Cardiovascular research: data dispersion issues

Ton J. Cleophas, Aeilko H. Zwinderman, Roya Atiqi, Joan van de Bosch

European Interuniversity College Pharmaceutical Medicine, Lyon, France

Abstract

Biological processes are full of variations and so are responses to therapy as measured in clinical research. Estimators of clinical efficacy are, therefore, usually reported with a measure of uncertainty, otherwise called dispersion. This study aimed to review both the flaws of data reports without measure of dispersion and those with over-dispersion.

Examples of estimators commonly reported without a measure of dispersion include:

1) number needed to treat;
2) reproducibility of quantitative diagnostic tests;
3) sensitivity/specificity;
4) Markov predictors;
5) risk profiles predicted from multiple logistic models.

Data with large differences between response magnitudes can be assessed for over-dispersion by goodness of fit tests. The χ^2 goodness of fit test allows adjustment for over-dispersion.

For most clinical estimators, the calculation of standard errors or confidence intervals is possible. Sometimes, the choice is deliberately made not to use the data fully, but to skip the standard errors and to use the summary measures only. The problem with this approach is that it may suggest inflated results. We recommend that analytical methods in clinical research should always attempt to include a measure of dispersion in the data. When large differences exist in the data, the presence of over-dispersion should be assessed and appropriate adjustments made.

Correspondence: Ton J. Cleophas, Department Statistics, Circulation, Boston MA, c/o Department Medicine, Albert Schweitzer Hospital, Box 444, 3300 AK, Dordrecht, Netherlands.
E-mail: ajm.cleophas@wxs.nl

Key words: clinical research, uncertainty, standard error, confidence interval, sensitivity, specificity, reproducibility, Markov model, numbers needed to treat, logistic models, risk profiles, over-dispersion, variance inflating factor.

Received for publication: 27 January 2010.
Revision received: 28 January 2010.
Accepted for publication: 19 May 2010.

Heart International 2010; 5:e9
doi:10.4081/hi.2010.e9

Introduction

Biological processes are complex and therefore full of variations. Clinical responses as captured in clinical research are, therefore, equally complex. Statistics give no certainties, only chances and, consequently, their results are often reported with a measure of dispersion, otherwise called uncertainty. Generally, standard errors are calculated as a measure for dispersion in the data. For example, in a hypertension study a mean systolic blood pressure after active treatment of 125 mm Hg compared to 135 mm Hg after placebo treatment may indicate that either the treatment was clinically efficacious or that the difference observed is due to random variation. To address this important distinction, the standard errors of the mean are found to be 5 mm Hg each, and a pooled standard error is calculated: $\sqrt{(5^2+5^2)} = 7.07$ mm Hg. According to the Student's t-test this result is statistically insignificant: the t-value = (135-125)/7.07=1.4, and should have been larger than approximately 2. With such a result it is, usually, concluded that the treatment effect does not differ from a placebo effect, and that the calculated mean difference is due to random variation rather than a true treatment effect.

It is sometimes difficult to assess complex estimators of clinical efficacy for standard errors. Consequently, they are then reported as mean results without further statistical test or P value. For example, number needed to treat (NNT) in clinical trials, reproducibility of quantitative diagnostic tests, sensitivity and specificity, Markov estimators, and risk profiles from multiple logistic models are routinely reported without measure of uncertainty. Clinical decisions made from such estimators are, therefore, not entirely in agreement with evidence-based medicine. As recommended by the Standards for the Reporting of Diagnostic Accuracy Studies (STARD) steering group,[1] ample efforts should be given to include a measure of uncertainty in any research result in order for predictions to be more accurate.

Another dispersion issue is the use of traditional standard errors in situations where the data are over-dispersed. Over-dispersion describes the phenomenon in which the spread in the data is wider than compatible with Gaussian modeling. This phenomenon is particularly common with logistic models but can also occur with continuous real data samples.[2] Traditional statistical tests overestimate the precision of over-dispersed data, meaning that the calculated P values are too small, potentially supporting the erroneous conclusion of a significant effect. To date, statistical software programs do not routinely include tests for over-dispersion. Therefore, it is incumbent upon investigators to recognize over-dispersion and make their own assessments prior to the analysis.

