The spurious correlation between concentration and creatinine-corrected concentration in urine

Simon Brown, Delwyn G. Cooke, Leonard F. Blackwell and David C. Simcock

Abstract—The use of urinary analytes to monitor physiological processes relies on making the correct measurement. Three alternatives are commonly contemplated: concentration, creatinine-corrected concentration and excretion rate. Of these, the latter is the most reliable, but is perceived by some to be difficult to measure. This has led to the more frequent reliance on concentration and one of the justifications for this is the reported linear relationship between the concentration and the creatinine-corrected concentration. We show that this correlation is spurious in that the magnitude of the correlation coefficient depends on the ratio of the standard deviations of the creatinine and analyte concentrations. As an example urinary analyte we use pregnanediol (Pd) which is an important tool for women wishing to monitor their own fertility. Urinary Pd concentration is not a reliable substitute for creatinine-corrected Pd concentration or the Pd excretion rate.

Keywords—creatinine correction, menstrual cycle, spurious correlation, urinary analyte concentration.

I. INTRODUCTION

It is fundamental to the meaningful use of any data that the measurement on which they are based is reliable, appropriate and free from confounding factors. However, there are instances where data quantity is taken to be a reasonable substitute for data quality. Moreover, the continued use of a measurement known to be defective is sometimes justified by an argument that it is ‘difficult’ or ‘inconvenient’ to do a better measurement. Problems of this sort are widespread, but are especially common in measurements of urinary analytes, including those involved in the measurement of reproductive hormones in urine by women monitoring their own fertility.

The quantity of an analyte (A) in a urine sample has been expressed in many ways, including (i) concentration ([A]), (ii) [A] normalised by the creatinine (Cr) concentration ([A]/[Cr]) and (iii) the excretion rate ($J_A$). We have shown that $J_A$ can be related to the rate of production of A and that [A] is, at best, a poor estimate of $J_A$ [1]. This is because the variability of both the volume of urine accumulated ($V$ in mL) and the time between voids ($\Delta t$ in h) means that [A] changes between voids. Variation of this sort results from environmental and lifestyle factors. The most direct measure of the physiological urinary output of an analyte is its excretion rate ($J_A$, in g h$^{-1}$ or mol h$^{-1}$) which, as we have outlined previously [2], is the product of [A] and the urine production rate ($J_V$, in mL h$^{-1}$)

\[
J_A = \frac{[A]}{[Cr]} \frac{\theta_A}{V} \frac{V}{\Delta t} = \frac{\theta_A}{\Delta t} 
\]

where $\theta_A$ is the quantity (in mol or g) of A in the void. An alternative measure that is often used is the ratio of [A] to the concentration of creatinine (Cr). It is widely assumed that Cr is excreted at a constant rate [3-5]. This approach follows from (1): if $J_C = [Cr]J_V$ is constant, then $J_V \propto 1/[Cr]$ and

\[
J_A \propto \frac{[A]}{[Cr]},
\]

which is the basis of the widespread use of [Cr] to ‘correct’ for $J_V$.

However, the perception that it is difficult to measure $J_V$ and the desire to avoid determining [Cr] have motivated many to assume that concentration ([A]) is a reasonable means of monitoring a urinary analyte. Two recent ‘justifications’ for this are that plots of (a) $\ln([A])$ versus $\ln(J_A)$ [6] and (b) $\ln([A])$ versus $\ln([A]/[Cr])$ [7] are ‘linear’. In both of these cases [6, 7], the urinary analyte (A) is pregnanediol-3-glucuronide (PdG) which is a metabolite of the reproductive hormone progesterone, although Roos et al. [7] also applied this analysis to oestrone-3-glucuronide (E1G) which is a metabolite of the reproductive hormone oestradiol. The combination of $J_{PdG}$ and $J_{E1G}$ provides a powerful means of monitoring the menstrual cycle and fertility [8, 9]. However, there is a recent trend, based in part on these ‘linear’ plots, towards a reliance on [PdG] and [E1G] [7, 10-14], despite the very substantial literature based on excretion rates [15-43].

The notion that this sort of analysis provides some support for the idea that [PdG] might be a reasonable substitute for $J_{PdG}$ [6] or even [PdG]/[Cr] [7] has prompted us to examine the evidence. We do so using some numerical experiments and also using measurements of the urinary concentration of pregnanediol (Pd, which is obtained by hydrolysis of PdG), [Cr] and $J_V$.

