ELSEVIER

Information and Software Technology 40 (1998) 811–821

INFORMATION AND SOFTWARE TECHNOLOGY

Combining empirical results in software engineering

Lesley M. Pickard[a,*], Barbara A. Kitchenham[a], Peter W. Jones[b]

[a]*Department of Computer Science, University of Keele, Staffordshire ST5 5BG,UK*
[b]*Department of Mathematics, University of Keele, Staffordshire ST5 5BG,UK*

Abstract

In this paper we investigate the techniques used in medical research to combine results from independent empirical studies of a particular phenomenon: meta-analysis and vote-counting.

We use an example to illustrate the benefits and limitations of each technique and to indicate the criteria that should be used to guide your choice of technique. Meta-analysis is appropriate for homogeneous studies when raw data or quantitative summary information, e.g. correlation coefficient, are available. It can also be used for heterogeneous studies where the cause of the heterogeneity is due to well-understood partitions in the subject population. In other circumstances, meta-analysis is usually invalid. Although intuitively appealing, vote-counting has a number of serious limitations and should usually be avoided.

We suggest that combining study results is unlikely to solve all the problems encountered in empirical software engineering studies, but some of the infrastructure and controls used by medical researchers to improve the quality of their empirical studies would be useful in the field of software engineering.

Keywords: Empirical studies; Meta-analysis; Vote-counting; Software engineering

1. Introduction

In most scientific disciplines, experiments and empirical studies are a standard means of furthering scientific understanding. Formal experiments are used to test scientific hypotheses in the knowledge that the results of an experiment will generalize to the population from which the experimental subjects/objects are drawn.

We usually assume that the same ideas apply to empirical studies in software engineering. We attempt to use empirical studies to investigate the efficacy of our software engineering methods and/or the impact of various project/personnel factors on project productivity or product quality. However, we often find the different empirical studies of the same phenomenon report different and sometime contradictory results. There are a number of reasons why this might occur. One reason is that empirical studies in disciplines such as software engineering, that are strongly influenced by individual differences among human subjects, do not usually have 'key experiments' allowing us to refute our hypothesis in the way physics or chemistry does. Another reason is that we usually struggle to find experimental subjects for our empirical studies. This means that we might not have sufficient data points to detect a phenomenon even if it really existed.

Another problem concerns the extent to which our empirical studies contribute to our understanding of, and ability to control, software engineering phenomena. Although we can apply standard scientific techniques such as formally designed experiments or statistical data analysis techniques, it is not clear that our results have the same generality as empirical results in other disciplines. The problem arises from the difficulty of defining the population of software engineering subject and objects to which any results can properly be said to generalize.

For example, if we perform a formal experiment intended to evaluate different testing techniques in a university setting, we will use some standard (but relatively small programs) seeded with defined defects, and we will use student volunteers to act as experimental subjects. In such circumstances, if we find that one testing technique is superior to another what do our results really mean?

If all we can say is that with the specific group of student subjects, and the specific programs and the specific set of defects, one testing method has performed better than another, our experiment has not told us very much. However, if we want to make some general statement about the superiority of a particular testing method, we must be confident that our students are a random sample from the population of software developers, our

* Corresponding author. Fax: + 44 1782 713082; e-mail: lesley@cs.keele.ac.uk

0950-5849/98/$ - see front matter
PII: S0950-5849(98)00101-3

experimental programs are a random sample from the population of software programs, and our defects are a random sample from the population of software defects. It seems clear from descriptions of experiments in software engineering that our notions of the populations of programmers, programs and defects are very sketchy and that our selection of subjects is seldom random (although experimenters are usually careful to allocate subjects randomly to treatments).

For other forms of empirical study, the problems of generalization are even worse. If we collect data about projects in a particular company our results can only be applied to that company. Datasets composed of projects from a variety of different companies do not solve the problem, unless the basis on which the dataset is derived is statistically valid. It must be recognized that data provided by a self-selected group of companies on a similarly self-selected set of projects violates the basic principle of randomness that is necessary for generalization to be possible, and fails to define any population to which results could be generalized.

If it is not possible to generalize the results of a single empirical study, is there any point in undertaking empirical studies in software engineering? In our view, empirical studies are still valuable for evaluating the efficacy of software engineering methods, but we need to consider the 'weight of evidence' rather than rely on single experiments. By weight of evidence, we mean the extent to which empirical results are consistent across a variety of different studies. This issue is also important in medical studies. We believe that there are enough similarities between the problems faced in medical research and the problems faced in empirical software engineering research to warrant an investigation of the techniques the medical area use for combining evidence and their possible applicability to software engineering.

In a medical investigation, whether it is an experiment, a trial, a case-study or an observational study, researchers are trying to detect an effect, e.g. whether a particular treatment has an effect on reducing the severity of a disease, whether a particular characteristic or environmental factor increases the chances of getting disease, etc. We also wish to detect effects in software engineering, e.g. whether the use of a particular design method or tool will increase development productivity more than an alternative or what factors will affect product quality. As well as confirming the existence of an effect we often wish to know its magnitude. This is called the *effect size*. The effect size can be measured in two ways — by a standardized difference (if the study has comparisons) or by a measure of association (i.e. correlation coefficient). An effect size is an indicator of the average magnitude of an effect.

