample.
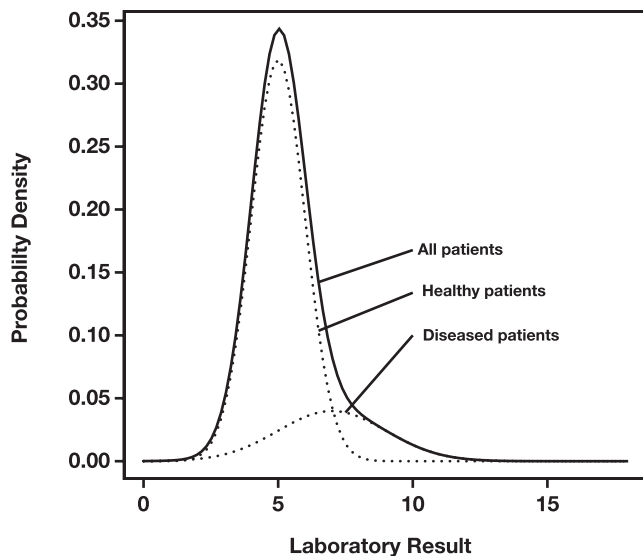

Figure 1 Density of a distribution of mock patient data for a hypothetical analyte consisting of a mixture of Gaussian distributions representing healthy (subscript H) patients ($\mu_H = 5$, $\sigma_H = 1$) and diseased (subscript D) patients ($\mu_D = 7$, $\sigma_D = 2$). Solid line, overall density; dotted lines, component densities.

However, a critical component of the procedure is that the CDF be plotted using a Gaussian probability scale for the y-axis or, equivalently, that corresponding quantiles of a standard Gaussian distribution be plotted on a linear y-axis **Figure 3**. Such a plot is known as a "normal probability plot" and is a specific type of quantile-quantile (QQ) plot for which the comparator distribution is the standard Gaussian distribution.[2]

The Hoffmann method has been widely applied in the field of clinical chemistry for reference interval derivation and reference interval validation.[3-22] There is also interest in automating the procedure.[14,15]

Early papers describing or applying the method were true to the methodology as originally described by Hoffmann and clearly explained the purpose and necessity of using a Gaussian probability scale for the cumulative probabilities plotted on the y-axis.[23-25] Unfortunately, with only a few notable exceptions,[18,26,27] a host of modern papers[8,12-17,19] and even some textbooks[11] explicitly show the use of an incorrect method that involves plotting a CDF of the sample in purely linear space.

While it may seem inconsequential, this methodological deviation results in incorrect reference interval estimates in all circumstances, even in the ideal case of a single, unmixed Gaussian distribution representing measurements from only healthy individuals. Correspondingly, the parameter estimates of μ_H and σ_H may be poor, the latter being generally (but not universally) too small,

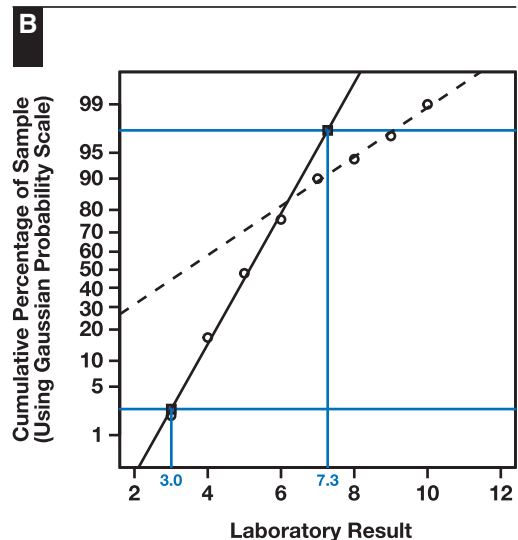| Result	Tally	Count	Cumulative Frequency	Cumulative %
1-2		0	0	0
2-3	II	2	2	2
3-4	IIIIIIIIIIIIIII	15	17	17
4-5	IIIIIIIIIIIIIIIIIIIIIIIIIIIIIII	31	48	48
5-6	IIIIIIIIIIIIIIIIIIIIIIIIIIII	28	76	76
6-7	IIIIIIIIIIIIII	14	90	90
7-8	IIII	4	94	94
8-9	III	3	97	97
9-10	II	2	99	99
10-11	I	1	100	100
11-12		0	100	100