II. BACKGROUND

To examine the ‘linearity’ of $\ln(y)$ versus $\ln(y/x)$ and $\ln(y)$...
versus ln(x/y) we summarise both by writing them as ln(y) versus ln(g(x, y)), where g(x, y) = x/y and γ = ±1, although the analysis is not restricted to these values of γ. A linear relationship of this type would imply

\[ \ln(y) = \beta_0 + \beta_1 \ln(g(x, y)) = \beta_0 + \beta_1 \ln(y) + \gamma \beta_1 \ln(x), \]

(3)

from which it is clear that if x = 1, then β0 = 0 and β1 = 1. The ordinary least squares (OLS) estimates of β0 and β1 are

\[ \hat{\beta}_0 = \langle \ln(y) \rangle - \hat{\beta}_1 \langle \ln(g(x, y)) \rangle \]

\[ \hat{\beta}_1 = R_y - \frac{s_{\ln(y)} s_{\ln(g(x, y))}}{s_{\ln(g(x, y))}}, \]

(4)

where \( <z> \) and \( s_z \) are the sample mean and sample standard deviation of z, respectively [44]. In (4) \( R_y \) is the correlation coefficient between ln(y) and ln(g(x, y))

\[ R_y = \frac{\text{cov}(\ln(y), \ln(g(x, y)))}{s_{\ln(y)} s_{\ln(g(x, y))}} \]

(5)

where \( \text{cov}(w, z) \) is the covariance of w and z. The mean and standard deviation of ln(g(x, y)) are

\[ \langle \ln(g(x, y)) \rangle = \langle \ln(y) \rangle + \gamma \langle \ln(x) \rangle \]

(6)

and

\[ s_{\ln(g(x, y))}^2 = \gamma^2 s_{\ln(x)}^2 + s_{\ln(y)}^2 + 2\gamma \text{cov}(\ln(x), \ln(y)), \]

(7)

respectively, and

\[ \text{cov}(\ln(y), \ln(g(x, y))) = s_{\ln(y)} s_{\ln(g(x, y))} + \gamma \text{cov}(\ln(x), \ln(y)). \]

(8)

Substituting (7) and (8) into (5) yields

\[ R_y = \frac{1 + \gamma R_x}{\sqrt{1 + \gamma^2 + 2\gamma R_x}}, \]

(9)

where γ = \( s_{\ln(g(x, y))} / s_{\ln(y)} \) > 0 and we have written the correlation coefficient between ln(x) and ln(y) as

\[ R_x = \text{cov}(\ln(x), \ln(y)) / s_{\ln(x)} s_{\ln(y)}. \]

Substituting (6) into (4) yields

\[ \hat{\beta}_0 = (1 - \hat{\beta}_1) \langle \ln(y) \rangle - \gamma \hat{\beta}_1 \langle \ln(x) \rangle \]

(10)

and, using (4), (5), (7) and (8) gives

\[ \hat{\beta}_1 = \frac{\gamma R_x \lambda + 1}{\gamma^2 \lambda^2 + 2\gamma R_x \lambda + 1}. \]

(11)

In general (i) \( \hat{\beta}_1 \) tends to decline with increasing \( \lambda \), although the behaviour is more complex for \( \gamma = -1 \) if \( R_x \) is large (Figure 1A), (ii) if \( \lambda \) is small \( \hat{\beta}_1 \approx 1 \) and if \( \lambda \) is large \( \hat{\beta}_1 \approx 0 \) (Figure 1A), (iii) if \( \lambda \) is small \( R_x \approx 1 \) and if \( \lambda \) is large \( R_x \) is smaller and can be negative depending on \( \gamma \) and \( R_x \) (Figure 1B) and (iv) neither \( R_x \) nor \( \hat{\beta}_1 \) depends systematically on \( <\ln(x)> \) or \( <\ln(y)> \). The corollaries are that (i) if \( \lambda \) is small \( \hat{\beta}_0 \) approaches \( -\gamma <\ln(x)> \) and (ii) as \( \lambda \) increases \( \hat{\beta}_0 \) approaches \( <\ln(y)> \).

III. METHODS

Measurements of urinary concentration of pregnanediol (Pd), which is quantitatively derived from PdG [45], and Cr, and of \( J_{PV} \) were obtained from the DIY trial carried out in the late 1980s in Melbourne. The data we analyse here are a subset of these and comprise periovulatory measurements of Pd and Cr for 26 menstrual cycles from 12 subjects, yielding a total of \( n = 190 \) complete records.