Three methods of combining results from individual studies are commonly used in medical research: combining the test statistic values, categorizing the outcome of tests of hypothesis (vote-counting) and estimating treatment effects across studies (meta-analysis). The results of an individual study are often given in the form of '*p*' values that give the probability of obtaining a significant difference between treatments if the true treatments were really the same. Combining test statistic values is based on combining the results of the '*p*' values from different studies, i.e. combining the test statistic values [13]. However, there is an interpretation problem with such tests. Rejection of the combined hypothesis only means that the null hypothesis cannot be accepted, not that the alternative hypothesis is true in every study. This means that if you are interested in whether there is a difference in productivity between two design methods, this type of test may give the result that (for a particular significance level) we can reject the combined null hypothesis that there is no productivity difference shown in the studies. This does not imply that there is a common productivity difference in the population as a whole, we can only conclude that there is a productivity difference between the two design methods in at least one study. Therefore, the test does provide any information about whether the effect is consistent across different studies. Because of this difficulty we concentrate, in this paper, on investigating meta-analysis and vote-counting. We describe each technique and consider whether it would be useful in the context of empirical software engineering studies. The choice of which technique to use is determined by the amount of information available to you but we will highlight in this paper some other factors that influence the applicability of the techniques. We illustrate the use of each technique on an software cost estimation example. In the example we use we have access to the raw data, but often this is not available. We discuss some of the problems that might occur if you have only the results of the individual studies not the raw data.

Recently there have been some attempts to use meta-analysis [14] and informal vote-counting [3], in software engineering studies. Before software engineering researchers adopt these techniques, we believe that it is important to be aware of their risks and limitations. Thus, we use our example to explain the potential problems of the techniques.

2. Meta-analysis

Meta-analysis is a technique for pooling data from different studies [12,13,21]. It is mainly used in Clinical Trials (i.e. controlled experiments) but there has been some use in epidemiology and observational studies [6,9,15,22–24,26]. In order to use a meta-analysis technique, you must have a quantitative measure of effect size for each individual study.

The aim of meta-analysis is to provide a quantitative and objective procedure for combining the information from different studies. With a subjective review of previous studies, a reviewer can influence and bias the review and different reviewers can come to different conclusions. The use of meta-analysis is intended both to resolve the uncertainty when the results of studies disagree, and to increase

the confidence in the results obtained from individual studies. The outcome of a meta-analysis is an average effect size with an indication of how variable that effect size is between studies. This section discusses the use of meta-analysis techniques and its potential problems.

Meta-analysis involves three main steps:

1. Decide which studies to include in the meta-analysis.
2. Estimate an effect size for each individual study.
3. Combine the effect sizes from the individual studies to estimate and test the combined effect.

2.1. Study selection

Which study you chose to include in the meta-analysis is crucial since it will influence the rest of the analysis and the results. Study selection has two, sometime conflicting, objectives: to include only appropriate, valid studies and to include as many studies as possible. The creation of an objective ''inclusion criteria'' helps the selection process. Inclusion criteria are usually based on the type of empirical study, the test hypothesis, the choice of effect measures, and the available explanatory variables.

2.1.1. The type of empirical study

It is important to base any meta-analysis on individual studies of the same type, for example, all case studies, all cohort studies or all formal experiments. The greater the degree of similarity between the studies the more confidence you can have in the results of a meta-analysis.

It should be noted that a meta-analysis using case-control studies assumes that the controls from the different studies are comparable. Even with an individual study there is a problem in software engineering about choosing an appropriate control, [18]. Therefore it is difficult to assemble a set of studies where all the controls are comparable and, if they are not comparable, a meta-analysis is compromised.

2.1.2. The test hypothesis

In medicine an example hypothesis might be ''Is the death rate from breast cancer with a new treatment lower than that of an existing treatment?'' This means that all the studies included in the meta-analysis must have a measure of the death rate (or raw data from which it can be calculated).

In software engineering, an example hypothesis is ''Does the use of a new tool improve the productivity of a development (without any detrimental effect on the quality) compared to the use of an existing tool?'' You must decide which productivity and quality measures are appropriate and assure that all studies have used the same measure of productivity and the same measure of quality.

2.1.3. Common explanatory variables

Two known explanatory variables that influence breast cancer are age and gender. Therefore we may decide to include only studies that have information about age and gender included.

Project factors could be used to control a meta-analysis in the same way age and gender are used in the breast cancer example. For example, we might consider characteristics like application type, development method or implementation language. Whether the existence of information about these factors should be used as an inclusion criterion depends on whether software engineers can agree that they have an influence on productivity or quality of the development process or product.

2.1.4. Common measures

Although a meta-analysis should include as many studies as possible, it must only include studies that have comparable measures. The choice of which measure to use for a particular software attribute depends on which is most appropriate in a particular environment and may differ substantially between organizations. This will limit the number of studies that can be included in any meta-analysis.

2.1.5. Study selection problems

Selection of studies can be compromised by publication bias. Publication bias results from the preference of journal editors to publish results that demonstrate significant differences between the treatments and reject manuscripts that report insignificant results [13,20]. If publication bias has occurred, a combination of the published individual study results in a meta-analysis will result in an over-estimation of the size of effect and an inflated probability that the difference is significant.

Publication bias can be checked for during the meta-analysis in a variety of ways:

- Calculate number of studies required to refute the conclusions of the meta-analysis [13]. If the number of studies is small compared with this value, selective sampling may have influenced the results. In this situation, the meta-analysis results are likely to be biased and cannot be generalized.
- Produce a Funnel plot. Observed effect sizes are plotted against the sample size. The points should scatter around an underlying 'true' value, producing a funnel pattern. Gaps indicate potential publication bias [27,28].
- Begg's quantitative method. This uses the sample size of study and an estimate of the size of the source population, [27]. The problem with this method is that it needs information about specific incident rates and proportion of population who would enroll in the trial. Such values are difficult to obtain.

Medical researches have access to a large database (MEDLINE). This database is a major reference for identifying existing studies on a specific phenomenon. Access to this type of facility reduces the chance of publication bias. In the absence of such a facility, good access to relevant work is limited to personal knowledge and literature searches.

The decision as to whether to include a particular study in a meta-analysis should include an investigation of whether the study is of a high enough quality to provide confidence in its results. Medical researchers attempt to provide a consistent view of quality, between experiments, by deriving quality criteria for empirical studies. Some researchers have suggested going a step further and using quality criteria as a means of weighting individual study results in meta-analysis. Although using quality criteria as weighting factor is controversial, deriving quality criteria for software engineering studies might be beneficial since such criteria provide background information that is useful when undertaking any assessment of previous research.

