Eur J Epidemiol (2008) 23:237–242
DOI 10.1007/s10654-008-9230-x

METHODS

Data splitting as a countermeasure against hypothesis fishing: with a case study of predictors for low back pain

Fredrik A. Dahl · Margreth Grotle · Jūratė Šaltytė Benth · Bård Natvig

Received: 11 October 2007 / Accepted: 7 February 2008 / Published online: 21 February 2008

Abstract There is growing concern in the scientific community that many published scientific findings may represent spurious patterns that are not reproducible in independent data sets. A reason for this is that significance levels or confidence intervals are often applied to secondary variables or sub-samples within the trial, in addition to the primary hypotheses (multiple hypotheses). This problem is likely to be extensive for population-based surveys, in which epidemiological hypotheses are derived after seeing the data set (hypothesis fishing). We recommend a data-splitting procedure to counteract this methodological problem, in which one part of the data set is used for identifying hypotheses, and the other is used for hypothesis testing. The procedure is similar to two-stage analysis of microarray data. We illustrate the process using a real data set related to predictors of low back pain at 14-year follow-up in a population initially free of low back pain. "Widespreadness" of pain (pain reported in several other places than the low back) was a statistically significant predictor, while smoking was not, despite its strong association with low back pain in the first half of the data set. We argue that the application of data splitting, in which an independent party handles the data set, will achieve for epidemiological surveys what pre-registration has done for clinical studies.

Keywords Data splitting · Hypothesis fishing · Data dredging · Two-stage analysis · Low back pain

Contribution Grotle and Natvig have contributed with medical expertise on musculoskeletal disease, and have performed the model building. Natvig has also been responsible for the collection of survey data. Dahl has developed the data splitting method, in cooperation with the other three authors, and performed the final hypotheses tests. Saltyte Benth has handled the data splitting, and assisted in the model development and hypothesis testing. All four authors have taken active part in the writing of the manuscript.

F. A. Dahl (✉)
Helse Sør-Øst Health Services Research Centre, Akershus University Hospital, Mail drawer 95, 1474 Lorenskog, Norway
e-mail: fredrik.dahl@ahus.no

M. Grotle
From the National Resource Center for Rehabilitation in Rheumatology, Department of Rheumatology, Diakonhjemmet Hospital, Oslo, P.O. Box 23, Vinderen 0319, Oslo, Norway

M. Grotle
Division for Neuroscience and Musculoskeletal Medicine (FORMI Section), Ullevaal University Hospital, 0407 Ullevaal, Oslo, Norway

J. Šaltytė Benth
Helse Sør-Øst Health Services Research Centre, University of Oslo, Mail drawer 95, 1474 Lorenskog, Norway

B. Natvig
Section of Occupational Health and Social Insurance Medicine, Institute of General Practice and Community Health, Faculty of Medicine, University of Oslo, P.O. Box 1130, Blindern 0318, Oslo, Norway

Introduction

The concept of statistical significance may be the single-most important mathematical invention for applied science. Its use has become so widespread and commonplace that many non-mathematical readers may not be actively aware of its true meaning. To briefly review, the statement "X is correlated with Y at significance level alpha" signifies, "If no true correlation between X and Y exists, the probability of obtaining the observed correlation is less than alpha." The *P*-value of a test is therefore a measure of surprise; the smaller the *P*-value, the greater the surprise. Standard practice has been to set alpha at 0.05, which literally allows for a

5% chance of erroneously reporting a significant finding (Type I error). One cannot interpret the *P*-value as a probability of *having made* a Type I error, so 5% significance does not imply that the conclusion is correct with a 95% probability. Such statements are meaningful only in a Bayesian context where one assigns a priori probabilities to hypotheses. The present article addresses non-Bayesian (frequentist) analysis, which is by far the most common in epidemiology.

Ioannidis [1] purports that most scientific findings are likely to be false, despite being reported as statistically significant. One of his arguments, which we support, is that the pressure to publish creates an incentive for researchers to simultaneously address a large number of hypotheses and selectively report only "significant" results. This conduct is labelled *hypothesis fishing* or *data dredging*. (In older papers the term *data mining* has also been used, but the meaning of this has shifted toward discovery of valid patterns in databases.) Also, the investigator may run different statistical tests (e.g. parametric and non-parametric tests) for any given hypothesis, which is a less recognized form of hypothesis fishing.