Figure 2 Original Hoffmann method applied to a simulated sample of $n = 80$ healthy and $n = 20$ diseased individuals from the mixture distribution shown in Figure 1. **A**, Tally of sample into a set of measurement ranges together with cumulative percentages. **B**, Hoffmann plot of right-midpoint of range (x-axis) vs cumulative percentage on Gaussian probability scale (y-axis), with extrapolated linear regions corresponding to healthy (solid line) and diseased (dashed line). Gray horizontal lines represent the 2.5th and 97.5th percentiles, and where those lines intersect the extrapolated "healthy" linear region, the x-coordinates give the estimated reference interval.
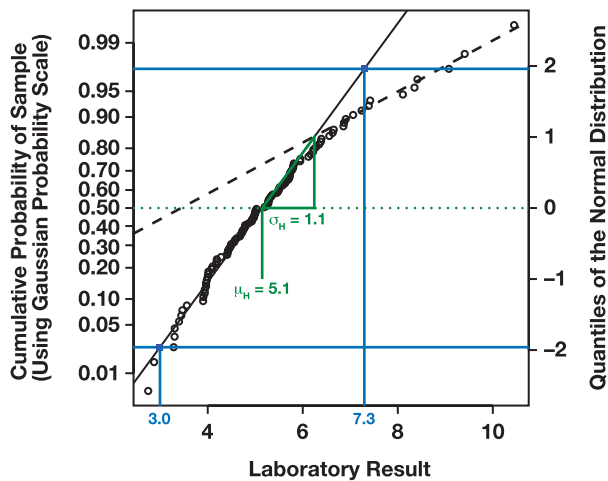
■**Figure 3**■ Correct modern variant of the Hoffmann method, plotting the measurement values (x-axis) against the cumulative probabilities of the sorted sample using a Gaussian probability scale (left y-axis) or equivalently against the corresponding quantiles of a standard Gaussian distribution (right y-axis). Extrapolation of the linear segment for the "healthy" linear region and estimation of the reference range may proceed as with the original method (blue), or by first estimating the mean and using the slope against the right (linear) y-axis to calculate σ_H by $\sigma_H = 1/\text{slope}$ (green).

resulting in reference intervals much narrower than those generated by the correctly formulated Hoffmann method.

Although many approaches exist for Gaussian mixture model decomposition,[28] in clinical chemistry attention has been focused primarily on the strategies of Hoffmann and Bhattacharya,[29] likely because both are graphical in nature. The Bhattacharya method relies on first categorizing the measurements of laboratory quantity x, into equally spaced classes (bins) of width h. This results in frequencies, denoted y, for each class. The quantity $\Delta \log(y) = \log(y(x + h)) - \log(y(x))$ is then plotted against x and, for appropriately chosen widths h, the resulting plot will have regions where the curve straightens and has a negative slope. Extrapolation of these linear regions allows the determination of the x-intercept and the angle formed with the negative x-axis from which the mean, standard deviation, and proportion p for each contributing mode can be calculated.[29] The method remains popular.[30]

In contrast, a modern (nongraphical) computational strategy for this problem is the use of maximum likelihood (ML) through the expectation maximization algorithm.[31] A description of how ML methods work is beyond the scope of this paper and the interested reader is directed to a review on the topic.[28] Importantly, however, the ML strategy for this and other problems has been implemented in relatively easy to use open-source computational packages for the R statistical programming language.

In what follows, we will demonstrate the magnitude of the error associated with the use of a linear CDF, both algebraically and with random number simulations. We will also compare different methods for parameter estimation: the Hoffmann method (both correctly and incorrectly implemented), the Bhattacharya method, and modern ML approaches using a select two[32,33] of several freely available software packages[32-36] implemented using the R statistical programming language, which can be applied to this and other related problems.

Materials and Methods

Using algebraic methods, we analyzed and compared the reference interval estimates generated by both a correct modern variant of the Hoffmann method and the commonly used but incorrect implementation using a linear CDF. We further characterized the theoretical behavior of estimates generated by each method in a number of extreme, but illustrative, cases: a single Gaussian distribution representing a population of healthy individuals only, and examples with both negligible and very large differences in healthy and diseased subpopulation means. We also described a typical illustrative case with a large healthy subpopulation and a small diseased subpopulation having modest separation between the subpopulation means.

We then compared approaches with random number simulations. Samples of size n = 100,000 from two-component Gaussian mixture distributions were randomly generated to simulate a variety of situations: with $\mu_H = 10$ and $\mu_H = 1$ fixed, μ_D and σ_D ranged from 12 to 18 and 1 to 3, respectively, and ρ_H ranged from 0.7 to 0.9 (45 combinations in total). Each simulated sample was evaluated using multiple procedures to recover μ_H, σ_H, the reference interval $\mu_H \pm 1.96\sigma_H$ and other parameters where applicable. The procedures employed were: a correct modern variant of the Hoffmann method, the incorrect Hoffmann variant, the Bhattacharya method, and a modern ML method from the R mixtools package.[33] Hoffmann and Bhattacharya linear sections were identified using visual oversight as prescribed. ML methods were provided crude starting parameter estimates of $\mu_H = 11$, $\mu_D = 15$, $\sigma_H = 0.5$, and $\sigma_D = 1.5$ for all simulations. Procedures were evaluated for their performance in reproducing the correct results based on the known simulation parameters.

We also applied the procedures to deidentified, routine clinical laboratory datasets to illustrate performance differences between approaches. The datasets used included a hemoglobin (Hb) dataset with easily resolved and

approximately Gaussian healthy and diseased subpopulations, a thyroid stimulating hormone (TSH) dataset with a highly skewed distribution and multiple (hyperthyroid and hypothyroid) diseased subpopulations, and a plasma calcium dataset with poor separation between the healthy and diseased subpopulations. Hb analyses were performed on a Sysmex XN-3000, TSH analyses were performed on a Roche Cobas e601, and calcium analyses were performed using the Arsenazo III method on the Siemens Advia 1800. All analyses were performed according to manufacturer specifications. Results for Hb, TSH, and calcium were extracted from laboratory information system from the entire 2016 calendar year. For clinical data, in addition to the parameter estimation methods applied to simulation data, the R mixdist[32] package was employed because it permits the use of other distributions, including the gamma distribution which, depending on its so-called shape and rate parameter values, can both exhibit skewing and approximate the normal distribution.[37]

This study was waived from ethics review by the research ethics boards of St Paul's Hospital and the University of British Columbia as a quality initiative. All computational analyses were performed using R version 3.4.4 (supplemented with the dplyr, magrittr, mixtools, mixdist, and xtable R packages). Our work is an example of reproducible research,[38] and the Supplemental Data includes the article in the form of a literate program (including all text and source code in a single document) written using the rmarkdown package[39] (all supplemental data are available at *American Journal of Clinical Pathology* online).