IV. RESULTS

A. An example

The relationship between ln([Pd]) and ln([Pd]/[Cr]) is approximately linear (Figure 3A) and OLS regression yields

\[ \hat{\beta}_0 = -0.04 \pm 0.09 \quad [95\% \text{ CI}] \quad \text{and} \quad \hat{\beta}_1 = 0.94 \pm 0.08 \quad [95\% \text{ CI}] \]

\[ (R_0 = 0.855 \quad [95\% \text{ CI: 0.811, 0.889}], \quad p < 0.001). \]

As the independent variable is uncertain, Deming regression might be a more appropriate approach but it yields similar
estimates of the intercept (0.06 ± 0.09 [95% CI]) and slope (1.11 ± 0.09 [95% CI]) assuming a precision ratio of one. In neither case is the slope significantly different from one ($p \geq 0.891$).

Figure 2. Distribution of urinary Pd (A: $<\text{ln}([\text{Pd}])>$ = -0.57 ± 0.07 (SD), $s_{\text{ln}([\text{Pd}])}$ = 1.03 ± 0.05 (SD)) and Cr (B: $<\text{ln}([\text{Cr}])>$ = 0.00 ± 0.04 (SD), $s_{\text{ln}([\text{Cr}])}$ = 0.54 ± 0.03 (SD)) concentration ($n$ = 190). In each panel the curve is the lognormal cumulative distribution function fitted to the data by maximum likelihood. The insets show the corresponding QQ plot in which the straight line indicates equality between the theoretical lognormal and observed quantiles.

For these data, $s_{\text{ln}([\text{Pd}])}$ = 1.03 (Figure 2A), $s_{\text{ln}([\text{Cr}])}$ = 0.54 (Figure 2B), $R_1$ = 0.424 and cov($\text{ln}([\text{Pd}])$, $\text{ln}([\text{Cr}])$) = 0.23, so $\lambda = s_{\text{ln}([\text{Cr}])}/s_{\text{ln}([\text{Pd}])}$ = 0.52 is small and it follows from (9) that no matter the value of $R_1$, the correlation between $\text{ln}([\text{Pd}])$ and $\text{ln}([\text{Cr}])$ is likely to be high (Figure 1B), which is the case ($R_0 = 0.855$). To examine this point, we randomly sampled the [Cr] data without replacement using the sample function in R, so that each [PdG] was ‘corrected’ (2) by a random [Cr] but $n$, $s_{\text{ln}([\text{Pd}])}$, $s_{\text{ln}([\text{Cr}])}$ and $\lambda$ were identical for each iteration. For each of 1000 iterations $R_0$ was calculated and the distribution of these values is shown in Figure 3B. While $R_0$ = 0.855 for the data shown in Figure 3A, the randomised [Cr] values yielded $R_0$ that were all high (Figure 3B, $<R_0>$ = 0.887 ± 0.007 (SD)).

For these data, $s_{\text{ln}([\text{Pd}])}$ = 1.03 (Figure 2A), $s_{\text{ln}([\text{Cr}])}$ = 0.54 (Figure 2B), $R_1$ = 0.424 and cov($\text{ln}([\text{Pd}])$, $\text{ln}([\text{Cr}])$) = 0.23, so $\lambda = s_{\text{ln}([\text{Cr}])}/s_{\text{ln}([\text{Pd}])}$ = 0.52 is small and it follows from (9) that no matter the value of $R_1$, the correlation between $\text{ln}([\text{Pd}])$ and $\text{ln}([\text{Cr}])$ is likely to be high (Figure 1B), which is the case ($R_0 = 0.855$). To examine this point, we randomly sampled the [Cr] data without replacement using the sample function in R, so that each [PdG] was ‘corrected’ (2) by a random [Cr] but $n$, $s_{\text{ln}([\text{Pd}])}$, $s_{\text{ln}([\text{Cr}])}$ and $\lambda$ were identical for each iteration. For each of 1000 iterations $R_0$ was calculated and the distribution of these values is shown in Figure 3B. While $R_0$ = 0.855 for the data shown in Figure 3A, the randomised [Cr] values yielded $R_0$ that were all high (Figure 3B, $<R_0>$ = 0.887 ± 0.007 (SD)).

B. Numerical experiments

To examine the effect of changes to specific parameters we carried out numerical experiments in which $<\text{ln}(y)>$, $s_{\text{ln}(y)}$, $<\text{ln}(x)>$ and $s_{\text{ln}(x)}$ were varied independently. For simplicity, we concentrate on $g(x, y) = y/x$ (so $\gamma = -1$), but an analogous treatment can be given for $g(x, y) = xy$ (3). In each case, 1000 random values of each of $x$ and $y$ were generated from the lognormal distribution and (3) was fitted to the values by OLS regression to obtain estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$.

