Measuring Success

When we observe the world, we sometimes make mistakes. ***Michael Wallace****, on behalf of the measurement error topic group of the STRATOS Initiative, explains the potentially severe consequences of this often overlooked issue, and how statistics can help bring us back - or at least a little closer - to the truth.*

What did you eat yesterday?

It's a simple question, with a seemingly straightforward answer. In my case, I had toast for breakfast, a sandwich for lunch, and pasta for dinner. Do snacks count? I had an apple during the afternoon and, if you must know, a doughnut with my morning coffee (I was teaching first thing). Oh, do you need to know what I drank as well? I think along with the coffee there were two, no, maybe three cups of tea? I drank water throughout the day too, but didn't measure it...

Whenever we conduct statistical analyses, we make numerous assumptions. Many of these assumptions are familiar to anyone who has taken an introductory statistics class. We may assume our data "follow a particular distribution" (such as the famous 'bell curve'), or that the effect of a treatment is the same for different patients. Arguably the most common assumption of all, however, is barely mentioned: that the measurements we take are without error. In other words, when we ask for someone's height, or weight, or blood pressure, or simply what they had for dinner, we assume the measurement we get is exactly the same as whatever number we are *trying* to measure.

In some cases this is a reasonable assumption. Age, at least where reliable birth records are available, will be accurate, as will a handful of other variables such as sex or employment status. We may also be prepared to assume that our measurements are 'close enough' to what we're trying to observe that it is not worthwhile worrying about it. However, such cases tend to be the exceptions rather than the rule, and to assume our measurements are perfect (or sufficiently 'near-perfect') when in reality they are not can have grave statistical consequences.

A dramatic example occurred in the study of cervical cancer and human papillomavirus (HPV) infection.[1] The link between these two conditions is now well established, with HPV vaccinations a common recommendation to mitigate this risk. For many years, however, this relationship was dismissed, thanks in part to a number of studies which suggested no clear association between the two conditions. The cause? Measurement error.

An inconvenient truth

Measurement error occurs when the value of an observation does not equal the 'truth' we, the analysts, are trying to observe. In some cases this is easy to understand: if you're 160cm tall that represents the truth, but if I make a muddle of my tape measure I may end up with a measurement of 161cm. Often, though, the mere definition of 'truth' is surprisingly difficult to pin down. Your weight fluctuates throughout the day, week, month, and year. If I want to use weight as a variable in an analysis, what is the 'true' weight I'm trying to measure? Your weight right now, the average across the day, or something else altogether?

Furthermore, such errors come in many forms. At its simplest, we may view what we observe as equal to the truth plus some 'random' noise. This is the *classical* measurement error model, and includes the appealing scenario where we are just as likely to under- as over-estimate what we wish to observe. Often, however, errors may be systematic in nature. A common example is *white coat hypertension*, where blood pressure measure readings - often made in doctor's surgeries or other high-stress environments - are typically elevated. Asking people how often they smoke cigarettes or drink alcohol (or eat doughnuts) may result in other systematic errors.

A different type of model we may be willing to assume for the error is known as *Berkson* error. For example, if we wish to study the effect of air pollution on the health of a population, we might attempt to measure the levels of particulate matter in the environment. Air quality monitors will record such data, but it is likely unknown how much an individual person might be exposed to. We can therefore view the 'truth' (how much particulate matter an individual breathes in, for example) as equal to the error-prone observation from the air quality monitor, but with noise added. This is in contrast to the classical sense where the error-prone observation may be thought of as the truth with noise added. What type of error we encounter, whether it is systematic or truly random, and what variables in our analysis it affects, all impact both the problems it can cause and the methods available to us to address them.

While Berkson and classical errors are often seen to dichotomize the measurement error universe), there are more nuanced considerations. Errors in the outcome, such as misclassification of disease diagnosis, can pose unique challenges. Moreover, the outcome itself can sometimes influence the accuracy of reporting, creating what is known as *differential* measurement error. For example, someone who has received a lung cancer diagnosis may then overestimate their historical cigarette consumption as a result, potentially leading to exaggerated effect estimates.