In an epidemiological and health services setting, complex data sets based on statistical surveys are commonplace, in which hundreds of variables are collected from thousands of people. Data collection for a survey requires enormous effort, with both individual and collective demands on the researchers and respondents. From a purely economic point of view, it appears logical to "turn the data set upside down," searching for anything of interest buried in it. Therein lies the temptation to launch a large-scale fishing expedition for all potentially interesting hypotheses. Similar to real-life commercial ocean fishing, which requires adjusting the mesh size of the nets upward, researchers investigating multiple hypotheses need to adjust the level of statistical significance down in order to preserve the meaning of statistical significance. The general rule of thumb is to divide alpha by the number of hypotheses, referred to as a Bonferroni correction [2]. However, in accordance with Ioannidis [1], we contend this method is rarely applied. The limited use of Bonferroni correction may result from uncertainty surrounding the exact number of hypotheses "fished for," but some researchers may also be motivated to keep the "catch." In epidemiological research, upwards of 100 possible hypotheses are common, and setting the alpha at 0.0005 eliminates much of the fun in a fishing expedition. As a consequence, researchers may be tempted to discuss only a handful of "significant" results, failing to mention the 95 that proved non-significant.

Hypothesis fishing renders *P*-values almost completely meaningless, and we consider it to be a serious problem for epidemiological survey analysis. The present paper recommends a very simple countermeasure against hypothesis fishing, based on data splitting. The procedure is similar to two-stage analysis of microarray data. We argue that if our method could be applied for survey data in general, it could do for epidemiological studies what pre-experiment registration of RCTs has done for medical experiments.

The rest of the article is laid out as follows: First we review some mathematical methods that might be considered useful for counteracting hypothesis fishing. Then we explain our method in detail, and compare it to two-stage analysis of microarray data, followed by a case study where the method is applied to an analysis of low back pain. The article ends with discussion and conclusion.

Mathematical remedies (that fail to solve the problem)

If one considers hypotheses fishing to be a mathematical problem, it is reasonable to look for mathematical solutions. There is a large mathematical literature that relates to model building and multiple hypotheses, and we will only try to point out the main themes.

The simplest way of handling the problem of multiple hypotheses is to reduce the significance level through Bonferroni correction. It is also possible to use so-called false discovery rate (FDR) [3], which is a way of controlling the expected percentage of rejected null-hypotheses (discoveries) that are falsely rejected. If all the null-hypotheses are true, FDR is equivalent to Bonferroni correction, but otherwise it is less strict. As we mentioned in the introduction, it can be hard to keep track of the number of hypotheses one is really addressing, which makes the use of Bonferroni or FDR cumbersome.

Complex mathematical methods exist that adjust for the effect of model selection on inference (for example, see Madigan and Raftery [4]). In order to utilize these methods, however, one must describe the model-building strategy in a formal mathematical way. Faraway [5] has analysed various data splitting approaches similar to the one we apply, and compares them to mathematical inference adjustment methods. Based on his simulation study, he concluded that data splitting costs more in reduced accuracy than it gains in "honesty". This conclusion is to be expected, because adjustment methods utilize information that the splitting methods disregard. In our present setting, however, accurate information on modelling procedure is not available, so adjustment methods are not an option.

The task of choosing a set of hypotheses is—at least superficially—related to the problem of choosing the set of predictors to include in a statistical model for the data. An overview of this field is given by Hastie et al. [6], and we only mention a few key concepts. An obvious goal in statistical modelling of a given set of data is to develop a model that fits the data well, and a model will always fit better when the set of predictors is increased. However, if one includes too many predictors, the model is prone to

imitating random properties of the data set, which are not present in the underlying sampling distribution of the data. Therefore, one needs ways to make trade-offs between model fit and model size. This can be done directly through various information criteria, such as Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), or the Divergence Information Criterion (DIC). A different approach called *cross validation* is to split the data set in two parts, using one part for estimating model parameters, and the other for evaluating model fit. A problem with this method is that only half of the data is used for each task. A clever remedy is to leave one data point out at the time, estimate the model from the remaining data points, and measure model fit by taking the average fit on the points left out (*leave-one-out cross validation*).

We mention these methods mainly to point out that they are not very relevant for controlling hypotheses fishing. The use of information criteria or cross validation helps in trading model fit for model size, but does not produce adjustments for the total number of variables. The cross validation procedure of splitting the data set, using one part to estimate model parameters and the other for validation is deceptively similar to the method we apply. But the purpose of validating a model's predictions is entirely different from our purpose of conducting sound hypothesis testing.

In addition to the inherent difficulties related to mathematical ways of dealing with hypothesis fishing, we do not see the problem as mainly a mathematical one. Some researchers may not be aware of the fact that *P*-values lose their meaning when the hypotheses are chosen from a large pool of undocumented ones. Others may be vaguely aware of the problem, but choose not to address it unless reviewers demand it. Reviewers, on the other hand, may not feel inclined to insist on purity beyond what they have done in their own scientific work.

