A Model for the Effect of Bias for Cholesterol on the Population at Risk

Martin H. Kroll, Mark Ruddel, and Ronald J. Ellen

The location of the Reference Value for an analyte within the population distribution affects the magnitude of error due to methodological bias. Using the gaussian distribution, we evaluated the effects of systematic and proportional biases of the method (positive and negative), mean value, and standard deviation on the magnitude of error. We chose four Reference Values for cholesterol as a model. For a population with a mean of 2.0 and SD of 0.36 g of cholesterol per liter, a 3% positive proportional bias causes sixfold more error at the 50th percentile than at the 97.5th. In general, the error for a given bias (proportional or systematic) is greater for a Reference Value within the body than at the tails of the distribution. Further, the magnitude of the error varies as a function of the mean and standard deviation of the population.

The National Institutes of Health's Consensus Conference on lowering cholesterol (1) selected Reference Values for cholesterol based on the probability of developing coronary heart disease rather than values near the 2.5th and 97.5th percentiles (the tails) of the population distribution. These Reference Values for cholesterol are near the 50th and 75th percentiles, which are within the body of the distribution. Furthermore, many of the enzymatic methods show a methodological bias compared with the Reference Methods (2). The location of the Reference Value in the distribution affects the number of patients misdiagnosed (the error) owing to methodological bias.

The population distribution for cholesterol approximates a gaussian distribution sufficiently to allow the gaussian distribution equation to be used as a model. By moving the Reference Values toward the mode of the distribution, the error (the population at risk) for a given bias is increased. Here we quantify this relationship and extend our observations beyond cholesterol to the general case.

Methods

The percentage of the population deemed at risk for coronary heart disease is the area under the curve between two Reference Values. The Reference Values cut the population distribution near the 75th and 50th percentiles. For our model, we used the gaussian normal distribution curve, so that the area under the curve is given by the integral.

We used the following equation to determine the error (percentage misdiagnosed):

\[
\text{Error} = 100\% \left( \int_{R_2}^{R_1} \left( \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{(x - m - b)^2}{2\sigma^2} \right) \right) - \int_{R_1}^{R_2} \left( \frac{1}{\sqrt{2\pi} \sigma} \exp \left( -\frac{(x - m)^2}{2\sigma^2} \right) \right) \right)
\]

where \( \sigma \) is the SD of the distribution, \( m \) is the mean, \( R_1 \) and \( R_2 \) are the Reference Values, \( x \) is the concentration of analyte, \( b \) is the systematic bias expressed in the same units as the mean, and \( p \) is the proportional bias. For a purely systematic bias, \( p = 1.00 \), and for a purely proportional bias, \( b = 0 \). A 5% positive bias is \( p = 1.05 \) and a 5% negative bias is \( p = 0.95 \).

In general, to calculate the percentage of the population above a Reference Value, \( R_1 \), is the Reference Value and \( R_2 \) is infinity. Between two Reference Values, \( R_1 \) is the lower and \( R_2 \) the higher value. To calculate the percentage of the population below a Reference Value, \( R_2 \) is the Reference Value and \( R_1 \) is minus infinity.

We examined two types of bias: systematic and proportional. Even though not all biases are purely one or the other, analyses based on the isolated types will give a good estimate of the combination. For the systematic biases, we used 0.06 and 0.10 g of cholesterol per liter. For the proportional biases, we used 3% and 5%. The effects of bias on error in the population at risk were examined for both positive and negative biases at the 50th, 75th, 90th, and 97.5th percentiles for a population with a mean of 2.0 and an SD of 0.36 g cholesterol per liter. The proportional and systematic biases are equal, respectively, at the 50th percentiles, because of the way they were chosen.

We calculated the error as generated by varying bias, both proportional and systematic, over all percentiles. Also, we calculated the error caused by varying the mean and standard deviation at fixed values for bias. In our generalized studies of the effects of the different variables, we have varied the values of one variable and kept the rest fixed.

Thus, when bias was varied, the mean was kept at 2.0 and the SD at 0.2, and when the mean or SD was varied, the proportional bias was kept at ±5% or the systematic bias was kept at ±0.10.

Results and Discussion

The location of the Reference Value in the distribution of the population greatly affects the magnitude of the population considered at risk for both systematic and proportional biases. The percentage of the population misdiagnosed as being at risk for coronary heart disease for one subpopulation is presented in Table 1. The percentage of the population misdiagnosed is rather small for the 97.5th percentile, ranging from 0.8% to 1.5% for positive bias and from 1.2% to 3.2% for negative bias. At the 50th percentile, the error due to bias may be directly compared, because the values selected to represent systematic biases (0.06 and 0.10) are the same as those for the proportional ones (3% and 5%, respectively; i.e., 3% of 2.0 is 0.06 and 5% of 2.0 is 0.10). Proportional bias causes a slightly greater error in the population at risk than does systematic bias at the 75th, 90th, and 97.5th percentiles, and negative bias causes a slightly greater error than does positive bias at these same percentiles (Table 1). More important, as the Reference Value moves away from the 97.5th percentile, toward the interior of the distribution, the error becomes larger. To aid in the comparison of the magnitude of the errors in the
population at risk, we have presented the normalized values of the error in parentheses. In each category, the error at the 97.5th percentile was set at 1. At the 75th or 50th percentile, the relative error is 5- to 5.5-fold greater for positive bias, and 3.6- to 6.5-fold greater for negative bias, than at the 97.5th percentile (Table 1). Thus, the error in population at risk due to methodological bias for a Reference Value depends on the location of the value in the population distribution.