Results

Algebraic Results

As shown in Supplemental Appendix 1, the correct modern variant of the Hoffmann method (as illustrated in Figure 3) will give estimated upper and lower limits of normal (LLN and ULN) approximated by the formula:

$$\text{Correct Limits} \simeq \mu_H + \frac{A \times B}{\frac{1}{\sqrt{2\pi}}\frac{p_H}{\sigma_H} + \frac{p_D}{\sigma_D}\phi(z_D)}$$

$$\text{where } A = \left(\pm z_{\alpha/2} - \Phi^{-1}\left(\frac{1}{2}p_H + p_D\Phi(z_D)\right)\right)$$

$$B = \phi\left(\Phi^{-1}\left(\frac{1}{2}p_H + p_D\Phi(z_D)\right)\right)$$

and where $\Phi(t)$ and $\phi(t)$ are the CDF and probability density function, respectively, of a standard Gaussian distribution, $z_D = (\mu_H - \mu_D)/\sigma_D$, and α determines the size of the $(1-\alpha)\times 100\%$ normal range (eg, $\alpha = 0.05$ for a 95% normal range).

In contrast, if the operator employs an incorrect variant of the Hoffmann method by plotting an empirical CDF in purely linear space and fitting a line to its apparently linear section, the resulting LLN and ULN will be given by the approximate formula:

$$\text{Incorrect normal limits} \simeq \mu_H + \frac{\frac{1}{2} \pm \frac{1-\alpha}{2} - \left(\frac{1}{2}p_H + p_D\Phi(z_D)\right)}{\frac{1}{\sqrt{2\pi}}\frac{p_H}{\sigma_H} + \frac{p_D}{\sigma_D}\phi(z_D)}$$

While these results have a number of ramifications (Supplemental Appendix 1), the most striking pertains to the situation where the sample population is predominantly healthy, which is the usual context for use of indirect reference interval estimation methods.[40,41] Specifically, if p_H is close to 1, then, as shown in Supplemental Appendix 1, the correct and incorrect LLN and ULN estimates (for $\alpha = 0.05$) will approach:

$$\text{Correct normal limits} \quad \begin{aligned} &\simeq \mu_H \pm z_{\alpha/2}\sigma_H \\ &\approx \mu_H \pm 1.96\sigma_H \end{aligned}$$

$$\text{Incorrect normal limits} \quad \begin{aligned} &\simeq \mu_H \pm \sqrt{\pi/2}(1-\alpha)\sigma_H \\ &\approx \mu_H \pm 1.19\sigma_H \end{aligned}$$

This means that the reference intervals generated by the incorrect variation will be approximately 40% too narrow, with the LLN and ULN delineating the central 77% of the distribution rather than its central 95%. In contrast, the Hoffmann approach as originally described and its correct modern variant will produce LLN and ULN estimates that converge to $\mu_H \pm 1.96\sigma$ under these same circumstances. This finding is confirmed using random number simulations in Supplemental Appendix 1 along with a number of other illustrative examples.