In the case of HPV and cervical cancer, measurement error occurred in the detection of HPV infection itself. Identifying a patient as HPV-free when they were not, or vice versa, constitutes measurement error: the value we observe (whether HPV infection is detected or not) is different to the truth (whether the patient is actually infected). Sometimes, such error is referred to as *misclassification* because a categorical variable is under consideration.
Regardless of what we call it, measurement error represents a ubiquitous problem, especially in medical statistics where much of the literature relies on measuring attributes of people. Beyond health research, it can be just as problematic, affecting laboratory observations in chemistry or physics, or questions of econometrics (consider, for example, asking someone about their salary, or the value of their house). Regardless of discipline, the word 'error' can sometimes cause consternation among researchers, especially those using the best available tools or techniques. Unfortunately, truly perfect measurements may be impossible to take. Despite its prevalence, however, measurement error remains a relatively new - and comparatively specialist - area of research. Moreover, it is often ignored in practice.

But why?

A convenient untruth

For years, there has been one key suspect: the claim that measurement error's impact is 'not that bad'. This general belief stems from a convenient result that only holds in a few special cases. A common goal in research is to estimate the strength of a relationship between two (or more) variables, such as between HPV infection and cancer. When errors are consistent with the classical measurement error model, and not systematic (in other words, when what we observe is equal to the truth plus some noise that, on average, is zero), we can construct analyses where measurement error will at worst lead us to *under-estimate* the strength of such relationships (see "Attenuation and a Post-Hoc Fix", page 6). One perception is that the over-estimation of an effect is a far greater crime than underestimation: imagine, for example, if a pharmacist over- rather than under-estimated the efficacy of a new drug. There are also some statistical tests which can remain theoretically valid despite such attenuation (that is,

reduction in magnitude) of estimated effects. Consequently, we may also underestimate the negative impact of measurement error itself.
Whether these observations alone are sufficient grounds to view measurement error as an ignorable concern is a matter for debate in itself. Indeed, attenuated effects are not necessarily harmless. A particularly notable example comes from a 1990 article in *The Lancet*.[2] The authors studied the association between blood pressure and the risks of stroke and heart disease. Measurement error, and the aforementioned attenuation of effects, led to these risks being severely underestimated. Moreover, regardless of whether attenuation is ever acceptable, it can be shown - and especially in the case of systematic error - that the strength of an effect can be *over-estimated* as well. Anticipating the extent to which measurement error must be addressed can swiftly become a complex challenge.

The tendency to overlook the impact of measurement error due to the belief it only causes a weakening in effects is at best puzzling and at worst deeply concerning. In the case of HPV and cancer, for example, such a weakening of the relationship delayed implementation of health practices designed to prevent the former causing the latter (see "HPV and Cancer", page 7). More generally, the idea that this could be the only effect of errors in the increasingly complex models used by statisticians and non-statisticians alike would be laughable were the consequences not potentially so severe.

That is not to say, however, that the underappreciation of measurement error stems only from this misapprehension. Another factor is the perception that it is a problem too difficult to solve. When I raise the issue of measurement error with collaborators on scientific projects, for example, I am sometimes met with a resigned shrug that the measurements they have are the best they can do (or sometimes offence at the mere suggestion their measurements are not perfect!). Furthermore, while statistical methodologies to analyze error-prone data do exist, they are seldom known to a research team. With proper planning these techniques can be easily incorporated into an analysis, and ideally into the design of the study itself.

More data

One of the biggest obstacles we face in the study of measurement error is in characterizing the size and structure of the measurement error itself. If a set of patients each has their blood pressure measured once there isn't much we can do to detect - or correct for - the measurement error that is surely present. In general we need some additional data from which we can learn more.

Typically, the cheapest form of such data is known as *repeated measures* or *replicates*. As the names suggest, rather than taking a single measurement, we take repeated measures on at least a subset of our sample, giving us valuable insights into how accurate our data are. For example, if the difference between two measurements is very small, this would suggest our measurements are fairly accurate. A large discrepancy would give us greater concern that measurement error may be particularly problematic. At a very simplistic level, we could imagine taking two measurements and using their average, but more sophisticated techniques can take advantage of statistical theory to improve our analyses further.

A major problem repeated measures cannot (necessarily) solve is that of systematic error. After all, if your blood pressure is *always* elevated when visiting the doctor, an average of several measurements will also be elevated. The ideal source of additional information, therefore, are *validation* data, where some measurements are known to be perfect. This can sometimes be accomplished through more expensive or invasive procedures, or through the development of new techniques (as with the improvement in HPV testing; see "HPV and Cancer"). Often, however, this is simply not possible, either because no practical measurement mechanism exists, or whatever we are trying to measure in the first place is not well-defined. In the latter, we sometimes use what's known as a 'gold standard', where researchers determine the best measure they can, given practical or other constraints, reasonably obtain. Nevertheless, in such cases we must be careful in our interpretation, and ensure any associations are clearly set in the context of our gold standard and not lazily associated with some (unobtainable) 'perfect' measurement.