2.2. Size effect estimation for individual studies

You need to obtain a standardized indicator of effect size for each study before you can pool the information from individual studies into one meta-analysis. This effect size indicator needs to be independent of the particular unit or scales used in any individual study to obtain measures that are comparable across the different studies.

The choice of an indicator for effect size depends on the type of studies included in your meta-analysis. If the individual studies involve direct comparison between experimental conditions then it is likely that your effect size indicator will involve taking the difference between the mean values for each condition. The difference must be standardized to remove scale differences, i.e. divided by the combined standard deviation. If your studies do not involve a comparison, e.g. survey data, then your effect size is likely to be an association or correlation.

Size effects depend on standard, consistent measures. In epidemiology studies and clinical trials, meta-analyses use a defined measure of risk as their standardized measure of effect size. Unfortunately, software studies do not have a standard, easily interpreted measure of treatment effect that is recognized and agreed by all researchers.

Furthermore, extra information is required for software engineering studies in comparison with medical studies. This is because the same principal measures can be collected using many different definitions, and the actual definitions are needed to ensure that the measures are comparable across the studies. If the measures are incompatible (or are suspected to be) then the studies should not be combined and a meta-analysis is inappropriate.

What is often missing in both medical and software engineering research is access to the raw data used in the individual studies. This would greatly improve the validity of any results from a meta-analysis because it would allow the raw, unadjusted data to be used to construct individual effect sizes, instead of relying on summary information that may include unidentified adjustments. In software engineering, the Journal of Empirical Software Engineering, [2], is starting to address this problem. This journal maintains a repository of studies materials and raw data from the papers it publishes.

2.3. Combining effect sizes/estimating statistics and hypothesis testing

There are various different methods of combining the individual study effect sizes. Regardless of what you use for an effect size measure, the pooled effect size should be weighted. If the studies have different sample sizes then the estimates from larger studies are likely to be more reliable tan the smaller studies. In addition, studies with small variances are likely to be more precise. Thus, the effect size from the different studies is usually weighted either with respect to the study's sample size or its variance.

The decision of which method of combination is the most suitable depends on various factors, for example, the effect size measure used for individual studies, available computational facilities and precision required for estimate of the pooled effect size. However, these are minor considerations compared to the main issue of whether you can assume that the individual studies are homogeneous or not. This issue influences both the method of meta-analysis and interpretation of results.

Studies are assumed to be homogeneous, if all the studies are measuring the same underlying phenomenon and the effect size estimates only differ due to sample variability. In meta-analysis this is called a *Fixed Effect*. A fixed effect can be assumed when you are estimating one true effect for the population and the studies are representative samples of the general population of the investigation. For example, if you are investigating the effect of a new anti-inflammatory drug on rheumatoid arthritis in the general population of arthritis suffers, you can assume the studies are homogeneous if experimental subjects are drawn at random from patients attending general practice surgeries. The studies would not be considered a random sample if subjects were taken from a specialist clinic.

A meta-analysis estimates the true or population effect size by calculating an average value of the individual study effect sizes (which are themselves averages). If the studies are not representative of the population then the meta-analysis may give a misleading result. For example, if your studies mainly included very extreme cases of joint inflammation from rheumatology units, the meta-analysis estimate of the general population effect of arthritis suffers will be biased towards the effect of the drug on severe inflammation. How much this bias influences the meta-analysis results depends on whether the drug has the same effect on all inflammations regardless of severity. However, this information will not be available from your meta-analysis if you have only samples taken from severe cases.

The assumption that the studies are all representative samples of the overall true effect and only differ due to sampling error is not always valid. In this situation the studies are said to be heterogeneous.

There are tests available to detect whether the studies are heterogeneous. In the case when the effect size is a correlation coefficient, there are two tests for homogeneity of correlations; the *Q* test (that is based on Fisher's *z* transformation of the correlation coefficient) and the Likelihood Ratio Test (that uses the maximum likelihood estimate of *p*). These tests check whether the amount of variation between the study effect size estimates is more than would be expected if the studies come from a single population.

If you detect that the studies are heterogeneous it is important to identify the cause of the heterogeneity. There are two main reasons for the presence of heterogeneity: incomparable measures and the existence of sub-populations. If the measures are incomparable (for example, different definitions have been used to derive the measure), then the effect sizes should not be combined. Heterogeneity can also occur when the overall population would naturally be partitioned into different sub-populations. We will use a simple example to show the effect of analysing studies that have sub-populations.

Suppose you are interested in the incidence rate of breast cancer and have sampled from two crowds — a football match and Harrod's Sale day. It is likely that the football crowd would be pre-dominantly male whereas the Harrod's crowd would be pre-dominantly female. The different studies would give different results because gender has a major influence on the incidence of breast cancer.

However, suppose your were not able to recognize gender (e.g. you were an alien) then you would not be able to tell why the study results were different. If you combined the studies in a meta-analysis you would obtain an estimate for the population as a whole. The use of a Fixed Effect model would underestimate the variance because there is significant difference between the means of the two sub-populations. In this situation a Random Effects model may be useful to estimate the average population incidence rate for breast cancer. A Random Effects model allows for variability due to an unknown factor influencing in the effect sizes for different studies and produces a larger estimate of the population variance than the Fixed Effect model. It incorporates estimates for both the sampling error and the variability in the sub-populations.

There is a danger in applying meta-analysis to heterogeneous studies because it is difficult to tell if the studies used in the meta-analysis are representative of the population. In the above example if the proportion of independent studies was approximately 50% male-dominated studies and 50% female-dominated studies, the overall estimate of the incidence of breast cancer would be a valid estimate because the distribution of the studies is equivalent to the distribution of gender in the population. However, if you had more pre-dominantly female studies than male-dominated studies, the average, based on all the studies, would not provide a valid estimate of the breast cancer in the population as a whole. If you cannot recognize gender you would not know that the type of study would cause a bias in a meta-analysis.