Simulation Results

Supplemental Figures 1-9 show the results of 45 simulated scenarios (five scenarios per figure). Each is a sample of size $n = 100,000$ simulated from a two-component Gaussian mixture with simulation parameters as summarized in Supplemental Table 1. For each simulation, the corresponding figure shows a histogram of the resulting mixture distribution and the application of the Bhattacharya method, a correct modern variant of the Hoffmann method, and the incorrect Hoffmann variant. The estimates of the means, standard deviations, and reference intervals (LLN and ULN) for these three graphical methods and modern ML methods, together with the error associated with each method, are shown in Supplemental Table 1.
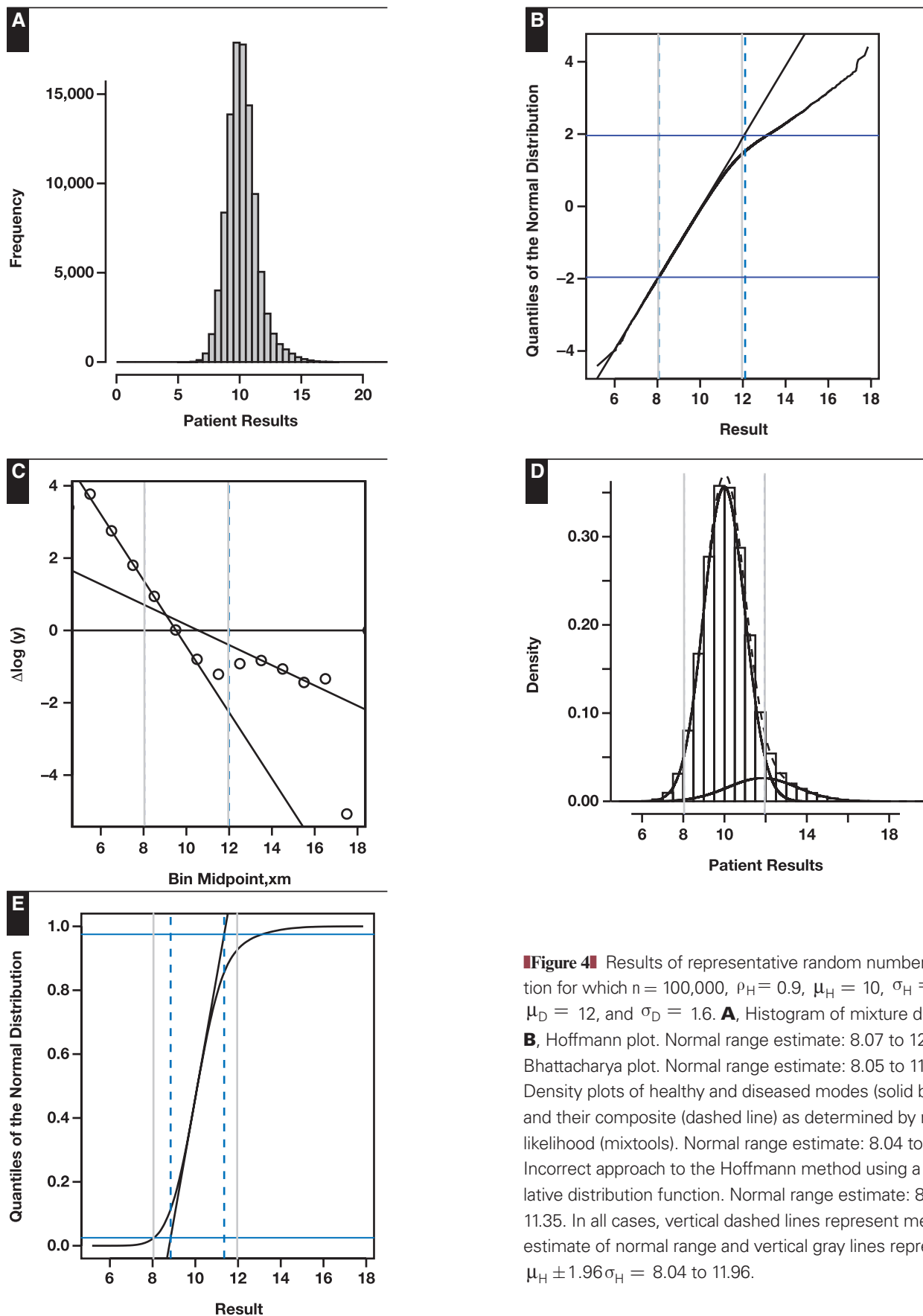
Figure 4 Results of representative random number simulation for which n = 100,000, $\rho_H = 0.9$, $\mu_H = 10$, $\sigma_H = 1$, $\mu_D = 12$, and $\sigma_D = 1.6$. **A**, Histogram of mixture distribution. **B**, Hoffmann plot. Normal range estimate: 8.07 to 12.11. **C**, Bhattacharya plot. Normal range estimate: 8.05 to 11.99. **D**, Density plots of healthy and diseased modes (solid black lines) and their composite (dashed line) as determined by maximum likelihood (mixtools). Normal range estimate: 8.04 to 11.95. **E**, Incorrect approach to the Hoffmann method using a cumulative distribution function. Normal range estimate: 8.84 to 11.35. In all cases, vertical dashed lines represent method estimate of normal range and vertical gray lines represent $\mu_H \pm 1.96\sigma_H = 8.04$ to 11.96.

The tendency of the incorrect variant of the Hoffmann method to produce reference intervals that are too narrow as $p_H \rightarrow 1$ is illustrated in ❚Figure 4❚ and Supplemental Figure 1, Figure 4, and Figure 7 (where $p_H = 0.9$). The simulated example in Supplemental Appendix 1, Section A.4.1 illustrates the same phenomenon in the extreme case where $p_H = 1.0$, representing samples taken from an entirely healthy population.

Across the 45 random simulations, the median absolute errors of the estimates using the correct modern variant of the Hoffmann method were 2.2% for μ_H and 7.2% for σ_H. In comparison, the errors for the widely used but incorrect variant using a CDF in linear space were 3.1% and 25%, respectively; errors for the Bhattacharya method were 0.5% and 9.9%; and errors for modern ML methods using the mixtools package were 0.0% and 0.3%. These results are demonstrated graphically in ❚Figure 5❚.

Application to Real Laboratory Data

Results of applying these methods to real patient data—adult male outpatient Hb, adult TSH (inpatient and outpatient), and adult plasma calcium (inpatient and outpatient)—are shown in ❚Table 1❚, with additional detail given in Supplemental Appendix 2. For ML analyses, TSH results greater than 15 mIU/L and calcium results below the 1st percentile and above the 99th percentile were removed prior to analysis to facilitate convergence.