Besides replication and validation data, there is a handful of other ways to acquire information about the measurement error itself, but these tend to be less reliable or more complicated to implement. A common problem arises in practice where, owing to measurement error not being anticipated in the design of a study, no additional data are available to address this problem at the analysis stage. In such cases our last resort is often *external* data: measurements - or knowledge of measurement error - from other sources or studies. Of course, with this comes additional uncertainty about how accurately such external information translates to our own setting, but it may be our only choice.

All correct

There are therefore many options available to us, and what decisions we make at both the design and analysis stages of any study should be informed by measurement error considerations. We should begin by asking what types of error we anticipate, and in which variables, and then identify where in our design we can make useful accommodations.

For illustration, let's suppose that you're conducting a study and, having read this article and been convinced of the perils of measurement error, decide to anticipate its effects. You determine that while one of your variables will be measured with error, it will be with random, additive noise (in other words, classical and not systematic). Luckily, while validation data are too expensive to obtain, repeated measures can be collected relatively cheaply, so you measure every subject twice.

In the simplest of settings (where the myth of attenuation is, in fact, a reality), we can carry out a typical, or 'naive', analysis and apply a retrospective *method of moments* correction. For this we use our additional data to estimate some aspects of our problem - including the size of our measurement error - and then use a simple formula to correct our naive results (See "Attenuation and a Post-Hoc Fix").

In more general settings a popular technique is known as *regression calibration*. For this we again conduct our usual analysis, such as a linear or logistic regression, but replace the error-prone observations with our 'best guess' of the truth for each individual (or at least, our best guess based on what's observed or measurable). This is a little like using the mean of each subject's observed values, but takes the additional information about the size (and distribution) of the measurement error into account to yield more accurate results. It can even be used in the case of systematic error, if we know enough about that aspect of our error process.

Another popular method, known as *simulation extrapolation*, or SIMEX, takes a rather different but nevertheless intuitively appealing approach. Having learned about the structure of the measurement error in the system, we are able to simulate new datasets as if they were subject to more and more severe (that is, larger) measurement error. This can lead to a pattern in our resulting estimates: for instance if the strength of a relationship gets gradually smaller as the measurement error increases, we can follow this pattern backwards to the hypothetical scenario where no error is present. SIMEX is usually applied in the case of classical error, but can be used in other settings, such as with spatial data (see "Simulation Extrapolation", page 7).

There are, naturally, a whole host of techniques available for addressing the problem of measurement error in practice. If you have a dataset where measurement error is present, chances are there's a correction method that will fit your particular scenario. Many of these methods have implementations in common software environments such as Stata's **merror** package, `simex` in R and SAS macros from the United States National Cancer Institute. Simpler approaches, such as the method of moments correction and regression calibration, can even be implemented directly with relative ease (especially if you have a friendly statistician at hand). Of course, the use of any statistical methods must be approached with care, and close attention paid to any underlying assumptions upon which they rely.

Future perfect

Regardless of your statistical background (or the friendliness of any statistically-inclined colleagues), the most important lesson is that accounting for measurement error is by no means impossible. Indeed, it is often the case that with careful planning, measurement error can be minimized or avoided altogether. This can be achieved, for example, through the use of more precise measurement techniques, or by selecting variables that are less susceptible to mis-measurement.

Failing that, additional data - though not essential - vastly enhance our capacity to address the problems measurement error might cause. What's more, they can often be collected with comparatively little cost. From here, there is a vast - and expanding - array of methods available for correcting our analyses.

Ultimately, however, those who do not take advantage of such approaches must face a hard truth: failure to take measurement error into account can completely invalidate their findings. While we can seldom expect perfection in measurement *or* modelling, there is much that can - and should - be done in getting us a little bit closer to it.

References:

1 Schiffman M H, Schatzkin A. Test reliability is critically important to molecular epidemiology: an example from studies of human papillomavirus infection and cervical neoplasia. Cancer Research, 1994; 54:1944s-1947s.

2 MacMahon S, Peto R, Cutler J, Collins R, Sorlie P, Neaton J, Abbott R, Godwin J, Dyer A, Stamler J. Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. Lancet, 1990; 335:765-774.

Further Reading:

For more on measurement error, see:

- Carroll R J, Ruppert D, Stefanski L A, Crainiceanu C M. Measurement Error in Nonlinear Models: A Modern Perspective. 2nd Edition. Chapman and Hall/CRC Boca Raton FL, 2006.