The solution: data splitting

We recommend splitting the data set randomly into two sections (Parts 1 and 2). This allows the investigators to identify hypotheses in Part 1 of the data set, while remaining blind to Part 2 until the hypotheses are specified. True hypothesis testing is then performed using only Part 2 of the data. At this point there is no second-guessing. If the alpha-level is set to 0.05, and the *P*-value in Part 2 is 0.051, the result is by definition not significant, even if it received a *P*-value of 0.001 in Part 1. In such cases, there will be a temptation to "make a compromise" by computing the average of the *P*-values from Parts 1 and 2, but this is not allowed. Because the data in Part 1 was used to construct the hypotheses, it is tainted, and cannot take any part in the hypotheses testing.

This procedure is strict with respect to identifying statistical significance. Once a hypothesis is supported, however, the entire data set is used for estimating the effect size. Thus, the purpose is to ensure the proper use of the term *statistical significance*. Once a significant finding is established, though, it is preferable to obtain the most accurate parameter estimates possible.

Models: hypothesis variables, and confounders

Epidemiological hypotheses are usually formulated within the framework of a model. Assume the hypothesis is that eating mushrooms increases the risk of cancer. To test this hypothesis, one would build a model with predictors like age, gender, smoking status, as well as mushroom habits, in order to control for these confounding factors. (If old people eat more mushrooms, excluding age from the model would give an incorrect positive association between mushroom eating and cancer.)

From a purely computational point of view, no differences exist between predictors associated with hypotheses and confounders, yet the semantics are very different. The confounders are included only as a means of estimating the causal link between the cause (mushrooms) and its hypothesized effect (cancer).

Researchers often use *P*-values as a road map (in addition to literature reviews and general medical knowledge), when deciding which variables should be included as confounders. A common piece of advice is to include confounders demonstrating an association at a *P*-value below 0.1 or 0.2. Despite this rationale for including a confounding factor, no similar demands are made for inclusion of the confounder in Part 2. The chosen set of confounders provides the framework within which the hypothesis is defined, and it is not the framework that is being tested. In the mushroom example, the hypothesis is that mushroom eating is associated with cancer when controlling for age, gender, and smoking, not that controlling for each of these factors is necessary.

Size of Part 1 and Part 2

When deciding upon the relative size of Parts 1 and 2, a trade-off exists between the need to identify hypotheses by exploration in Part 1, and the need to achieve statistical significance in Part 2. An even split may be reasonable in cases where the need for exploration is high, particularly if the data set is large, so that half of the data set is sufficient to achieve statistical significance for stronger effects. In cases where greater domain knowledge is available based on the existing literature, a smaller Part 1 is reasonable, especially when the sample size is small.

Multiple hypotheses

It is possible to investigate multiple hypotheses within our splitting regime, using Bonferroni corrections. Assume, in the mushroom-cancer example, that the analysis of Part 1 also provided strong support for the hypothesis that eating bananas protects against cancer. One might then choose to include both mushroom habits and banana habits as hypothesis variables, and consequently divide alpha by 2 (the number of hypotheses). If either mushroom or banana habit fails the significance test in Part 2, it will still be in the model, as a confounder.

It might be the case that both banana and mushroom habits get *P*-values that fall between alpha/2 and alpha. In this situation, both hypotheses would have passed the significance test individually, but the choice to include two hypotheses resulted in failure of both variables to reach significance. Many non-statisticians would argue that scientific procedures and statistical analysis should be objective, with conclusions based on 'hard facts', independent of arbitrary choices of hypotheses. Unfortunately, this is not the case if we wish to claim statistical significance.

The mushroom-banana example is a clear case in which investigators should reduce the alpha-level to account for multiple hypotheses. At the other extreme, if independent research groups investigating different research questions based on independent data sets, their combined effort is obviously not a case of 'multiple hypothesis testing'. A grey area exists with partially overlapping data sets, hypotheses, and research groups, often making it difficult to decide whether Bonferroni corrections are called for. A pragmatic solution may be to view a published article as a unit, and apply Bonferroni corrections within each one.

Relation to two-stage analysis in genetics

Readers who are familiar with microarray analysis will recognize that our data splitting method is similar to two-stage analysis, as it is routinely performed in genetics [7].