For the Hb and calcium datasets, where the reference interval estimation problem is more straightforward, the incorrect variant of Hoffmann's method using a CDF in linear space produces narrower reference intervals and
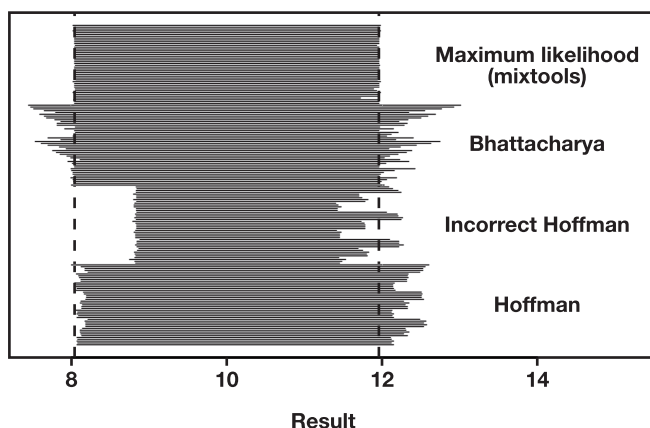

❚Figure 5❚ Reference interval estimates across 45 random simulations (n = 100,000 each) as represented by 45 horizontal lines per method spanning the range of the calculated upper and lower limits. The vertical dashed lines represent the target results of 8.04 and 11.96.

significantly different parameter estimates, compared to the other methods. For the TSH example, comparison of methods is complicated by the highly skewed distribution and the multiple diseased subpopulations (hypo- and hyperthyroidism). The distribution of TSH data does not satisfy the assumptions of the Hoffmann, Bhattacharya, or modern Gaussian-based ML methods, and all methods produce estimates that appear clinically incorrect, with ULN that seem too low. However, using modern ML methods with non-Gaussian model assumptions (a mixture of gamma distributions in this case), provides a superior fit and produces clinically plausible reference interval estimates (Supplemental Appendix 2).

Discussion

Random simulations illustrate a number of points. First, the normal QQ plot used in the correct implementation of the Hoffmann method is very sensitive to deviations from a Gaussian distribution. This has the effect of clearly resolving the linear segments corresponding to the healthy and diseased subpopulations and facilitates proper identification of the former. In contrast, the incorrect approach typically produces a sigmoidal shape in all situations, with no clear delineation of the healthy and diseased subgroups. The apparently linear region in the middle portion of the CDF does not necessarily correspond preferentially to the healthy subpopulation as the contributions of the two subpopulations tend to blend imperceptibly. Extension of this linear section has no particular meaning beyond identifying the steepest tangent line to the CDF, and the resultant parameter estimates are related in uninformative ways to the correct results. This effect can be appreciated by inspecting the progressions seen as μ_D approaches μ_H in the Supplemental Figures and in a representative simulation shown in Figure 4.

Second, while the Hoffmann method, when correctly implemented, is imperfect and will generally overestimate the ULN when $\mu_D > \mu_H$ (or, conversely, underestimate the LLN when $\mu_D < \mu_H$), its estimates are fairly accurate provided $p_H \gtrsim 0.7$ as shown in Figure 5. Moreover, these biases tend to be both modest and predictable in nature. In contrast, with the incorrect variant of the method, the limits of the normal range may be underestimated or overestimated depending on the specifics of the proportions, means, and variances of the healthy and diseased modes (as illustrated in the Supplemental Figures, Supplemental Appendix 1, and Figure 5).

In general, our algebraic results and simulations (Figure 5) demonstrate that use of the incorrect variant of

▊**Table 1**▊

Results of Mixture Model to Clinical Data for Three Representative Scenarios: Well-Resolved Populations (Male Outpatient Hb), Skewed and Multimodal (Adult TSH), and Poorly Resolved (Total Calcium)

Analyte	Method	μ_H	σ_H	μ_D	σ_D	p_H	LLN	ULN
Hb	Hoffmann-QQ	13.90	1.40	10.6	1.5		11.10	16.70
Hb	Bhattacharya	14.50	1.20	9.3	1.3	0.7	12.10	16.90
Hb	ML-mixtools	14.50	1.30	10.0	1.6	0.63	12.00	17.10
Hb	ML-mixdist	14.50	1.30	10.0	1.6	0.63	11.90	17.00
Hb	Hoffmann-CDF	13.50	1.30				11.00	16.10
TSH	Hoffmann-QQ	1.60	0.80				0.00	3.15
TSH	Bhattacharya	1.22	0.90				−0.54	2.98
TSH	ML-mixtools-normal	1.62	0.91	0.10, 4.85	0.08, 2.70	0.78	−0.15	3.40
TSH	ML-mixdist-gamma	1.98	1.31	0.07, 7.82	0.05, 2.86	0.92	0.29	5.25
TSH	Hoffmann-CDF	1.70	0.70				0.36	3.04
Ca	Hoffmann-QQ	9.10	0.70				7.80	10.35
Ca	Bhattacharya	9.13	0.60				7.96	10.30
Ca	ML-mixtools-normal	9.14	0.62	7.79	0.45	0.90	7.93	10.35
Ca	ML-mixdist-normal	9.15	0.62	7.79	0.45	0.90	7.94	10.35
Ca	ML-mixdist-gamma	9.21	0.58	8.00	0.53	0.83	8.12	10.38
Ca	Hoffmann-CDF	9.10	0.40				8.31	9.86

Ca, calcium; CDF, cumulative distribution function; Hb, hemoglobin; LLN, lower limit of normal; ML, maximum likelihood; QQ, quantile-quantile; TSH, thyroid stimulating hormone; ULN, upper limit of normal; μ_D, mean of the diseased population ; μ_H, mean of the healthy population; σ_D, standard deviation of the diseased population; σ_H, standard deviation of the healthy population.