If you could recognize gender, you would be able to partition the studies into sub-populations of male and female and so reduce the heterogeneity. A meta-analysis using the sub-populations would estimate the effect size for the female incidence of breast cancer and the effect size for the male incidence of breast cancer separately. This leads to a rather more meaningful assessment of breast cancer incident rates than those related to the entire population.

A basic problem is that if you detect heterogeneity you may not know what is causing it. There are many circumstances the effect size does not appear to be constant across all the studies but the reason for the difference cannot be explained by one (or several) known indicators. This is often the case in software engineering. For example, many researchers have suggested that different types of applications or use of different languages have an influence on productivity, but there is no agreed set of language and application categories that are used in all productivity studies.

Medical researchers have set up Cochran groups, [7] where they agree on a standard experimental protocol for investigating a particular phenomenon. In particular, they agree on a standard dimensionless value for reporting results. The aim of the group is to accumulate a set of homogeneous studies that can be used in a meta-analysis. When a researcher completes an empirical study, the results are reported to the relevant Cochran group. The study report often includes the raw data (definitions identified within the group) as well as the results. Cochran groups are responsible for performing a meta-analysis on all of the studies reported to them and updating the meta-analysis as and when new empirical results are reported.

2.4. Example of a meta-analysis

In medicine, meta-analysis is considered more appropriate for combining the results of replicated experiments than for combining the results of observation studies. However, the examples of meta-analysis is software engineering have been related to combining the results of observational studies, so we have selected an example of the same type.

In order to investigate the practical use of meta-analysis techniques, we investigated the relationship between project effort and product size found in a number of different software engineering studies. The datasets were all collected as observational studies, not as formal experiments. It should be noted that the results of this meta-analysis are inherently biased because the studies were chosen from readily available datasets, not by a thorough review of the literature.

The first step in any meta-analysis is to identify which studies to include in the analysis using some defined inclusion criteria. The criteria we used were very simple since the datasets were collected for a variety of reasons and were not collected with respect to any particular experimental design. The criteria for the inclusion of studies were:

- Studies which included both effort and size information.

Table 1
Summary of $\hat{\rho}$ values for individual studies

Study[a]	Sample size (n_i)	$\hat{\rho}$	95% confidence interval for $\hat{\rho}$	
			Lower	Upper
1	33	0.8870	0.7817	0.9431
2	19	0.9612	0.8998	0.9852
3	17	0.6405	0.2310	0.8572
4	15	0.7227	0.3342	0.9013
5	63	0.8605	0.7788	0.9135
6	15	0.7980	0.4833	0.9301
7	33	0.6677	0.4209	0.8225
8	15	0.9327	0.8051	0.9778

[a] The sources for the studies are given in Appendix A.

This criterion was used in order to ensure comparability with respect to study scope.

- Studies with project-based information. This criterion was used in order to ensure that the studies used a common object of study.
- Studies with effort collected in months and size collected in lines of code. This criterion was used in order to ensure that the studies used comparable measures.

Several problems were identified because of the nature of the datasets chosen:

- large variation between the different companies, in type of development process and type of software produced;
- lack of any information about the measurement error;
- little information about the actual measurement definitions used to collect the raw data;
- apart from the effort and size information, little consistency between the studies about other explanatory variables;
- lack of normality in the data — most of the meta-analysis assume data is distributed normally. In order to normalize the data, values were transformed using a natural logarithmic transformation.

The purpose of the meta-analysis was to investigate the relationship between effort and size, measured in lines of code. The effect size measure used was the Pearson correlation coefficient. The mean difference was not applicable because it was an investigation of an association not an investigation of comparative treatments.

Table 1 shows the estimated correlations (i.e.$\hat{\rho}$ values) for the individual studies. The results are also shown in Fig. 1 that includes one extra point (i.e. point 9). This point shows the meta-analysis results.

Fig. 1 and Table 1 provide information about how similar the different individual studies are. However, there are two issues that need to be addressed if you are using meta-analysis:

- Whether or not the results of a study should be removed from the analysis if it is different. If the meta-analysis is intended to show a general result, a study with an atypical result may greatly influence the meta-analysis because it is so different.
- An individual company can exhibit a relationship between effort and size but not when using lines of code as the size measure but using some other measure of size, e.g. number of modules. This is a particular problem in software engineering studies which are attempting to combine information from different companies. There is no universal appropriate size measure, it depends on many different factors including type of development and type of final product. It is meaningless to use an inappropriate measure for a company

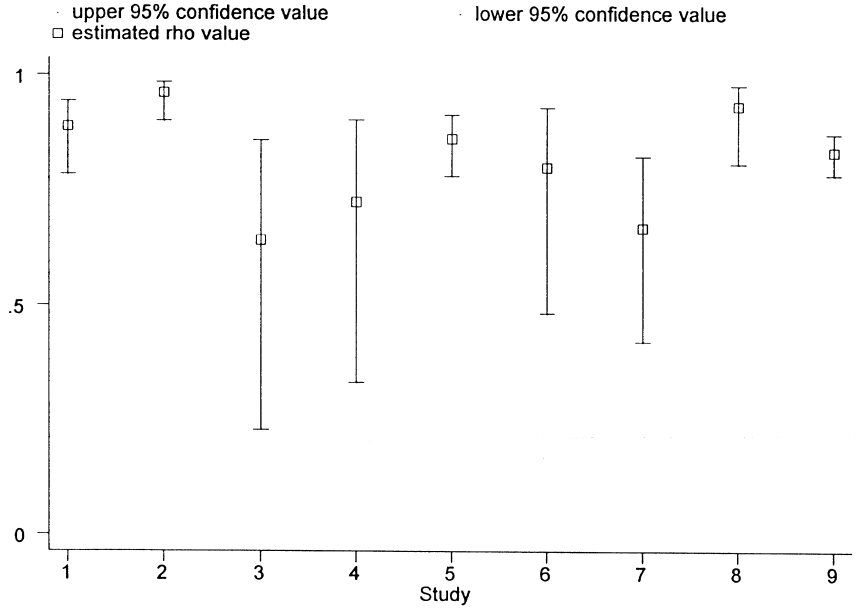

Fig. 1. Estimated correlation coefficient with associated 95% confidence intervals.

