Evidence-Based Decision Making

What's Evidence?

We all make rational decisions that are based upon evidence don't we? After all, the last time we ate a pepperoni pizza for supper we had bad dreams, so that's clearly good evidence that in future we should take care to avoid eating pepperoni after 6pm!

Except of course, that it obviously isn't. What else did we eat that evening, and were the bad dreams related to what we ate or did they occur because of something entirely different. Yet, in everyday life much of our 'evidence' is rather like that, anecdotal in nature, far from objective, and likely to be rather selective.

That doesn't matter too much when deciding which pizza to order, since that sort of decision is not one that needs to be objective (well, assuming no allergies), and there will be many other factors that will influence it. However, if when we are walking in the hills we come to a fork in the path, it might be useful to consult the map to decide which one we need to take if we are to get back before dark. And now of course, we want the map to help us work out where we are by using objective evidence about what is around us in order to make our decision (are we opposite a wood, or near to a bridge, or…).

So when the stakes get higher, for both personal and professional decisions, then we may (should) seek out good quality evidence that can be used to help make a choice. The evidence may take different forms, but usually we will want to find the best and most reliable sources wherever possible. And evidence that is based upon objective measurements, such as the energy rating for a washing machine that we are considering buying, will usually be a better basis than the salesman's promises!

Using data from empirical studies to make decisions

When we need information about *physical* properties to help make a decision (weight, energy consumption, brightness of illumination, etc.), then the result of taking a measurement will usually be sufficient. However when our decisions relate to information about how humans behave and react, as occurs in clinical medicine, education, psychology, and in software engineering too, then we are likely to need rather more than individual measurements and have to turn to empirical studies.

Why is this so? Well, experimental studies (trials) in these subjects make extensive use of human participants. And for these studies, we can expect to find quite a wide spread of values in the results obtained when measuring the effect of some intervention, even when a study is performed rigorously. This 'spread' stems from the natural variation that occurs between humans, whether it be physiological (affecting clinical studies) or of experience and ability (affecting computing studies). Because of this, it is necessary to take many measurements, and to use a variety of participants. This is in contrast to experimental studies in the natural sciences, where any variation in the results is usually small, largely attributed to measurement error, and usually has a normal distribution.

The effect of this is conveniently illustrated by the use of *box plots* when reporting the outcomes from a study that involves human participation. A box plot 'groups' the results in quartiles, each containing 25% of the values and then plots these to show how the data is distributed. The range of the middle 50% of values is shown as a 'box', and the outer 'quartiles' as long lines. Within the box, the middle line indicates the value of the *median* (not the mean). Figure 1 shows a simple illustration of this, where the vertical axis measures the outcome variable (in this case time) and we then draw separate plots for the results with and without intervention. (Position on the x axis and the width of the boxes are arbitrary.)

<insert Figure 1 here>

As we can see, in this case the participants allocated to the group who were using the intervention (a software tool) mostly took less time to perform their task than those in the control group, although a few did take longer (so the intervention box plot has a wider spread, and is also rather skewed).

Box plots provide a useful visualisation of how widely the results may vary due to individual characteristics of the participants. However, it is still necessary to employ statistical tests to determine whether or not the difference between the two datasets is significant or could simply have occurred randomly.

As well as this naturally-occurring variability, there may also be some variation arising from possible experimental errors that might bias the results of a study. When the effects of these are combined we can see that relying upon the results of a single trial of an intervention in order to make decisions may be somewhat risky. This however was how medicine was often practised in the past, with doctors making clinical decisions about a patient's treatment that were based on the results of those trials that they were aware of, or that were recommended by experts as being most appropriate. Unfortunately though, *expert judgement* is itself an unreliable way of identifying relevant evidence (although it may be useful when interpreting it). Experts usually have opinions, which will inevitably be apt to bias their selections.

The Evidence-Based Paradigm and Systematic Reviews

What is now considered to be the 'model' of evidence-based decision-making first became established in clinical medicine. Much of the stimulus for this came from the work of Archie Cochrane (1909-1988) who was a pioneer in seeking ways to determine the best evidence about treatments. He was highly critical of the lack of sound evidence for many commonly used medical interventions and particularly encouraged the use of greater rigour in clinical trials ("randomize till it hurts"). In honour of his contributions to thinking about evidence, the not-for-profit organisation that now oversees the clinical review process was named after him (the *Cochrane Collaboration*, www.cochrane.org).

A key step in the development of an evidence-based approach was the realisation that aggregating the results from separate studies could help reduce the effects of both natural variability and also of any bias that might occur in individual studies. But to be really effective this process of aggregation needs to be performed in a rigorous and systematic way. At this point, we need to shift from the medical past to the

software engineering present and discuss what we do in computing. Before doing so though, we should note that should you have to consult your doctor, it is highly likely that he or she will address your needs by drawing upon knowledge derived through this process of aggregation.

For computing, the evidence-based process used to derive sound and unbiased evidence from empirical studies consists of five steps (Kitchenham et al., 2004).

1. Convert a need for information about some intervention (technique, process etc.) into an answerable question.
2. Track down the best evidence relating to the question in a systematic, objective and unbiased way.
3. Critically appraise the evidence for validity, impact and applicability (how useful it is)
4. Integrate the critical appraisal with domain expertise about the given topic and the needs of any stakeholders.
5. Evaluate the outcomes and use to improve the preceding steps.