In the current paper, we will review both the flaws inherent in data without a measure of dispersion and that of data with over-dispersion. As there are almost no real data examples assessing these flaws in the cardiovascular literature, we will give hypothetical examples. Simple Gaussian distribution based methods for assessment are used, and most of them can be readily found in major statistical packages like SAS,[3] and special software programs for the calculation of confidence intervals like Confidence Interval Analysis.[4]

Data without measure of dispersion

Number needed to treat in clinical trials

In order to decide whether the results of a study are important for future patient care the

NNT is often calculated. As an example, in a clinical trial of β-blocker versus placebo for the prevention of post-infarct arrhythmias, the proportion of post-infarct arrhythmias is significantly lower with the β-blocker than with placebo, with 51 of 748 (6.8%) in the β-blocker group and 126 of 764 (16.8%) in the placebo group (relative risk 2.4, 95% confidence interval 1.8-3.3). With this result, it is interesting to extrapolate these results to future populations. The number needed to treat in order to prevent one arrhythmic patient is often used for that purpose, and is calculated according to:
NNT =1/ (0.168-0.068) =1/0.1=10

We will need to treat 10 patients with a β-blocker in order to prevent one arrhythmic patient. This conclusion, however appealing to readerships of articles, is not justified, because it is based on the assumption that the proportions are 100% certain, but the proportions do have boundaries of uncertainty, the 95% (or 99%) confidence interval, which indicates that the number could differ considerably from 10.

Using the equation "proportion ±2√ [p (1-p) /n]" the 95% confidence intervals are calculated as follows:
0.068 is between 0.051 and 0.085
0.168 is between 0.126 and 0.210

If we include this uncertainty in the calculation of the NNT, then we can be 95% sure that the numbers required to prevent one arrhythmic patient range between 1/(0.210-0.051) =6.3 and 1/(0.126-0.085) =24.4. As we consider the treatment of future patients, it is more accurate to think of NNT between 6 and 25 instead of an NNT of 10 patients. We should add that the NNT can also be derived from the risk difference. The risk difference and its 95% confidence interval can be calculated in SAS,[3] Confidence Interval Analysis,[4] and other software programs.

Reproducibility of quantitative diagnostic tests

Reproducibility, otherwise called reliability, of diagnostic tests or questionnaires is an essential prerequisite for implementation. A routine but incorrect method for that purpose is the following. We calculate the mean value of the first set of tests and then from the second set of tests. If the difference is small, then we conclude that the results of the two tests are reproducible. As an example, in a diagnostic study of patients with Raynaud's phenomenon, the reliability of venous occlusion plethysmography is assessed by duplicate testing of 6 patients (Table 1). The mean difference between the duplicate tests is as small as 0. Yet, the test is poorly reproducible, with a range of differences between two tests of no less than -11 to +10 mL/min.

The mean difference between two sets of tests is, obviously, not good enough for demonstrating a high level of reproducibility between tests. More adequate for that purpose are methods that assess the spread of differences between repeated measurements like, for example, the duplicate standard deviation (Table 2). For adequate reproducibility the magnitude of the duplicate standard deviation should equal 10-20% of the averages of the test results. Also adequate is the repeatability coefficient that is calculated by the standard deviation of the individual differences between tests 1 and 2: a result equal to 10-20% of the test averages is considered to be adequate.

Sensitivity and specificity

In clinical research, use of a "gold standard" test for making a diagnosis is often laborious and, sometimes, impossible. Frequently, simple and non-invasive tests are used.

Disease Present	Yes	No
Test positive	a	b
Test negative	c	d

In the above 2×2 contingency table: a = the number of truly positive patients in such a simple non-invasive test, b = the number of false positives, c = the number of false negatives, and d = the number of truly negatives.