Hoffmann's method will generate reference interval estimates that are too narrow, particularly when p_H is close to 1, the very context in which indirect reference interval strategies are recommended to be constrained.[40,41] This observation has obvious implications for a number of published studies[8,12-17,19,22] and a reference textbook.[11] In addition, there are undoubtedly many unpublished internal studies conducted in clinical laboratories where use of an incorrect method has yielded invalid reference interval estimates.

Some authors, without identifying the use of an incorrect variant, have found fault with the resulting reference intervals and have noted their poor performance in comparison to directly determined values.[42] Other authors have identified the "linear CDF" variant as a deviation from Hoffmann's original method[27,43,44] without explicitly specifying the significance. These results have obvious implications for strategies purported to improve the Hoffmann method[14,15] that have used the incorrect variant as a starting point; the proposed refinements have the practical effect of making the flawed variant produce wider reference interval estimates, as Hoffmann's method would have done.

It must be acknowledged, though, that even when correctly performed, the Hoffmann and other indirect a posteriori methods are imperfect. The International Federation of Clinical Chemistry and Laboratory Medicine Committee on Reference Intervals and Decision Limits has rightly criticized these methods on the basis of their assumption of normality[41] and the errors that inevitably ensue. Likewise, the Clinical Laboratory Standards Institute advises that indirect methods are, at best, tools for rough estimation.[40]

Even when the Gaussian assumption is met, the random simulations we present show that the graphical methods of Hoffmann and Bhattacharya did not perform as well as ML estimates. ML estimates, while sophisticated in their methodology, have been implemented in R and other languages in a manner that is reasonably easy to use. These packages are also capable of simultaneous fitting of multiple modes and in the case of mixdist,[32] there is no need to assume the underlying distributions are normal. This affords the fitting of skewed data without application of a normalizing transformation such as Box Cox[45] (Supplemental Appendix 2).

However, neither the traditional graphical methods nor ML methods represent a panacea for decomposition of mixture distributions. As with any decomposition method, fitted results may not be meaningful from a physiological standpoint. For example, the fitted diseased mode may paradoxically extend well into the range of the healthy or the reference interval estimate may deviate from those established by traditional means. The establishment of a fit that successfully converges and makes clinical sense may require exclusion of extreme outliers or fixing certain parameters as constant (eg, μ_D and/or σ_D) resulting in a solution that is at least, in part, heuristic.

This creates the risk that one may introduce arbitrary constraints until one finds what one expects to find. These kinds of trimming assumptions were required to achieve convergent and clinically meaningful fits for both TSH (TSH > 15 mIU/L were excluded) and calcium (results less than the 1st and greater than the 99th percentile of the raw data were excluded), as discussed in Supplemental Appendix 2.

It should be noted that ML fit of Gaussian mixture models has been proposed for the problem of "data mining" reference intervals previously.[46] However, it is probable that most laboratorians would find the necessary computations intimidating. This motivated our use and provision of open-source R code, which can be applied to mixtures comprised of normal or skewed distributions (see Supplemental Data). It should also be mentioned that other computational strategies employing ML to fit the healthy mode to a truncated normal distribution have been previously described and implemented using custom R code.[47]

Irrespective of the decomposition method, and even when (1) skewing is appropriately accounted for, (2) mixture model decomposition is successful, and (3) fitted model is good, the reference intervals obtained with these procedures may not match what is found in traditional reference interval studies performed on healthy populations. For example, all the decomposition methods assessed yielded lower limits of normal for outpatient male Hb of approximately 11 to 12 mg/dL (110-120 g/L), that is lower than values obtained from healthy populations, which are typically approximately 13.5 to 17.0 g/dL (135 to 170 g/L).[48] Likewise, the lower limit of normal for uncorrected plasma calcium was also lower than typically reported at approximately 8.0 mg/dL (2.0 mmol/L), which is right on the cusp of levels at which severe symptomatology may appear.[49]

The validity of indirect reference interval estimation is a matter of debate.[43,50] On the basis of our own observations, we believe the recommendation of the Clinical and Laboratory Standards Institute EP28-A3c[40] is prudent: indirectly calculated reference intervals derived from routine analyses should be used cautiously. What *is* certain is that if one is to use Hoffmann's method, it should be undertaken as originally described using a normal QQ plot because use of a linear CDF generates inaccurate and possibly erratic parameter estimates of the healthy mode. However, because more flexible and accurate contemporary computational methods for mixture decomposition are freely available,[32-36] some of which can be performed free of a Gaussian assumption, it may mean that purely graphical procedures like Hoffmann's have had their day.