Table 2
Summary of $\hat{\rho}$ values for individual studies

Study[a]	Sample size (n_i)	$\hat{\rho}$	95% confidence interval for $\hat{\rho}$ Lower	Upper
1	33	0.8870	0.7817	0.9431
2	19	0.9612	0.8998	0.9852
3	17	0.6405	0.2310	0.8572
4	15	0.7227	0.3342	0.9013
5	63	0.8605	0.7788	0.9135
6	15	0.7980	0.4833	0.9301
7a	10	0.4855	− 0.2075	0.8541
7b	12	0.7097	0.2292	0.9121
7c	11	0.8181	0.4284	0.9504
8	15	0.9327	0.8051	0.9778

[a] The sources for the studies are given in Appendix A.

just because other companies have used it, but if a company's information is to be included in the meta-analysis it must use the same measure.

Point 9 on Fig. 1 represents the estimate of the combined effect size (i.e. the mean of the individual $\hat{\rho}$ values), with associated 95% confidence intervals, using the fixed effect model. The fixed effects model assume the studies are homogeneous, i.e. assumes that there is a common underlying correlation ρ, and each individual study delivers an independent estimate of ρ. The confidence intervals are based on the *standard error* of the *combined* estimate. However, in software engineering it is often important to know what is likely to happen in an individual study. Individual study results are more variable than the average effect over different studies. A *standard deviation* is required to obtain a prediction interval for *individual studies*.

We used a fixed effect model in the first stage of the meta-analysis, although it unlikely that a fixed model would be appropriate due to variation between the studies. If a fixed effect model is appropriate, it is easier to interpret the results. The combined $\hat{\rho}$ value for the meta-analysis indicated a strong correlation of 0.83 between effort and size (measured in lines of code) with a 95% confidence interval of $0.78 \leq \rho \leq 0.87$. However, in a specific study the estimate of ρ could vary from 0.4 to 0.98. The *Q* test suggested the presence of heterogeneity between the individual studies indicating the need for a random effects model (*Q* was 21.98 compared to a critical value of 14.07, the 95% point of the χ^2distribution with 7 degrees of freedom).

Villar recommends a sensitivity analysis to check the robustness of the meta-analysis results with respect to the choice of statistical methods used to combine the studies and the inclusion of lower quality studies [28]. This is because the reliability or trustworthiness of the meta-analysis depends on the rigour of the application of the technique. If there is any doubt about whether to include a study, especially on the grounds of quality, the way forward is to include the study and check its impact on the meta-analysis results using a sensitivity analysis.

A sensitivity analysis of our example meta-analysis showed that no single individual study had a major impact on the estimate of ρ, but studies 2 and 7 had a major influence on the heterogeneity between the studies. When these studies were removed, the *Q* test indicated that the rest of the studies were homogeneous, i.e. were estimating the same underlying ρ value, (*Q* was 8.52, which is less then the critical value of 11.07). There is no real justification for removing individual studies from a meta-analysis just because they contribute significantly to heterogeneity. However, if studies that behave differently have some common characteristic, e.g. included different types of project to the projects in the other studies, this may indicate the existence of a sub-population.

One of the problems of combining empirical dataset results in many situations is the lack of information about the study and the raw data it was based on. This may affect the results of any analysis. For example, study 7 data is composed of data from three different environments. Combining the data together provides an average relationship but reduces the amount of information in the meta-analysis. It may also provide a misleading result if three different environmental conditions and languages exhibit different relationships between effort and size. Table 2 shows study 7 divided into three individual studies. Studies 7b and 7c show correlations that are significantly different from zero, but study 7a shows a non-significant correlation.

For this set of studies there is very little change in the results when the meta-analysis is performed using study 7 as three individual studies. There is still a strong correlation (0.85 compared to 0.83) between effort and size, with a 95% confidence interval of $0.80 \leq \rho \leq 0.88$ for the mean correlation. The *Q* test still suggested the presence of heterogeneity between the individual studies (*Q* was 20.56 compared to a critical value of 16.92, the 95 percent point of the χ^2distribution with 9 degrees of freedom).

A sensitivity analysis indicated that the removal of study 2 resulted in the *Q* test indicating that the remaining studies were homogeneous (*Q* was 12.48 compared to a critical value of 15.51, the 95 percent point of the χ^2distribution with 8 degrees of freedom). The *Q* test still suggested the presence of heterogeneity between individual studies when any other individual study was removed, even the only non-significant study, 7a (*Q* was 16.77 compared to a critical value of 15.51). The inclusion of a non-significant study has suggested that the very high correlation of study 2 is unlikely to be an estimate of the same parameter value of ρ.

Study 5 could also be sub-divided. However, in this case the breakdown is by mode of development, [4], instead of environment or organization and, therefore, would introduce another level of detail that is unknown for the other studies.

If your studies are heterogeneous, it is necessary to use a random effects model. Using a random effects model on the

full set of 8 studies, the results were very similar to using the fixed effect model except that the mean correlation was slightly less (0.81 compared to 0.83). For the random effects model, the variance of the estimated mean ρ values was 0.0147 and the estimate of the sample variance was 0.0112. This gave an estimate of the variance of the mean population correlation due to unknown causes of 0.0035, which corresponds to a standard error of 0.06. The results are not presented in terms of a confidence interval, because the presence of heterogeneity implies that the individual effects in the different studies are dependent on the individuals study conditions and not representative samples of the general study population.

The example has shown the vulnerability to heterogeneity of meta-analysis of observational studies. To reduce heterogeneity, medical statisticians make use of:

1. Objective selection criteria for inclusion of studies in the meta-analysis.
2. Sensitivity analysis of meta-analysis results to assess the sensitivity of the results, both to the individual studies and to the method of analysis.
3. Standard size effect metrics (such as odds ratios) that reduce problems of comparable measures.
4. Well-established explanatory variables that partition the population into homogeneous sub-sets.