There are a few differences, however. In a microarray context the set of possible hypotheses is given by the number of genes, and the FDR method is normally used to limit the number of incorrect findings. Rather than primarily counteracting hypothesis fishing, microarray two-stage analysis is usually motivated by cost effectiveness: By screening out promising candidates first, and then evaluating them, researchers can make a higher number of valuable discoveries for each monetary unit spent. In a microarray setting the procedure is also likely to be more automatic, as interesting genes are filtered out in two more or less mechanical steps of analysis. In our epidemiological application, on the other hand, there will be a man-in-the-loop, as the researcher builds a model with hypothesis variables and confounders based on a combination of his domain knowledge and Part 1 of the data.

Case study of low back pain in the Ullensaker study

Study sample and setting

Data material consisted of adults enrolled in an epidemiological survey for musculoskeletal pain (MSP) in the Ullensaker municipality, 40 km northeast of Oslo in Norway. In 1990, 4050 inhabitants born in 1918–1920, 1928–1930, 1938–1940, 1948–1950, 1958–1960 and 1968–1970 (age 20–70 years) were sent a postal questionnaire about MSP. Of these people, 67% responded. Individuals who reported low back pain (LBP) during the past year (1990) were excluded from this material ($N = 1439$), such that the original sample consisted of 1283 participants who were free of LBP in 1990. In 2004, a 14-year follow-up was conducted. A total of 763 participants (59%) responded and formed the present study sample. These 763 participants were randomly divided in two samples with $n = 369$ (Part 1) and $n = 394$ (Part 2), respectively.

Outcome measures

To identify respondents with LBP, we used the answer to the question, "During the past year, have you experienced pain or discomfort in your lower back?". This item was based on the Standardised Nordic Pain Questionnaire [8], which is a self-report questionnaire frequently used in Scandinavian epidemiological studies.

Independent variables (potential risk factors)

In 1990, the survey questionnaire contained a number of socio-demographic and health-related factors, which could be included as risk factors in the present study. Socio-demographic variables were gender, age, marital status, and work status. Health-related variables were body mass index (BMI), smoking status, number of MSP sites other than the low back, duration of previous MSP, use of medication due to MSP, having been examined by a health care provider due to MSP during the last year, comorbidity, family history of musculoskeletal problems, emotional distress, leisure physical activity, participation in competitive sports, sleeping problems, and self-perceived health.

Model and hypotheses

A logistic regression model was developed based on Part 1 of the data set and medical expertise. The number of pain sites and smoking status were included as independent variables. Smoking status was dichotomised as smoking and non-smoking. Number of pain sites was operationalized using participant responses on the Nordic Pain Questionnaire [8]. Specifically, respondents reported whether they had experienced any pain or discomfort from the following 10 areas during the previous year: head, neck, shoulder, elbow, hand/wrist, upper back, low back, hip, knee and ankle/foot (responses were "yes/no"). The total number of pain sites was computed and categorized into the following four categories: no pain sites, 1 or 2 sites, 3 or 4 sites, and 5 or more pain sites.

We also included age, which was categorized into values corresponding to the six birth cohorts: 1918–1920, 1928–1930, 1938–1940, 1948–1950, 1958–1960, and 1968–1970. In addition, gender and marital status (dichotomised into married/partnership versus living alone) were included. Results of the logistic regression model are presented in Table 1.

We hypothesized that smoking would be positively associated with LBP. Therefore, a 1-tailed hypothesis test was conducted. It was also hypothesized that individual pain sites would be positively associated with LBP. To limit the number of hypotheses, though, we hypothesized that the total number of pain sites would affect LBP probability, rather than run analyses for each level of the variable.

Hypotheses testing

The significance level alpha was set to the usual value of 0.05. With two hypotheses, the critical *P*-value becomes 0.025. Results for Part 2 of the data set are illustrated in Table 2. The *P*-value of the pain sites variable (0.015) was below the critical value of 0.025, and therefore it is concluded that the number of pain sites was significantly associated with LBP at the 14-year follow-up.

Table 1 Parameter estimates from Part 1, controlling for age, gender, and marital status

Predictor	OR estimate	95% CI for OR	*P*-value
Number of pain sites[a]			0.012
1 or 2 pain sites	2.292	(1.248–4.208)	0.007
3 or 4 pain sites	2.690	(1.406–5.147)	0.003
5 or more pain sites	2.944	(1.193–7.262)	0.019
Smoking	2.079	(1.285–3.363)	0.003

[a] The reference category for number of pain sites was *no pain sites*

Table 2 Parameter estimates from Part 2, controlling for age, gender, and marital status

Predictor	OR estimate	95% CI for OR	*P*-value
Number of pain sites[a]			0.015
1 or 2 pain sites	1.328	(0.793–2.224)	0.281
3 or 4 pain sites	1.598	(0.857–2.979)	0.141
5 or more pain sites	3.941	(1.700–9.136)	0.001
Smoking	0.993	(0.627–1.571)	0.487*