Conclusion

The Hoffmann method is frequently applied in a manner divergent from the original description. For distributions satisfying the assumptions of the Hoffmann method, the error of using a CDF plot in linear space rather than a normal QQ plot (equivalent to a normal probability plot) typically leads to reference intervals that are too narrow. The behavior of this erroneous approach is generally dependent on the specifics of the distributions of diseased and healthy individuals and may produce reference intervals that are too wide in select circumstances. Among the methods evaluated (Hoffmann using a QQ plot, Hoffmann using a CDF, Bhattacharya, and ML), ML most consistently recovered the correct values from random number simulations and, depending on the tool employed, has the added benefit of being able to fit skewed distributions without the use of normalizing transformations.

Corresponding author: Daniel T. Holmes, MD; dtholmes@mail.ubc.ca.

References

1. Hoffmann RG. Statistics in the practice of medicine. *JAMA.* 1963;185:864-873.

2. Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis of data. *Biometrika.* 1968;55:1-17.

3. Murthy JN, Hicks JM, Soldin SJ. Evaluation of the Technicon Immuno I random access immunoassay analyzer and calculation of pediatric reference ranges for endocrine tests, T-uptake, and ferritin. *Clin Biochem.* 1995;28:181-185.

4. Krafte-Jacobs B, Williams J, Soldin SJ. Plasma erythropoietin reference ranges in children. *J Pediatr.* 1995;126:601-603.

5. Soldin SJ, Morales A, Albalos F, et al. Pediatric reference ranges on the Abbott IMX for FSH, LH, prolactin, TSH, T4, T3, free T4, free T3, T-uptake, IGE, and ferritin. *Clin Biochem.* 1995;28:603-606.

6. Soldin OP, Miller M, Soldin SJ. Pediatric reference ranges for zinc protoporphyrin. *Clin Biochem.* 2003;36:21-25.

7. Soldin OP, Hanak B, Soldin SJ. Blood lead concentrations in children: new ranges. *Clin Chim Acta.* 2003;327:109-113.

8. Soldin OP, Bierbower LH, Choi JJ, et al. Serum iron, ferritin, transferrin, total iron binding capacity, hs-CRP, LDL cholesterol and magnesium in children; new reference intervals using the DADE dimension clinical chemistry system. *Clin Chim Acta.* 2004;342:211-217.

9. Soldin OP, Hoffman EG, Waring MA, et al. Pediatric reference intervals for FSH, LH, estradiol, T3, free T3, cortisol, and growth hormone on the DPC IMMULITE 1000. *Clin Chim Acta.* 2005;355:205-210.

10. Soldin SJ, Soldin OP, Boyajian AJ, et al. Pediatric brain natriuretic peptide and N-terminal pro-brain natriuretic peptide reference intervals. *Clin Chim Acta.* 2006;366:304-308.

11. Soldin S, Brugnara C, Wong E. *Pediatric Reference Intervals.* 7th ed. Washington, DC: AACC Press; 2011.

12. Soldin OP, Dahlin JR, Gresham EG, et al. IMMULITE 2000 age and sex-specific reference intervals for alpha fetoprotein, homocysteine, insulin, insulin-like growth factor-1, insulin-like growth factor binding protein-3, C-peptide, immunoglobulin E and intact parathyroid hormone. *Clin Biochem.* 2008;41:937-942.

13. Soldin OP, Sharma H, Husted L, et al. Pediatric reference intervals for aldosterone, 17alpha-hydroxyprogesterone, dehydroepiandrosterone, testosterone and 25-hydroxy vitamin D3 using tandem mass spectrometry. *Clin Biochem.* 2009;42:823-827.

14. Katayev A, Balciza C, Seccombe DW. Establishing reference intervals for clinical laboratory test results: is there a better way? *Am J Clin Pathol.* 2010;133:180-186.

15. Katayev A, Fleming JK, Luo D, et al. Reference intervals data mining: no longer a probability paper method. *Am J Clin Pathol.* 2015;143:134-142.

16. Grecu DS, Paulescu E. Quality in post-analytical phase: indirect reference intervals for erythrocyte parameters of neonates. *Clin Biochem.* 2013;46:617-621.

17. Feng Y, Bian W, Mu C, et al. Establish and verify TSH reference intervals using optimized statistical method by analyzing laboratory-stored data. *J Endocrinol Invest.* 2014;37:277-284.

18. Shaw JL, Cohen A, Konforte D, et al. Validity of establishing pediatric reference intervals based on hospital patient data: a comparison of the modified Hoffmann approach to CALIPER reference intervals obtained in healthy children. *Clin Biochem.* 2014;47:166-172.

19. Hackenmueller SA, Grenache DG. Reference intervals for intestinal disaccharidase activities determined from a non-reference population. *J Appl Lab Med.* 2016;1:172-180.

20. Strich D, Karavani G, Levin S, et al. Normal limits for serum thyrotropin vary greatly depending on method. *Clin Endocrinol (Oxf).* 2016;85:110-115.

21. Clark ZD, Cutler JM, Pavlov IY, et al. Simple dilute-and-shoot method for urinary vanillylmandelic acid and homovanillic acid by liquid chromatography tandem mass spectrometry. *Clin Chim Acta.* 2017;468:201-208.

22. Han L, Zheng W, Zhai Y, et al. Reference intervals of trimester-specific thyroid stimulating hormone and free thyroxine in Chinese women established by experimental and statistical methods. *J Clin Lab Anal.* 2018;32:e22344.