If we use meta-analysis for combining software engineering observational studies, only the first two techniques are currently possible. Furthermore, because we do not have a standard, context-independent definition of software measures, we have an additional possible source of heterogeneity: use of incomparable measures.

One approach to study selection, being suggested in the medical community, is the use of good quality criteria to assess the validity of each study before it is included in a meta-analysis. In addition, we should make sure that we have a well-defined process for planning and performing meta-analyses. For example, [9] give a good standards and criteria list, suggested by [25], for undertaking a meta-analysis:

- Study design:
  1. The study design should be prepared before the study begins.
  2. The methods used to find relevant studies should be stated.
  3. A criteria for inclusion/exclusion of studies should be defined.
  4. Summary information on characteristics of study subjects should be provided.
- Combinability:
  1. Criteria for inclusion/exclusion should be defined in advance.
  2. The criteria should be reviewed if any studies exhibit very atypical effect sizes.
- Potential biases:
  1. Any potential sources of bias should be identified. For example:

     Are selection criteria valid?
     Are you able to extract of information from the individual studies (especially if the studies are not controlled experiments or trials).
- Statistical analysis:
  1. If a meta-analysis results are not significant the power of the test should be checked.
  2. Any possible sub-populations should be identified and separate meta-analysis should be performed within the categories.
- Sensitivity analysis:
  1. The studies should be analysed in two or more ways if possible.
  2. The quality of individual studies should be determined and an assessment of quality should be incorporated into final results.
  3. The studies should be checked for publication bias.
- Application of results:
  1. The conclusion should be established – whether the combined results provide a definitive, effective final answer, or a tentative one such that further individual studies are required.

[25] also recommend identification of any support for the results of the meta-analysis e.g. plausibility or indirect evidence from other studies.

3. Vote-counting methods

Vote-counting is a conceptually simple method. It uses the outcome of tests of hypothesis reported in different individual studies, e.g. whether a correlation was found to be significantly different from zero (either positively or negatively) or not significantly different from zero. Since the technique does not depend on the actual effect size values, it does not require all the stringent assumptions of the meta-analysis technique, e.g. comparable measures. Vote-counting is based on the assumption that there is one common underlying phenomenon e.g. a single underlying correlation coefficient. However, if only the significance levels of tests are known this assumption cannot be tested.

Vote-counting involves categorizing the different outcomes of the hypothesis tests into three groups:

- significant positive effect
- significant negative effect
- non-significant effect

Each study is classed as either a success or a failure. The classification depends on your *hypothesis*. For example, an

Table 3
Power of test of the null hypothesis if the alternative hypothesis is true

Study	Sample size	Value of ρ for alternative hypothesis			
		0.4	0.5	0.6	0.7
1	33	0.7639	0.9218	0.9862	0.9992
2	19	0.5410	0.7345	0.8887	0.9727
3	17	0.4981	0.6873	0.8526	0.9565
4	15	0.4520	0.6329	0.8059	0.9312
5	63	0.9520	0.9958	0.9999	~ 1
6	15	0.4520	0.6329	0.8059	0.9312
7	33	0.7639	0.9218	0.9862	0.9992
8	15	0.4520	0.6329	0.8059	0.9312

investigator, looking at whether the use of a new design tool will improve productivity, will allocate a success rating to any study that has a significant positive effect and a failure rating too all others. X is the number of successes in a set of k studies and is equal to the sum of the X_i where X_i takes the value 1 if study is a success and 0 if a failure. The proportion of successes is equal to X/k. We reject the hypothesis that there is no effect if X/k is greater than a predetermined cutoff value (CV).

There are two rather different ways of deciding on the cutoff on value: a formal statistical method and an ad-hoc method.

The statistical method of determining CV is based on the following argument:

1. The probability of rejecting the hypothesis of no effect in each individual study, when there really is no effect, is just the significance level of the individual test (e.g. 0.05).
2. If you have a set of k independent studies each using the 0.05 significance level *and there was no true effect*, 5% of the studies might have found a significant effect by chance.
3. Therefore if significantly more than 5% of the studies showed a significant effect, we can reject the hypothesis that there is no underlying effect.

For a particular value of k, we can work out the appropriate value of CV using the binomial distribution:

$$\text{Prob}\{\text{proportion of successes} > \text{CV}\} = \text{Prob}\left\{\frac{X}{k} > \text{CV}\right\}$$

$$= \sum_{i=[CV^*k]+1}^{k} \binom{k}{i} p^i (1-p)^{k-i}$$

where [CV*k] is the greatest integer less than or equal to CV*k and 'p' is the Binomial parameter which is set to 0.05 (i.e. the probability of obtaining a significant effect when there is no real effect). For $k = 9$, the value of CV*$k \leq$ l, i.e. we reject the null hypothesis if two or more studies demonstrated a significant positive effect. Thus, in our example, where all of the 8 studies showed a significant positive correlation we would reject the hypothesis of no effect.

In principle, if more than about 15% of studies show a significant effect, the vote-counting technique will usually reject the hypothesis that there is no underlying effect. Some researchers believe that a 15% level is too low a value to represent a true consensus and prefer to set an ad-hoc cutoff level themselves. For example, [19] suggest pre-setting CV to 0.33 when there are three outcomes (and to 0.5 when there are two). This means that you reject the hypothesis that there is no effect if more than a third of the studies are a success. In this case, you can use the binomial distribution to determine the significance level of the test

In our example, we have two outcomes: a positive correlation is considered a success and any other outcome is considered a failure. If we present CV to be 4, we need 5 or more successes in order to reject the hypothesis that there is no underlying effect. Using the binomial distribution we can identify the significance level of this test as follows:

$$P(x > 4 | \rho = 0) = \sum_{5}^{8} \binom{8}{i} (0.05)^i (0.95)^{8-i} = .000015.$$

Thus you are much less likely to reject the null hypothesis (i.e. $\rho = 0$) when it is true using Light's approach than if you use the cutoff value based on the statistical method. However, you are also more likely to accept the null hypothesis when false.