Validity of these kinds of tests is often assessed with sensitivity and specificity. Sensitivity = a / (a+c) = proportion in a sample of true positive patients, where the true positives are the patients with a positive test and the presence of disease; specificity = d / (d+b) = proportion of a sample of patients with a true negative test, where the true negatives are the patients with a negative test and without the presence of disease. However, most diagnostic tests have limited sensitivities and specificities. Levels around 0.5 (50%) mean that no more information is gained than by flipping a coin. Levels substantially higher than 50% are commonly accepted as documented proof that the diagnostic test is valid. However, sensitivity/specificity are estimates from experimental samples, and scientific rigor requires that the amount of uncertainty be included in the results of experimental sampling. Uncertainty is virtually never assessed in sensitivity/specificity evaluations of cardiovascular diagnostic tests. This is unfortunate as calculated levels of uncertainty could be used with statistical testing whether the sensitivity/specificity are significantly larger than 0.5 or whether their 95% confidence intervals are between previously set validation boundaries. If not, then it is appropriate to reject the diagnostic test, because it is too imprecise to predict the disease. As an example, a d-dimer test is used as a diagnostic test for the diag-

Table 1. In a diagnostic study of patients with Raynaud's phenomenon the reliability of venous occlusion plethysmography is assessed by duplicate testing of 6 patients.

Plethysmographic peripheral arterial flows (mL/min)			
Patient	Test 1	Test 2	difference
1	1	11	-10
2	10	0	10
3	2	11	-9
4	12	2	10
5	11	1	10
6	1	12	-11
Mean difference			0

Table 2. More adequate for assessing reproducibility between tests are methods that assess the spread of differences between repeated measurements, e.g. the duplicate standard deviation.

Plethysmographic peripheral arterial flows (mL/min)				
Patients	Test 1	Test 2	difference(d)	$(difference)^2$
1	1	11	-10	100
2	10	0	10	100
3	2	11	-9	81
4	12	2	10	100
5	11	1	10	100
6	1	12	-11	121
Averages	6.17	6.17	0	100.3

Duplicate standard deviation = $\sqrt{\frac{1}{2}\Sigma d^2/n} = \sqrt{(1/2 \times 100.3)} = 7.08$

nosis of pulmonary emboli.

Emboli	Yes	No
	(n. of patients)	
D-dimer Result		
Positive	2	18
Negative	1	182

The sensitivity and specificity in the above example is calculated to be 0.6666 and 0.911, respectively. These results could be interpreted as acceptable, because they are much larger than 0.5. However, in order to conclude that they are significantly larger than 0.5 their 95% confidence intervals should not cross the 50% boundary. Sensitivity / specificity are proportions and it is fairly straightforward to calculate standard errors from them.[5]

The equations are:

standard error sensitivity = $\sqrt{ac / (a+c)^3}$

standard error specificity = $\sqrt{db / (d+b)^3}$

The 95% confidence intervals of sensitivity and specificity can be calculated from:

95% confidence interval = sensitivity ±1.96× standard error

95% confidence interval = specificity ±1.96× standard error

sensitivity = 0.666±1.96×3.672 = between -5.4 and 7.8.

specificity = 0.911±1.96×0.286 = between 0.35 and 1.47.

These intervals are very wide and do not fall within the boundary 0.5 to 1.0 (50-100%). Thus, the sensitivity and specificity are insufficient.