These experiments indicate that $<\text{ln}(x)>$ and $<\text{ln}(y)>$ merely move the value of $\hat{\beta}_0$ in the $\text{ln}(y)$-$\text{ln}(y/x)$ plane (data not shown), as would be expected from (10). In contrast, increasing $s_{\text{ln}(x)}$ or $s_{\text{ln}(y)}$ increases the deviation from the regression line and also rotates the values clockwise around ($<\text{ln}(y/x)>$, $<\text{ln}(y)>$) thereby changing $\hat{\beta}_1$, consistent with (11). For example, increasing $s_{\text{ln}(x)}$ from about 0.1 to 1.0 to 2.0 (Figure 4) results in a decline in $\hat{\beta}_1$, from 0.941 to 0.036, and in $R_1$ (from 0.969 to 0.185) (Table 1). The covariance of $\text{ln}(y)$ and $\text{ln}(y/x)$ ($= R_1 s_{\text{ln}(y)} s_{\text{ln}(x)}$) is about 0.16 and $\hat{\beta}_1$ is about 4.61 over this range (Table 1). Given that $R_1 \leq 0.011$ for these simulations, (9) and (11) are

$$R_o \approx (\lambda + 1)^{1/2}$$

and

$$\hat{\beta}_1 \approx (\lambda + 1)^{-1},$$

respectively, so the correlation between $\text{ln}(y)$ and $\text{ln}(y/x)$ depends on $\lambda$ alone.
Figure 4. Relationship between ln(y) and ln(y/x) for $s_{\ln(x)} = 0.103$ (A), 0.997 (B) and 2.037 (C) and, for each, $s_{\ln(y)} = 0.407$. In each case 1000 lognormally distributed random values were generated for x and y. Further details of the simulations are given in Table 1.

Table 1. Details of the simulations shown in Figure 4. In each case $<\ln(y)> = 4.615$ and $s_{\ln(y)} = 0.407$.

<table>
<thead>
<tr>
<th></th>
<th>Figure 4A</th>
<th>Figure 4B</th>
<th>Figure 4C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt;\ln(x)&gt;$</td>
<td>4.607</td>
<td>4.620</td>
<td>4.637</td>
</tr>
<tr>
<td>$s_{\ln(x)}$</td>
<td>0.103</td>
<td>0.997</td>
<td>2.037</td>
</tr>
<tr>
<td>$&lt;\ln(y/x)&gt;$</td>
<td>0.008</td>
<td>-0.004</td>
<td>-0.021</td>
</tr>
<tr>
<td>$s_{\ln(y/x)}$</td>
<td>0.419</td>
<td>1.076</td>
<td>2.073</td>
</tr>
<tr>
<td>cov(ln(y), ln(y/x))</td>
<td>0.1653</td>
<td>0.1643</td>
<td>0.1561</td>
</tr>
<tr>
<td>cov(ln(x), ln(y))</td>
<td>0.0003</td>
<td>0.0013</td>
<td>0.0095</td>
</tr>
<tr>
<td>$\lambda = s_{\ln(x)/\ln(y)}$</td>
<td>0.254</td>
<td>2.450</td>
<td>5.007</td>
</tr>
<tr>
<td>$\hat{\beta}_0$</td>
<td>4.608</td>
<td>4.616</td>
<td>4.616</td>
</tr>
<tr>
<td>$\hat{\beta}_1$</td>
<td>0.941</td>
<td>0.142</td>
<td>0.036</td>
</tr>
<tr>
<td>$R_0$</td>
<td>0.969</td>
<td>0.375</td>
<td>0.185</td>
</tr>
<tr>
<td>$R_1$</td>
<td>0.007</td>
<td>0.003</td>
<td>0.011</td>
</tr>
</tbody>
</table>

To examine this further, the same approach was used to generate lognormally distributed x and y except that uniformly distributed values of $\mu$ and $\sigma$ were used to ensure an even distribution of $<\ln(x)>$ and $s_{\ln(x)}$ (Figure 5A). For each iteration the OLS regression coefficients and $R_0$ were determined (Figures 5, B and C). As shown in Figure 1, (9) and (11) indicate that both $R_0$ and $\hat{\beta}_1$ decline with increasing $\lambda$ (13), as is shown in Figure 5C. Consistent with (10), when $\lambda$ is large, so that $\hat{\beta}_1$ is small (11), $\hat{\beta}_0 \approx <\ln(y)>$ and when $\lambda$ is small, so that $\hat{\beta}_1 \approx 1$ (11), $\hat{\beta}_0 \approx -\gamma<\ln(x)>$ (Figure 5B).