- Gustafson P. Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments. Chapman and Hall/CRC Boca Raton FL, 2004.

- Coggon D, Rose G, Barker D J P. Measurement error and bias; Chapter 4 in Epidemiology for the Uninitiated. Fifth edition; BMJ Publishing Group, London UK. 2003. Access at http://www.bmj.com/about-bmj/resources-readers/publications/epidemiology-uninitiated

Attenuation and a Post-Hoc Fix If we consider simple linear regression, where a 'straight-line'

relationship is sought between a response Y and an exposure X, we are trying to find a relationship between the two of the form

$$Y = \alpha + \beta X + \epsilon$$

where ϵ represents 'noise'. Viewed this way, the strength of the relationship between X and Y is given by β: if it is large and positive, then Y will increase quickly as X increases, and vice versa. If X represented dosage of a drug, and Y the resulting health of a patient, we can see how estimating β tells us how well the drug is working. Standard statistical methods allow us to estimate β and determine the relationship between X (the *covariate*) and Y (the *outcome*).

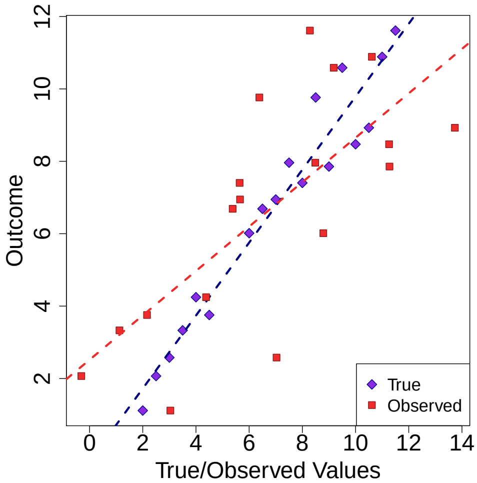
If our covariate is measured with error, then instead of X, we might observe $X^* = X + U$, where U represents the measurement error (and is assumed to be independent of X). Larger values of U would mean our observed value X^* is further from X, suggesting more severe measurement error. If we carry out our statistical analysis we will use X^* in our equation, instead of X, which can affect our resulting estimate of β.

In this case, if we assume that U follows a normal distribution with mean 0 and variance σ_u^2, and that X has variance σ_x^2, then our standard analysis would estimate not β, but

$$\beta^* = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\beta.$$

Because $\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2} < 1$, this means $|\beta^*| \leq |\beta|$: our estimate will be *attenuated.* In other words, we will underestimate the strength of the relationship between X and Y. This is illustrated by the true relationship being given by a steeper slope in the figure.

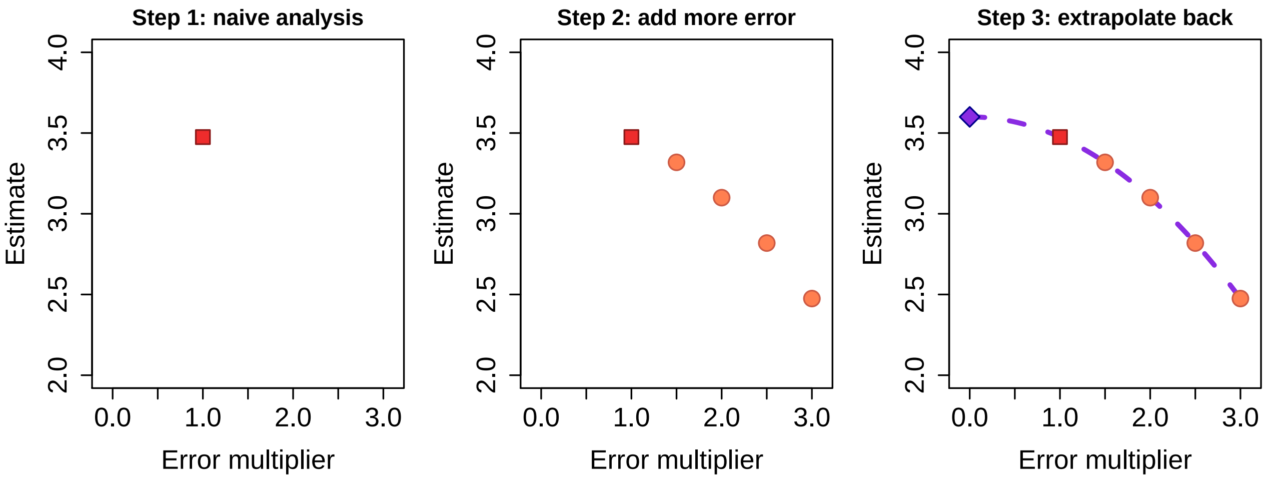
In this setting, however, if we can estimate σ_x^2 and σ_u^2 we can easily estimate β: conduct a 'naive' analysis to obtain an estimate of β^*, then multiply by our estimate of $\frac{\sigma_x^2+\sigma_u^2}{\sigma_x^2}$. This correction (a form of regression calibration) 'undoes' the attenuating effect of measurement error in this situation, returning us to an estimate of the true relationship between X and Y.