So far, we have addressed the effect of methodological bias on the population at risk and used cholesterol values from one specific subpopulation to illustrate the relationship. With a different subpopulation, the mean and SD about the mean would change, and thus alter the calculated errors for each percentile. By using Equation 1, the problem can be generalized to any population that could be modeled with the gaussian distribution. Currently, the Reference Values for the interpretation of cholesterol demonstrate this problem, but the problem could apply to any analyte. One can calculate the error from Equation 1 by inserting the proper values. We show the effects of varying the bias, mean, and SD on error in Figures 1 and 2.

For systematic bias (Figure 1), varying the bias or SD generates families of curves that are centrally symmetrical with respect to the positive and negative biases. The maximum percent error of the population at risk is a function of the methodological bias and mean (proportional bias only) and standard deviation of the population. This value is shown as a vertical line in the upper row and is plotted on the ordinate in the lower row of Figures 1 and 2. The plot of the maximum of the percentage error (y) and the bias (z) has a sigmoid shape; the curve can be described as $y = 100 \tanh (2.08z)$, where tanh is the hyperbolic tangent, and the region between $-0.03$ and $0.03$ could be adequately described with a straight line, $y = 19z$. The percent error varies inversely with the SD. A plot of the maximum percentage error (y) vs the SD (z) shows two hyperbolas, in the first and fourth quadrants. They can be represented as $xy = c$, where x is the SD, y is the maximum percent error, and $c = 9.57$ for positive bias and $-9.57$ for negative bias. Varying the mean has no effect on the error for systematic bias.

For proportional bias, varying the bias, mean, or SD generates families of curves that are centrally symmetric (Figure 2). The magnitude of the percentage error due to bias is directly proportional to the magnitude of the bias. A plot of the maximum percent error (y) vs bias (z) is nearly linear between $-10$ to $10\%$ bias, with slopes of 3.6 and 4.0 for positive and negative biases, respectively (Figure 2). As with systematic bias, this relationship could be expressed as $y = 100 \tanh (cz)$, with $c = 4.204$. The magnitude of the percentage error due to the mean is directly proportional to

### Table 1. Percentage of the Population Misdiagnosed (Error) Owing to Bias

<table>
<thead>
<tr>
<th>Percentile</th>
<th>Cholesterol, g/L</th>
<th>Proportional 3%</th>
<th>Proportional 5%</th>
<th>Systematic 0.06</th>
<th>Systematic 0.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>50</td>
<td>2.0</td>
<td>6.7(6.1)</td>
<td>11.0(7.4)</td>
<td>6.7(6.4)</td>
<td>11.0(8.5)</td>
</tr>
<tr>
<td>75</td>
<td>2.2</td>
<td>5.5(5.0)</td>
<td>8.9(5.9)</td>
<td>5.0(6.3)</td>
<td>8.0(6.2)</td>
</tr>
<tr>
<td>90</td>
<td>2.5</td>
<td>3.2(2.9)</td>
<td>4.7(3.1)</td>
<td>2.7(3.4)</td>
<td>4.1(3.2)</td>
</tr>
<tr>
<td>97.5</td>
<td>2.7</td>
<td>1.1(1.0)</td>
<td>1.5(1.0)</td>
<td>0.8(1.0)</td>
<td>1.3(1.0)</td>
</tr>
</tbody>
</table>

Negative   |                  |                 |                |                 |                 |
| 50         | 2.0              | 6.4(3.8)        | 11.0(3.4)      | 6.4(5.3)        | 11.0(5.2)       |
| 75         | 2.2              | 6.2(3.6)        | 10.6(3.3)      | 5.5(4.6)        | 9.6(4.6)        |
| 90         | 2.5              | 4.2(2.5)        | 7.4(2.3)       | 3.6(3.0)        | 5.9(2.8)        |
| 97.5       | 2.7              | 1.7(1.0)        | 3.2(1.0)       | 1.2(1.0)        | 2.1(1.0)        |

*Mean = 2.0 g/L, SD = 0.36 g/L, for men 35 to 39 y old.

*The relative magnitude compared with the 97.5th percentile is shown in parentheses.