23. Amador E, Hsi BP. Indirect methods for estimating the normal range. *Am J Clin Pathol.* 1969;52:538-546.

24. Neumann GJ. The determination of normal ranges from routine laboratory data. *Clin Chem.* 1968;14:979-988.

25. Gindler EM. Calculation of normal ranges by methods used for resolution of overlapping Gaussian distributions. *Clin Chem.* 1970;16:124-128.

26. Bolann BJ. Easy verification of clinical chemistry reference intervals. *Clin Chem Lab Med.* 2013;51:e279-e281.

27. Søeby K, Jensen PB, Werge T, et al. Mining of hospital laboratory information systems: a model study defining age- and gender-specific reference intervals and trajectories for plasma creatinine in a pediatric population. *Clin Chem Lab Med.* 2015;53:1621-1630.

28. Redner RA, Walker HF. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review.* 1984;26:195-239.

29. Bhattacharya C. A simple method of resolution of a distribution into Gaussian components. *Biometrics.* 1967;23:115-135.

30. Jones GR, Haeckel R, Loh TP, et al. Indirect methods for reference interval determination–review and recommendations. *Clin Chem Lab Med.* [published online ahead of print April 19, 2018]. doi: 10.1515/cclm-2018-0073.

31. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Series B Stat Methodol.* 1977;39:1-38.

32. Macdonald P, Juan DU. *Mixdist: Finite Mixture Distribution Models.* https://CRAN.R-project.org/package=mixdist. Accessed October 11, 2018.

33. Benaglia T, Chauveau D, Hunter DR, et al. mixtools: an R package for analyzing finite mixture models. *J Stat Softw.* 2009;32:1-29.

34. Leisch F. FlexMix: a general framework for finite mixture models and latent class regression in R. *J Stat Softw.* 2004;11:1-18.

35. Scrucca L, Fop M, Murphy TB, et al. Mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *R J.* 2016;8:289-317.

36. Lebret R, Iovleff S, Langrognet F, et al. Rmixmod: the R package of the model-based unsupervised, supervised and semi-supervised classification mixmod library. *J Stat Softw.* 2015;67:241-270.

37. Krishnamoorthy K. *Handbook of Statistical Distributions With Applications.* Boca Raton, FL: Chapman and Hall/CRC; 2006.

38. Baggerly KA, Coombes KR. Deriving chemosensitivity from cell lines: forensic bioinformatics and reproducible research in high-throughput biology. *Ann Appl Stat.* 2009;3:1309-1334.

39. Allaire J, Cheng J, Xie Y, et al. *Rmarkdown: Dynamic Documents for R.* https://CRAN.R-project.org/package=rmarkdown. Accessed October 11, 2018.

40. Clinical and Laboratory Standards Institute (CLSI). *Defining, Establishing, and Verifying Reference Intervals in the Clinical Laboratory; Approved Guideline.* 3rd ed. CLSI document EO28-A3c. Wayne, PA: CLSI; 2010:28.

41. Ichihara K, Boyd JC; IFCC Committee on Reference Intervals and Decision Limits (C-RIDL). An appraisal of statistical procedures used in derivation of reference intervals. *Clin Chem Lab Med.* 2010;48:1537-1551.

42. Roberts WL, Rockwood AL, Bunker AM, et al. Limitations of the Hoffman approach to determine pediatric reference intervals for two steroids. *Clin Biochem.* 2010;43:933-934; author reply 935.

43. Horowitz GL. Estimating reference intervals. *Am J Clin Pathol.* 2010;133:175-177.

44. Jones G, Horowitz G, Katayev A, et al. Reference intervals data mining: getting the right paper. *Am J Clin Pathol.* 2015;144:526-527.

45. Box GE, Cox DR. An analysis of transformations. *J R Stat Soc B Stat Methodol.* 1964;26:211-252.

46. Concordet D, Geffré A, Braun JP, et al. A new approach for the determination of reference intervals from hospital-based data. *Clin Chim Acta.* 2009;405:43-48.

47. Arzideh F, Wosniok W, Gurr E, et al. A plea for intra-laboratory reference limits. Part 2. A bimodal retrospective concept for determining reference limits from intra-laboratory databases demonstrated by catalytic activity concentrations of enzymes. *Clin Chem Lab Med.* 2007;45:1043-1057.

48. Adeli K, Ceriotti F, Nieuwesteeg M. Reference information for the clinical laboratory. In: Rifai N, Horvath A, eds. *Tietz Textbook of Clinical Chemistry and Molecular Diagnostics.* Saint Louis, MO: Elsevier; 2017:1745-1818.

49. Cooper MS, Gittoes NJ. Diagnosis and management of hypocalcaemia. *BMJ.* 2008;336:1298-1302.

50. Dorizzi RM, Giannone G, Cambiaso P, et al. Indirect methods for TSH reference interval: at last fit for purpose? *Am J Clin Pathol.* 2011;135:167-168.

HOLOGIC®

First and Only FDA Cleared
Digital Cytology System

Genius™ Cervical AI

Genius™ Review Station

Genius™ Digital Imager

Empower Your Genius With Ours

Make a Greater Impact on Cervical Cancer

with the Advanced Technology of the
Genius™ Digital Diagnostics System

Click or Scan
to discover more

_genius™
DIGITAL DIAGNOSTICS