The slope of a straight line fit to a dataset can give an indication of the strength of the relationship between the two variables: a steeper line suggests a stronger relationship. Here, a straight line fit to the error-prone observations (triangles) is less steep than that fit to the true values (circles), and the measurement error would therefore lead to an attenuated effect estimate. Intuitively, we can view the points as being 'stretched' horizontally (the error either increases or decreases the x-axis values) but not vertically, as the outcome is assumed to be measured without error. This leads to a shallower fitted line.

The curious reader may wonder what would happen if, in contrast, the outcome were subject to to measurement error of the same form detailed here, while the covariate remained error-free. In such a scenario it can be shown that the strength of the relationship would be maintained, but with greater uncertainty in its accuracy. Again, however, such a convenient result cannot be relied upon in more general settings.

Simulation Extrapolation



Simulation extrapolation begins with a 'naive' analysis which does not take measurement error into account, producing a (likely biased) estimate of the effect of interest. In addition, the size of the measurement error is estimated, so that new datasets may be simulated where additional error is added to the already error-prone measurements. For example, a dataset with an 'error multiplier' of 2 is one where the observed error-prone measurements have been replaced with simulated values estimated to suffer from twice as much measurement error as the original data. For each of these new datasets, the effect is estimated. This pattern is then used to extrapolate back to the scenario where no error is present.

HPV and Cancer

In their 1994 paper, Schiffman and Schatzkin studied the relationship between

HPV infection and cervical intraepithelial neoplasia (CIN, a precursor to cervical cancer). To do so, they analyzed two case-control studies, where individuals with CIN (the cases) were compared to those without (the controls). The first study used an older method to measure HPV infection, a version of the Southern Blot, while the second used a newer method called Polymerase Chain Reaction, or PCR. Although some good tests were available at the time of the first study - including a more reliable version of the Southern Blot - they were highly labour-intensive and thus unsuitable for large-scale epidemiological studies.

Repeat testing of HPV infection was available on a subset of individuals in each study, whereby evidence of the fallibility of the Southern Blot was derived. Of 51 participants successfully re-tested in the Southern Blot study, there were 11 'discordant pairs': individuals who received differing results. In contrast, all 43 of those re-tested in the PCR study produced the same diagnosis.

For both studies, the risk of developing CIN was assessed using odds ratios. An odds ratio is a numeric measure of how much one's risk of an outcome increases in the presence of another factor, compared to if that factor is absent. In this case, odds ratios were used to compare the risk of developing CIN between those with, and without, HPV infection. Here, the difference between the Southern Blot and PCR measures of HPV infection were stark: the odds ratio for CIN associated with a positive HPV test was 3.7 using Southern Blot, but 20.1 using PCR. Using the more reliable test revealed a much stronger association between HPV infection and CIN.

Later work went on to establish a causative relationship between certain strains of HPV and CIN, leading to the development of vaccination against HPV infection. Today, the Centers for Disease Control and Prevention (CDC) in the United States, and the United Kingdom's National Health Service, along with many other countries' health agencies, recommend vaccination against HPV for young women.

Acknowledgments:
This article is written on behalf of the Measurement error and Misclassification Topic Group of the international STRengthening Analytical Thinking for Observational Studies (STRATOS) Initiative (http://stratos-initiative.org). STRATOS aims to provide accessible and accurate guidance documents for relevant topics in the design and analysis of observational studies. The Topic Group is led by Laurence Freedman and Victor Kipnis, with members Hendriek Boshuizen, Raymond Carroll, Veronika Deffner, Kevin Dodd, Paul Gustafson, Ruth Keogh, Helmut Kchenhoff, Pamela Shaw, Anne Thiebaut, Janet Tooze, and Michael Wallace.