Formally, we say that the power of the test based on Light's method is less than the power of the test based on the formal statistical method. Informally, this can be appreciated by considering what happens if the true effect is difficult to detect (e.g. insufficient data points). In this case, a relatively high proportion of studies would not detect the true effect, so a meta-analysis using a cutoff point based on 50% of the individual studies being a success might incorrectly accept the null hypothesis. A meta-analysis using a cut-off value closer to 15% would be more likely to reject the null hypothesis. We confirm this more rigorously[1] in the following section.

The power of a test under the alternative hypothesis is the probability of rejecting the null hypothesis given that the *alternative* is true. The power is related to the number of data points in a study and the true value of the effect. For example Table 3 shows the power of the test of the null hypothesis for each individual study given several different values of ρ.

If there were a true underlying correlation (e.g. $\rho = 0.6$) and all our studies were based on the same number of data points (e.g. 33 data points), in the long term, we would expect 98.6% of studies to report a positive result. In fact, we have studies of different sizes, so we can use the average power of the individual studies i.e. 89%. Thus, we can use

[1] If you accept the informal argument, or do not want to consider statistical issues in greater depth, we recommend that you skip the next few paragraphs.

the Binomial distribution to assess the power of the two vote-counting methods given that the true value of $\rho = 0.6$:

Power of formal statistical method $= P(x > 1|\rho = 0.6)$

$$= \sum_{2}^{8} \binom{8}{i} (0.11)^i (0.89)^{8-i} = 0.9999.$$

Power of Light's method $= P(x > 4|\rho = 0.6)$

$$= \sum_{5}^{8} (0.11)^i (0.89)^{8-i} = 0.99.$$

Furthermore, if the actual value of ρ is smaller, the power of Light's method becomes much worse. For example, if $\rho = 0.4$, and we assume the power of each individual is reduced to 0.61, the power of the test based on Light's method is 0.62 whereas the power of the test based on the statistical approach is still 0.99.

4. Conclusions

Empirical studies of phenomena in software engineering often report different results. It would, therefore, be useful to combine the results of independent studies to obtain a common assessment of the nature of the phenomenon of interest. In this paper, we have considered two methods of combining the results of independent studies that have been proposed for use in the field of medicine: meta-analysis and vote-counting. Empirical studies in medicine have some similarities with empirical studies in software engineering, in particular results in both areas are both strongly influenced by individual differences between human subjects. Thus, an investigation of techniques for combining results used in medical statistics would seem to be relevant to software engineering. In addition, some researchers in software engineering are starting to use these techniques, albeit sometimes rather informally. In this paper we have attempted to explain how to apply these techniques with the help of a software engineering example.

Our example has identified a number of problems applying the techniques to observational software engineering studies. Some problems are inherent in the techniques, others are inherent in the application of the techniques to software engineering.

Vote-counting has a number of inherent problems which indicate that it should not be used as a method of combining empirical study results. In particular, it allows us to test only very weak hypotheses (in the example, we could only test whether the underlying correlation was 0 or not) and, counter-intuitively, a large number of independent studies does not provide us with any more confidence in our results than a smaller number, (in the example, once 2 or 3 studies have caused us to reject the null hypothesis further positive studies have little impact on our hypothesis test). In addition, vote-counting has a stringent requirement that the phenomenon under investigation is a single common phenomenon. If the effect is context dependent vote-counting is invalid. The only cases in which vote-counting might be appropriate are when either a software engineering phenomenon has been assessed using different measures in different studies, or the information reported from the studies is very limited (for example, significance levels are quoted but the values of the test statistics and the raw data are not). In both cases, meta-analysis cannot be performed and vote-counting is the only method available for combining results.

Meta-analysis allows us to assess common effect size, estimated from the effect sizes of each individual study. We are able both to test the hypothesis that the effect size is non-zero and to provide an estimate of the common effect size. Thus, meta-analysis leads to much stronger statistical inferences than can be made from the vote-counting. However, meta-analysis results are less trustworthy and more difficult to interpret if the individual studies exhibit heterogeneity. Since heterogeneity is usually found when different studies give different results (i.e. we have contradictory results), it appears that meta-analysis is of least use under the conditions when we would most like to use it. Furthermore, meta-analysis will not overcome basic deficiencies in the contributing studies. If the individual studies are of poor quality or are biased, any meta-analysis will be invalid. Thus, meta-analysis cannot help to overcome problems that result from individual studies being unable to draw subjects and objects at random from well-defined populations.

In our view, the lessons to be learnt from meta-analysis in medical research are not so much the statistical techniques but the infrastructure medical researchers have put in place to support meta-analysis. In particular, software engineering would benefit from:

- bodies that co-ordinate meta-analysis studies similar to Cochran groups that maintain records of replicated studies and update estimates of size effects as and when the results of new studies are reported to them;
- a database facility such as MEDLINE which maintains records of all experiments reported on phenomenon of interest;
- agreed quality standards for software engineering studies such as those reported by [25] (see Section 2);
- defined procedures for certain experiments that ensure that independent replications of an experiment can be combined.

Initiatives such as the Journal of Empirical Software Engineering which maintains a repository of studied materials and raw data of the papers it publishes are a useful starting point in this direction (see http://kapis.*www.wkap.nl/kapis/CGI-BIN/WORLD/journalhome.htm)*.

New statistical models (e.g. Multi-level Statistical Models [11] and Bayesian Hierarchical Models [8] have been developed that may help us to model the variations

between the studies. The use of these types of models could be beneficial in combining information from different empirical studies. However, they require detailed statistical knowledge. Software packages that perform multi-level models are becoming more readily available (e.g. Mln and BUGS) that will encourage their use but they must be used with caution; the use of a package is not a substitute for detailed knowledge of the technique. Also, the structure of the data must known in order to model the hierarchies and cross-classifications properly.

