Problems in orthopedic research

Dependent observations

Jonas Ranstam

Novo Nordisk A/S, Krogshøjvej 51, DK-2880 Bagsværd, Denmark. E-mail: jra@novonordisk.com
Submitted 02-01-08. Accepted 02-03-05

Repeated measurements on the same subject are common in orthopedic studies; several reports include statistical analyses on x knees, y hips or z finger joints. Counting hips, knees and finger joints instead of subjects may seem attractive to the investigator because statistical significance is more easily achieved as n, the sample size, becomes larger.

However, multiple measurements on a subject are typically more similar than measurements on different subjects because they have more in common. For example, loosening of arthroplasties may be affected by host factors like age, sex, weight, bone characteristics, physical activity, etc. Loosening times for bilateral arthroplasties in hosts may then, for each host, be more alike than those for unilateral arthroplasties in different hosts.

This is a problem because in traditional statistical analyses p-values and confidence intervals are generally calculated on the assumption that observed variability represents variability between independent observations. If we replace between-subject measurements (independent observations) with within-subject measurements (repeated observations), we will probably underestimate variability. Results of such analyses should rightly be regarded with scepticism.

To solve the problem, two main options are available. First, we can in fact use traditional statistical methods, but we must recognize dependencies in data and use the methods properly. An example of sound and flawed analyses using common t-tests is given below.

Secondly, we can use modern methods of statistical analysis specifically designed for analyzing data characterized by more than one source of variability. Computer programs for such analyses are available in many commercial statistics software packages. This is a more complex alternative but it has many advantages (see Brown and Prescott 1999, Pinheiro and Bates 2000) and readers of scientific medical journals will probably be increasingly exposed to papers presenting results based on these methods. The example is revisited with this methodology.

Example

Two types of knee prostheses increase bone mineral content (BMC) at the implantation site. We wish to find out whether the increase in bone mineral content differs between the two prostheses. A dataset with 10 patients and 15 knees is available for statistical analysis, see Table 1. We assume that the increase in bone mineral content is normally distributed.

Student's t-test is often inadequately used in this situation.

Flawed analysis

Ignoring that the sample has only 10 independent observations and testing the two groups of 7 and 8 knees (pretending that n = 15) with Student's t-test will show that prosthesis type 2 increases bone mineral content by 2.3% units more than type 1 ($p < 0.05$). The test is based on a variance estimate of 4.0% units.

The reason why this analysis is flawed is presented below.

Sound analyses

Traditional methods. The Student's t-test can be

Table 1. Increase in bone mineral content in 15 knees and 10 patients

ID[a]	PROSTH[b]	BMC[c]
1	1	4.2
1	1	2.9
2	1	2.1
3	1	3.3
4	1	4.5
4	1	5.1
5	1	8.7
6	2	6.1
7	2	9.9
8	2	5.4
8	2	3.7
9	2	6.1
9	2	6.3
10	2	8.1
10	2	7.7

[a] Patient id (1–10)
[b] Prosthesis type (1, 2)
[c] Increase in bone mineral content (%)

described with a statistical model (1) for the i patients (here i = 1, ... 10), with a common mean value, μ, a fixed factor measuring the different effects of each of the j prosthesis types, $PROSTH_j$ (j = 1, 2) and error terms, ε_{ij}, with independent and normally distributed errors having mean = 0 and variance = σ^2, in short: N (0, σ^2). All variance is assumed to be between patients, and is also assumed to be identical for the different types of prostheses (the two patient groups).

$$BMC_{ij} = \mu + PROSTH_j + \varepsilon_{ij} \quad (1)$$

If patients had had both types of prostheses, they could have been their own controls—i.e., comparisons could have been performed within subjects, without contamination of between-subject variance. To adapt the model to this situation, a fixed factor, ID_i, taking care of between-patient variation should be included. This model (2) corresponds to a paired t-test.

$$BMC_{ij} = \mu + ID_i + PROSTH_j + \varepsilon_{ij} \quad (2)$$

However, in the present example, the dataset consists of measurements on patients with both one and two prostheses (but none has both types of prostheses). Neither of the models will therefore be directly applicable.

To analyze the data adequately, two options are available. First, we can remove dependencies in the dataset (i.e., eliminate within-patient variation) by reducing data to one observation per patient. It could, for instance, be adequate to replace a patient's two measurements by their mean value, see Table 2. A Student's t-test would then be appropriate.

Testing the two groups of 5 subjects each, using the subject's mean BMC value (for patients with more than one prosthesis) shows that prosthesis type 2 increases bone mineral content by 2.3% units (p = 0.13) relative to type 1. This test is based on a variance estimate of 5.3 units.

Mixed models. Secondly, as a more attractive solution to the problem, we can use a model allowing k measurements per subject (here k = 1, 2) with one variance component for between-patient variation and another one for within-patient variation. The model would be similar to (2), but with one major exception: the ID_i terms can no longer be fixed, since that would eliminate between-subject variance. Instead, they have to be estimated from data by modeling between- and within-patient variance simultaneously. Such a factor, unlike the previously discussed fixed factors, is known as random.