[a] The reference category for number of pain sites was *no pain sites*

*1-sided *P*-value

Smoking status received a 1-sided *P*-value of 0.487, which exceeds the limit of 0.025 by a large margin, and this variable is thus deemed non-significant. This result may seem surprising, given the variable's strong association with the dependent variable in Part 1 of the data set (2-sided *P*-value = 0.003). This illustrates the dangers of hypothesis fishing: Our analysis suggests that the smoking status variable may only have been a "lucky winner" in the "*P*-value lottery" of Part 1.

Parameter estimates

Having concluded that *number of pain sites* has a significant effect on low back pain at follow up, we estimated the magnitude of the effect from the full data set (Table 3). Because the hypothesis test failed to give significance for smoking status, it is included only as a confounder together with age, gender, and marital status. The 2-sided *P*-value for smoking status was 0.040, so researchers following a hypothesis fishing procedure would probably have reported it as a statistically significant predictor.

Discussion

Our study was designed to illustrate a simple and straightforward data splitting method to counteract

Table 3 Parameter estimates of the hypotheses variable *number of pain sites*, from the complete data set controlling for age, gender, marital status and smoking status

Predictor	OR estimate	95% CI for OR	*P*-value
Number of pain sites[a]			0.000
1 or 2 pain sites	1.637	(1.116–2.400)	0.012
3 or 4 pain sites	1.983	(1.285–3.061)	0.002
5 or more pain sites	3.346	(1.846–6.067)	0.000

[a] The reference category for number of pain sites was *no pain sites*

hypothesis fishing in large-scale epidemiological surveys. This method involves splitting the data set, where the first half is used to identify hypotheses, while the remaining data is used to test the hypotheses. The data splitting procedure was illustrated using data material collected for a population-based health survey administered in Norway in 1990 and 2004. Results demonstrated that the number of pain sites ("widespreadness" of pain) was significantly associated with LBP following a 14-year follow-up. Smoking status was a strong predictor of LBP in Part 1 of the data set (hypothesis identification), but did not achieve significance in Part 2 (hypothesis testing). Therefore, this finding was dismissed as non-significant in our study. For the full data set, the *P*-value for smoking status was 0.040, so the traditional way of analysing epidemiological data would have given a different conclusion.

In this study, the investigators had free access to the entire data set prior to data splitting and during model development. However, any temptation to "peak" at the material was successfully avoided, as indicated by the discrepant results for smoking status. This indicates that the data splitting procedure can indeed function properly in the absence of strict external control of the data. Nevertheless, we recommend that the data set be handled by an independent party, so that researchers can document claims that only Part 1 of the data set was used for model and hypothesis development.

Ideally, the establishment of an independent international body is recommended to manage splitting of survey data. A fixed date for releasing data for Part 2 would be agreed upon, so that only those hypotheses specified prior to the release date would undergo a true significance test. Although several challenges and practical issues would inevitably need resolution (i.e., data collection, confidentiality, release of data), such an organization should be feasible and acceptable to the scientific community.

Conclusions

Results demonstrated that the number of musculoskeletal pain sites significantly predicts low back pain at a 14-year follow-up, when controlling for age, gender, marital status, and smoking. The application of the data splitting method in our study indicates its potential as an effective and useful method to counteract hypothesis fishing in population surveys. In our opinion, systematic data splitting administered by an independent party would accomplish for statistical surveys what pre-registration has already done for clinical trials.

References

1. Ioannidis JPA. Why most published research findings are false. PLoS Med 2005;2:696–701.
2. Abdi H. Bonferroni, Sidak corrections for multiple comparisons. In: Salkind NJ (ed) Encyclopedia of Measurement and Statistics. Thousand Oaks CA: Sage; 2007.
3. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J Roy Stat Soc Ser B-methodological 1995;57(1):289–300.
4. Madigan D, Raftery AE. Model selection and accounting for model uncertainty in graphical models using Occam's window. J Am Stat Assoc 1994;89:1535–46.
5. Faraway JJ. Data splitting strategies for reducing the effect of model selection on inference. Comput Sci Stat 1998;30:332–41.
6. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer, 2001.
7. Satagopan JM, et al. Two-stage designs for gene-disease association studies. Biometrics 2002;58:163–70.
8. Kourinka I, Johnsson B, Kilbom A, et al. Standardized Nordic questionnaires for the analysis of musculoskeletal symptoms. Appl Ergon 1897;18:233–7.