The first three of these steps form what we term a *systematic review*, while the other two form a process usually termed *knowledge translation* (KT). A systematic review is also termed a *secondary study*, since it does not involve any direct activity with participants (as occurs in a *primary study*). A secondary study is therefore one that aggregates the results from a set of primary studies.

A really important aspect of a systematic review is that as much as possible is *planned* in advance as it is important to avoid the risk of 'fishing' for useful patterns in any results. Particularly important elements to specify in advance are the search criteria for identifying relevant primary studies, and the inclusion and exclusion criteria used to determine which of these should be used as inputs. Because the application of these criteria does require some expertise about both the topic of the review and also the conduct of systematic reviews, it is also common to make decisions about which primary studies to include by using two analysts, carefully resolving any differences in their decisions (Kitchenham & Charters, 2007). It is worth noting that the 'reduction factor' is often quite high, it is quite normal to begin with several thousand candidate studies, and end up using around 20-50 of these. The example in the side box illustrates this more fully.

The way that the outcomes of the eventual set of primary studies is synthesised is also important, since it helps influence the quality of the eventual data from the review as well as any decisions made using this (Cruzes & Dybå, 2011). While statistical *meta-analysis* is the most powerful form, this can rarely be used in computing studies due to the heterogeneous nature of the primary studies typically found, so less powerful and more qualitative forms usually need to be employed.

How far can we use systematic review outcomes to make decisions?

So, if your doctor can make decisions about how best to treat you using the outcomes of a systematic review, can we make our technical decisions in the same way? In principle, the answer is yes, although in practice it is a bit less clear-cut than for your doctor. This is largely because clinical studies usually have a comparative form (treatment versus placebo) with the participants being *recipients* of the treatment, making it possible to use meta-analysis to synthesise the outcomes, and

hence providing sound statistically-based evidence. In contrast, computing studies usually have a range of forms, and the participants are often asked to perform skilled tasks within a range of organisational contexts, which may then introduce many other 'confounding factors', and complicate eventual synthesis.

In a recent study we examined 216 systematic reviews, published between 2004 (when the use of systematic reviews began to be adopted in software engineering) and the end of 2014, to see if they contained material that could potentially be used to support teaching of core software engineering ideas (and hence also, provide advice on practice). From these, we identified 59 studies that were considered to provide useful experience about such topics as cost modelling, requirements elicitation, model-driven engineering, agile methods, the needs of start-up companies, and so on. (An earlier version of this study was published as (Budgen et al., 2012).)

Few of the studies had very clear-cut results (and even then, there were caveats, usually relating to the quality of the primary studies used as inputs). However, we should also note that in computing, it is unusual for the use of any one practice to make a very big difference when compared with the effect of using another. Some of the studies also embody useful experiences derived from specific primary studies about when or how a particular technique might be particularly useful.

We should not be too surprised at this. Back in 1987 Fred Brooks Jr. explained why the nature of software made it unlikely that the 'silver bullet' solutions desired by managers could ever be feasible for software development activities (Brooks, 1987). So, when it comes to making decisions in computing, consulting the outcomes of a systematic review will rarely make the decision for you, but is likely to provide you with some important insight into what does work, and when (and possibly why). And knowing what factors may be important can still usefully make it possible to come to an *evidence-informed* decision. (Given that every patient the doctor sees has their own set of personal factors, clinical decisions are usually evidence-informed too.)

Will things change in the future? Well, this is still early days for secondary studies, with experience from only little more than a decade of performing systematic reviews available in software engineering, and with many of these being performed partly as a learning process. With growing experience, it is likely that more and better studies will be performed. This is also true for primary studies, which increasingly may be performed to meet needs identified by systematic reviews. But, while we can expect that more and better systematic reviews will emerge over time, as step 4 in the above process indicates, when using the outcomes from a systematic review to help with making decisions, there will always be the need to consider the user's own context as well.

So, evidence-informed decision making is feasible now, and is likely to become more so in the future. However (the inevitable caveat), if it is to be employed effectively, users of evidence also need to be educated about how to use it. In themselves, systematic reviews are no more of a silver bullet than all of the ideas that have gone before, but their use forms an important step towards putting software engineering education and practice on to a much sounder basis by providing more rigorously derived knowledge about our practices.

References

Brooks Jr., F. P. (1987), 'No silver bullet: essences and accidents of software engineering', *IEEE Computer* **20**(4), 10–19.

Budgen, D., Drummond, S., Brereton, P. & Holland, N. (2012), What scope is there for adopting evidence-informed teaching in software engineering?, in 'Proceedings of 34th International Conference on Software Engineering (ICSE 2012)', IEEE Computer Society Press, pp. 1205–1214.

Cruzes, D. S. & Dybå, T. (2011), 'Research synthesis in software engineering: A tertiary study', Information and Software Technology 53(5), pp. 440 – 455.

Kitchenham, B., Dybå, T. & Jørgensen, M. (2004), Evidence-based software engineering, in 'Proceedings of ICSE 2004', IEEE Computer Society Press, pp. 273–281.

Kitchenham, B. & Charters, S. (2007), Guidelines for performing systematic literature reviews in software engineering, Technical report, Keele University and Durham University Joint Report.