The incorporation of Bayesian statistics into the hierarchical model theoretically allows the incorporation of expertise into the model. However, the derivation of the required prior probabilities, and the resultant posterior, is too complicated for routine use. New methods have recently been developed to help (e.g. MCMC methods, [10] but are not readily available at present.

Even with new complex models being developed, it is unlikely that we will make much substantive progress until we address the issue of ensuring individual studies are properly conducted. There appear to be two critical issues:

1. We need proper definitions of software engineering populations and agreed methods of sampling those populations.
2. We need to agree a set of standard measures which are recorded for all empirical studies that will eventually allow us to define appropriate sub-populations (i.e. we need to define explanatory variables for use in software engineering studies analogous to gender and age in medical studies).

These concerns suggest that we need more research into the theory of empirical studies in software engineering as well as more empirical studies.

Acknowledgements

This research was undertaken as part of EPSRC research grant GR/L28371.

Appendix A. Data sources

Study 1: Belady–Lehman [5]
Study 2: Bailey–Basili [1]
Study 3: Yourdon [5]
Study 4: Wingfield [5]
Study 5: Boehm [4]
Study 6: Kemerer [16]
Study 7: Kitchenham–Taylor [17]
Study 8: Data made available to Mermaid project (Esprit)

References

[1] J.W. Bailey, V.R. Basili, A meta-model for software development resource expenditures. Proceedings 5th International conference on Software Engineering (1981) 107–116.
[2] V.R. Basili, W. Harrison (Eds.), Empirical Software Engineering: An International Journal, Kluwer Academic Publishers, 1996.
[3] R.D. Banker, C.F. Kemerer, Scale economics in new software development, IEEE Transactions on Software Engineering 15 (10) (1989) 1199–1205.
[4] B.W. Boehm, Software Engineering Economics, Prentice Hall, 1981.
[5] S.D. Conte, H.E. Dunsmore, V.Y. Shen, Software Engineering Metrics and Models, Benjamin/Cunnings Publishing Company, Inc., 1986.
[6] P.L. Canner, An overview of six clinical trials of aspirin in coronary heart disease, Statistics in Medicine 6 (1987) 255–263.
[7] Web-site: http://hiru.mcmaster.ca/cochrane.
[8] D. Draper, Bayesian Hierarchical Modelling (unpublished manuscript, Web-site: *http://www.bath.ac.uk/ ~ masdd*), 1998.
[9] J.L. Fleiss, A.J. Gross, Meta-analysis in epidemiology, with special reference to studies of the association between exposure to environmental tobacco smoke and lung cancer: A critique, Journal of Clinical Epidemiology 44 (1991) 127–139.
[10] W.R. Gilks, D.J. Spiegelhalter, Marcov chain Monte Carlo in practice, Chapman and Hall, 1996.
[11] H. Goldstein, Multilevel statistical models, Kendall's Library of Statistics 3 (1995).
[12] S. Greenland, Quantitative methods in the review of epidemiologic literature, Epidemiologic Reviews 9 (1987) 1–30.
[13] L.V. Hedges, I. Olkin, Statistical methods for meta-analysis, Academic Press, 1985.
[14] Q. Hu, Evaluating alternative software production functions, IEEE Transactions on Software Engineering 23 (6) (1997) 379–387.
[15] B. Jones, P. Jarvis, J.A. Lewis, A. F. Ebbutt, Trials to assess equivalence: the importance of rigorous methods, British Medical Journal 313 (1996) 36–39.
[16] C.F. Kemerer, An empirical validation of software cost estimation models, Comm. ACM 30 (5) (1987) 416–429.
[17] B.A. Kitchenham, N.R. Taylor, Software development cost estimatio, J. Systems and Software 5 (5) (1985) 267–278.
[18] B.A. Kitchenham, S.G. Linkman, D. Law, Critical review of quantitative assessment, Software Engineering Journal, 9 (2) (1994).
[19] R.J. Light, P.V. Smith, Accumulating evidence: Procedures for resolving contradictions among different research studies, Harvard Educational Review 41 (1971) 429–471.
[20] R.J. Light, D.B. Pilemar, Summing up: The science of reviewing research. Cambridge, MA: Harvard University Press, 1984.
[21] R.J. Lilford, J.G. Thornton, D. Braunholtz, Clinical trials and rare diseases: a way out of the conundrum, British Medical Journal 311 (1995) 1621–1625.
[22] M.P. Longnecker, J.A. Berlin, M.J. Orza, T.C. Chalmers, A meta-analysis of alcohol consumption in relation to risk of breast cancer, JAMA 260 (1988) 652–656.
[23] I. Romieu, M. Hernandez-Avila, M.H. Liang, Oral contraceptives and the risk of rheumatoid arthritis: A meta-analysis of conflicting literature, British Journal of Rheumatology 28 (Suppl. 1) (1989) 13–17.
[24] L. Rushton, D.R. Jones, Oral contraceptive use and breast cancer risk: a meta-analysis of variations with age at diagnosis, parity and total duration of oral contraceptive use, British Journal of Obstetrics and Gynaecology 99 (1992) 239–246.
[25] H.S. Sacks, J. Berrier, R. Reitman, V.A. Ancona-Berk, T.C. Chalmers, Meta-analysis of randomized controlled trials, N England Journal of Medicine 317 (1987) 450–455.
[26] R. Shinton, G. Beevers, Meta-analysis of relation between cigarette smoking and stroke, British Medical Journal 298 (1989) 789–794.
[27] T.D. Spector, S.G. Thompson, The potential limitations of meta-analysis, Journal of Epidemiology and Community Health 45 (1991) 89–92.
[28] J. Villar, G. Carroli, J.M. Belizan, Predictive ability of meta-analysis of randomised controlled trials, The Lancet 345 (1995) 754–772.