---

**Fig. 1. Error due to systematic bias**

The upper plots show the percentage error in the population due to varying the systematic bias of the method, the mean, or the standard deviation (SD) as a function of the percentile of the population (x-axis). For SD, the curves are above zero for positive bias and below zero for negative bias. The numbers in the curves represent the values of the bias or SD. The maximum percent error for a particular value of bias or SD is indicated by a small vertical tic-mark. The lower row plots the maximum percent error (y-axis) as a function of the bias, mean, or SD (x-axis). The mean = 2.0; when the bias is varied, the SD = 2.0; and when the SD is varied, the bias = ± 0.1
the magnitude of the mean (Figure 2), in contradistinction to systematic bias, where there was no change in error as the mean was varied. A plot of the maximum percent error (y) vs the mean (x) shows two curved lines, the magnitudes of which are directly proportional to the magnitudes of the mean (Figure 2). These relationships can be expressed as straight lines for a mean <7, y = cx, where c = 9.4 and 9.8 for positive and negative biases, respectively. The percentage error changes as the SD varies in a fashion similar to that of systematic bias, but because the mean is a variable for proportional bias, the changes in error observed with the SD or the mean are not independent and could be combined as the CV (SD/mean) for the distribution (Figure 2). The magnitude of this error is inversely proportional to the magnitude of the CV. A plot of the maximum percentage error vs the SD shows two hyperbolas, present in the first and fourth quadrants (Figure 2).

One can understand these effects by examining Equation 1 or picturing the gaussian distribution. For systematic bias, the biased method effectively shifts the entire curve, maintaining its shape without distortion. Of course, the absolute value for the bias increases, so does the magnitude of the absolute error. Changing the mean does not affect the relative distance between the biased and unbiased curves, which is given only by the systematic bias itself, so the error in this type of case is constant. The SD of the distribution is a measure of how wide the distribution appears. The wider the distribution curve, the flatter the slope, which then lessens the effect of the bias.

For proportional bias, the effect of varying the SD on the error is analogous to that for systematic bias. The case is different when the bias is varied. Bias does not shift the curve uniformly, but distorts it, because the shift is a percentage of the value for x. Thus, higher values are shifted more than lower ones. The effect of proportional bias on the error still is closely analogous to that found for systematic bias, but varying the mean does have an effect on the error. The higher the mean, the greater the error due to proportional bias for a fixed SD. If the SD is varied concomitantly with the mean (SD/mean = constant), the error is constant. Thus, in applying Equation 1, one must be aware of how much of the bias is systematic and how much proportional (Table 2). In general, these curves (Figures 1 and 2) show that the error is greatly increased when the Reference Value is in the body instead of the tails of the distribution.

Changes in the Reference Values cause a change in the error induced by a given bias, because the gaussian distribution is bell-shaped. The error is calculated from the area under the curve between two designated points on the x-axis (R1 = x1 and R2 = x2). The height of the distribution (y|x) specifies the degree of change for an inaccuracy in the value of the x-axis. In the tails of the distribution, which are the regions in the vicinity of the 2.5th and 97.5th percentiles, the height of the curve is low compared with the body of the distribution. The height is greatest at the 50th percentile, and still relatively high at the 25th and 75th percentiles. From Table 1, one can obtain a general idea how the error relates to the bias. At the 97.5th percentile, the error equals approximately one-third the percentage bias. At the 75th

Table 2. Summary of the Effects of Variables on Absolute Value of Error

<table>
<thead>
<tr>
<th>Variable</th>
<th>Systematic</th>
<th>Proportional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bias</td>
<td>Increase</td>
<td>Increase</td>
</tr>
<tr>
<td>Mean</td>
<td>No effect</td>
<td>Increase</td>
</tr>
<tr>
<td>SD</td>
<td>Decrease</td>
<td>Decrease</td>
</tr>
<tr>
<td>SD/bias</td>
<td>Decrease</td>
<td>Decrease</td>
</tr>
<tr>
<td>CV</td>
<td>Decrease</td>
<td>Decrease</td>
</tr>
</tbody>
</table>
and 50th percentiles the error is greater than the percentage bias. From the figure, it is apparent that the percentile where the maximum error occurs shifts toward, but never reaches, the tails of the distribution as the error due to the bias or mean (proportional bias only) increases or as the SD decreases. It is inappropriate to apply criteria obtained in the tails of the distribution to the body of the distribution.

A major departure from the usual approach to assigning Reference Values was made by the National Institutes of Health's Consensus Conference on coronary heart disease and cholesterol (1). The Conference advised that Reference Values based on the probability of developing coronary heart disease be used. The first set of recommendations were age-dependent and resulted in the Reference Values being near the 90th and 75th percentiles (1). The second set of recommendations were age-independent and resulted in Reference Values near the 75th and 50th percentiles (3). Allowable bias has been suggested as 5% now and 3% in five years (4). The crossing of the new Reference Values and the bias in the methods for cholesterol breed a new problem, which demands a more careful assessment than that for the more common modes of interpretation.

The use of Reference Values in the body of the distribution as opposed to the tails may not be unique for cholesterol. The situation for cholesterol arose out of the desire to base treatment on the probability of mortality from coronary heart disease (1, 5–7), which is an interpretation based on the outcome instead of the arbitrary cutoffs at the 2.5th or 97.5th percentiles.

One must be careful in assigning allowable bias for any test where the Reference Values fall in the body instead of the tail of the distribution. The assignment of allowable bias should be based on the permissible number of people that can be misdiagnosed. The error due to bias is dependent on the type of bias (systematic, proportional, positive or negative), the Reference Value, and the mean and SD of the density distribution; all these factors must be considered.

References