Figure 5. Relationship between $\lambda$ and (A) $<\ln(x)>$, (B) $\hat{\beta}_0$ and (C) $\hat{\beta}_1$ and $R_0$. Random values of $\mu$ and $\sigma$ were generated from the uniform distribution to ensure even representation (A). For each value of $\lambda$ 1000 lognormally distributed random values were generated for x and y. In (B) the horizontal line represents $<\ln(y)>$ and in (C) the curves are given by (13). The values of $s_{\ln(x)}$ range from 0 to 2.

V. DISCUSSION

We have shown that the correlation ($R_0$) between ln(y) and ln(x/y) is determined largely by the relative magnitude of $\lambda = s_{\ln(x)/\ln(y)}$ (Figure 1B). If $\lambda$ is small it is inevitable that $R_0$ is high (it can not be low), but even if $\lambda$ is larger it may be that $R_0$ is significant depending on $\gamma R_1$ (Figure 1B). Based on this analysis, the correlation between ln([PdG]) and ln([PdG]/[Cr]) shown in Figure 3A must be high simply because $s_{\ln(Cr)}$ is small. Given this, the relationship shown in Figure 3A, which is similar to that of Roos et al. [7], can provide no convincing support for the idea that [PdG] is ‘equivalent to’ [PdG]/[Cr]. Most importantly, this relationship can only be strong (Figure 1B), so the fact that this is the case ($R_0 = 0.855$ for the data in Figure 3A) conveys no significant information: it has no bearing on the equivalence or otherwise of the two measurements of urinary PdG.
Karl Pearson [47] pointed out that correlations of the form $y$ versus $x/y$ or $x/y$ versus $w/x$, among others, tend to be spurious and his point has been reinforced regularly ever since [48-53]. One of the best known examples of this is the correlation between the number of storks and the birth rate in a particular region which has been reported several times [54: 144-147, 55-57]. Despite the problem being well known, such analyses continue to be common [50, 51]. The relationship between $\ln(PdG)$ and $\ln((PdG)/[Cr])$ [7] is another example of a spurious correlation.

In essence the logarithmic transformation considered here (3) renders the correlation between $y$ and $x^y$ even more apparent. It is clear from (3) that the underlying relationship is just $\ln(y) = \ln(x)$, but where $\hat{z}$ is small (say $\hat{z} \leq 0.5$ or higher depending on $\gamma R$, Figure 1B), it is inevitable that $R_0$ is high (Figures 4A and 5C), but even if $\hat{z}$ is somewhat larger $R_0$ can be significant (Figures 4B and 5C). However, if $\hat{z}$ is large $R_0$ tends to be small (Figures 4C and 5C). Equation (9) indicates that it is not possible to observe a low $R_0$ for (3) if $\hat{z}$ is small (Figure 1B) and so it is incorrect to infer from data such as those shown in Figure 3A that [PdG] is a reasonable substitute for $[PdG]/[Cr]$ [7]. To draw this inference is to ignore the spuriousness of the correlation. While [PdG] may be a useful measurement in some circumstances, the apparent correlation between $\ln([PdG])$ and $\ln([PdG]/[Cr])$ [7] does not provide any significant support for the assertion.

Our general treatment of $\ln(y)$ versus $\ln(g(x, y))$ (3), as expressed in (9), includes as a particular case ($\gamma = 1$) the spurious correlation between $\ln([PdG])$ and $\ln([PdG]/[Cr])$ reported by Alliende et al. [6]. We defer to a later date consideration of the specific relationship between $\Delta g_{p0}$ and $[PdG]/[Cr]$ which has not yet been treated systematically despite the implicit assumption that they are equivalent [6, 7, 13].

VI. CONCLUSIONS
No matter how well data are analysed, if those data are flawed the analysis is also flawed. This is the case for what Pearson [47] called a “spurious” correlation. We have shown that the relationship between two measures of urinary PdG, the concentration ($\ln([PdG])$) and the creatinine-corrected concentration ($\ln([PdG]/[Cr])$) depends almost entirely on $\lambda$, the ratio of the standard deviations of $\ln([Cr])$ and $\ln([PdG])$ (9, 11). In practice, because $\sigma_{\ln([Cr])}$ is small, these two measures can only be highly correlated (Figure 1B) and so the fact that $R_0$ is high signifies nothing. Certainly, it can not be concluded from this relationship that [PdG] is as good a measure of urinary PdG as $[PdG]/[Cr]$. This is just one example of this class of spurious correlation, but it is a good reminder that a high correlation coefficient does not abrogate one’s responsibility to examine the data carefully.

REFERENCES