Hence, let ID_i^R be a random factor N (0, σ_b^2), where σ_b^2 represents between-patient variance while ε_{ijk} remains error terms N (0, σ^2). The resulting model (3) is mixed; it includes both a fixed and a random factor.

$$BMC_{ijk} = \mu + ID_i^R + PROSTH_j + \varepsilon_{ijk} \quad (3)$$

In contrast to models (1) and (2), the mixed model (3) does not require independent observations; it allows measurements on the same subject to be more similar than measurements on different subjects. This relationship can be described as intraclass correlation; the intraclass correlation coefficient, ICC, is defined as

$$ICC = \sigma_b^2 / (\sigma_b^2 + \sigma^2)$$

An analysis of all 15 knees using a mixed model (results from the statistical package R are presented here) yields variance component estimates

Table 2. Data structure in the statistical models discussed

Possible data combinations			Statistical model [a]		
Pat (*i*)	Type (*j*)	Obs no. (*k*)[b]	1	2	3
1	1	1	3.6	4.2	4.2
1	1	2	–	–	2.9
1	2	1	–	missing	–
1	2	2	–	–	–
2	1	1	2.1	2.1	2.1
2	1	2	–	–	–
2	2	1	–	missing	–
2	2	2	–	–	–
3	1	1	3.3	3.3	3.3
3	1	2	–	–	–
3	2	1	–	missing	–
3	2	2	–	–	–
4	1	1	4.9	4.5	4.5
4	1	2	–	–	5.2
4	2	1	–	missing	–
4	2	2	–	–	–
5	1	1	8.7	8.7	8.7
5	1	2	–	–	–
5	2	1	–	missing	–
5	2	2	–	–	–
6	1	1	–	missing	–
6	1	2	–	–	–
6	2	1	6.1	6.1	6.1
6	2	2	–	–	–
7	1	1	–	missing	–
7	1	2	–	–	–
7	2	1	9.9	9.9	9.9
7	2	2	–	–	–
8	1	1	–	missing	–
8	1	2	–	–	–
8	2	1	4.6	5.4	5.4
8	2	2	–	–	3.7
9	1	1	–	missing	–
9	1	2	–	–	–
9	2	1	6.2	6.1	6.1
9	2	2	–	–	6.3
10	1	1	–	missing	–
10	1	2	–	–	–
10	2	1	7.9	8.1	8.1
10	2	2	–	–	7.7

[a] Statistical models:
Model 1. Student's t-test: each subject contributes one observation. The mean value is used if there are two measurements on a patient .
Model 2. Paired t-test: each patient contributes a pair of measurements; their difference is used in the analysis. Since patients in this example do not have both types of prosthesis (indicated by 'missing'), the model can not be used.
Model 3. Mixed model approach: different number of observations per subject can be included in a mixed model since this models two sources of variation. This does not hold for Student's t-test.
[b] This index is redundant in models 1 and 2.

of $\sigma_b^2 = 4.8$ and $\sigma^2 = 0.5$; ICC = 4.8/(4.8+0.5) = 0.9. The additional increase in bone mineral content in prosthesis type 2 in this model is estimated at 2.4% units (p = 0.13).

Discussion

The identical p-values in the t-test, using aggregated data, and the mixed model are coincidental, but similar results should reasonably be expected.

In a comparison of the results of the three analyses, it is clear that the flawed t-test under-estimates the variability of data and produces too low a p-value, which is falsely significant.

The sound t-test produces the estimate $\sigma^2 = 5.3$, which is somewhat higher than the estimated between-subject variance in the mixed model above (it also includes some within-subject variation). However, in the flawed t-test $\sigma^2 = 4.0$, which is even lower than the mixed model estimate of between-subject variance $\sigma_b^2 = 4.8$. This shows the consequence of mixing within- and between-subject variance.

The results of the two sound analyses do not differ much. In this example, the two models can be considered equally good. In more complex analyses—e.g., requiring covariates—a mixed model might be the only alternative.

Disregarding repeated measurements, however, does not necessarily always lead to erroneous results. For instance, two research groups (Ripatti and Palmgren 2000, Schwarzer et al. 2001) independently studied loosening of hip prostheses in Germany and Finland. Both conclude that the bilateral prostheses they studied could have been treated as independent observations.

Statistical packages like SAS, S-plus, R, and SPSS contain programs for mixed models. The commands for the analysis in the example presented are shown in the Appendix. However, the analysis of data using mixed models is not easy; consultation of a professional statistician is recommended.

Conclusion

Repeated measurements on subjects should be recognized and analyzed adequately. Sound analyses can often be done using traditional methods of

analysis. On the other hand, methods specifically designed for the purpose are usually available.

Brown H. Prescott R. Applied mixed models in medicine. John Wiley & Sons, New York 1999.

Pinheiro J, Bates D M. Mixed-effect models in S and S-PLUS. Springer New York 2000.

Ripatti S, Palmgren J. Estimation of multivariate frailty models using penalized partial likelihood. Biometrics 2000; 56: 1016-22.

Schwarzer G, Schumacher M, Maurer T B, Ochsner P E. Statistical analysis of failure times in total joint replacement. J Clin Epidemiol 2001; 54: 997-1003.

Appendix

Syntax for the mixed model used in the example for the statistical packages SAS, S-plus, R, and SPSS. Variable names are in bold font.

SAS

```
proc mixed;
   class id prosth;
   model BMC = prosth / ddfm = satterth;
   random int/sub = id
```

S-plus and R

```
summary(lme(BMC ~ prosth, data = "data-
set", random = ~1 | id))
```

SPSS

```
mixed BMC by prosth
  / fixed = prosth
  / random = intercept| subject(id).
```

Note that these packages do not use identical algorithms for estimating random effects; their results might differ slightly, especially with small data sets.