Markov predictors

Regression models are only valid within an observed range of values. The Markov model goes one step further. It predicts beyond that range and, in addition, it does so without accounting for uncertainty. As an example, in an observational study the presence of heart failure defined as a B-natriuretic peptide test above 100 pg/mL is assessed in a group of 500 patients. At time 0 year none of 500 patients met the criterion. After one year, 50 of 500 (10%) had a positive test. An exponential pattern is assumed. It is concluded that, if after one year 90% had no heart failure, then:

after 2 years 90%×90%=81% will have no heart failure

after 3 years 90%×90%×90%=73% will have no heart failure

after 6.7 years = 50% will have no heart failure

Markov models are very popular for making predictions from health statistics or population-based studies like the Framingham studies. It is obvious that such models would be more accurate if uncertainty were included. Markov models using multiplication of proportions and standard errors of them can be calculated using a logarithmic transformation.[6] The natural logarithm of a proportion is given by ln [a/(a+b)]. We recommend that the standard error be approached from the equation (ln = natural logarithm):

standard error ln [a/(a+b)] = 1/a – 1/(a+b)

From the previously example, the 95% confidence interval of the proportion of patients who will have no heart failure after 6.7 years could be calculated:

ln [a/(a+b)]$^{6.7}$=6.7 ln [a/(a+b)]

The standard error of ln [a/(a+b)]$^{6.7}$=6.7 standard error [a/(a+b)] =6.7 [1/a–1/(a+b)].

The logarithmic transformed 95% confidence interval =

6.7 ln [a/(a+b)] ±1.96×6.7 [1/a–1/(a+b)].

The true 95% confidence interval is found by taking the antilogarithm.

Therefore uncertainty can be included in a Markov model resulting in more precise predictions from this clinical estimator.[6]

Risk profiles from multiple logistic models

Logistic models are often applied for determining individual and population risk profiles. As an example, we will use an observational study of myocardial infarction in females treated with estrogen. Additional risk factors are included (Table 3). The odds of myocardial infarct in patients with estrogen is 13.5 times that of patients without. As 4 of the risk factors are significant, we remove factor 5 and assume that all of the remaining 4 factors independently predict an increased risk and that, together, they predict the following risk:

the odds ratio (OR) of myocardial infarct with factors 1-4 = $OR_1 \times OR_2 \times OR_3 \times OR_4 = 75.9$

For an individual or a group of persons carrying all 4 risk factors the odds of suffering a myocardial infarction is 76 times that of the individual / group devoid of the risk factors. But is this true? Should we not include a boundary of uncertainty here? The standard error of each of the risk factors is given in Table 3 and needs to be incorporated in the final result for the purpose of accuracy and precision.

Logistic models for determining risk profiles use multiplications of odds ratios. If only significant predictors are included, we may assume that they are independent of one another and a fairly straightforward method is available for calculating the pooled 95% confidence interval of the multiplication products. The above example is used once more.

The pooled standard error of the natural logarithms of the odds ratio with the factors 1-4 (ln $OR_{\text{factors 1 - 4}}$) is given by (ln means natural logarithm):

standard error of ln $OR_{\text{factors 1 - 4}}$ =

$\sqrt{}$ (standard $error_1^2$ + standard $error_2^2$ + standard $error_3^2$ + standard $error_4^2$)

The logarithmic transformed 95% confidence interval is found by taking:

ln $OR_{\text{factors 1 - 4}}$ ± 1.96 × pooled standard error of ln $OR_{\text{factors 1 - 4}}$

In this way, uncertainty can be implied in the risk profile and better precision for predictions from data can be provided.

Data with over-dispersion

Over-dispersion depicts the phenomenon where the spread in the data is wider than compatible with Gaussian modeling. This phenomenon is particularly common with logistic models, but can also occur with continuous real data samples.[2] Over-dispersion can be detected by goodness of fit tests, for example the Pearson's χ^2 goodness of fit test or the Kolmogorov-Smirnov test.[7] To date statistical software programs do not routinely include tests for over-dispersion. Thus, investigators have to make their own assessments prior to the analysis.

Table 4 shows a hypothesized example of a 2 × 2 multicenter factorial clinical trial of the effect of a β-blocker and a calcium channel blocker on hypertension. The analysis requires the binary logistic model (ln = natural logarithm):

ln odds of responding = a + $b_1 \times_1$ + $b_2 \times_2$ + $b_3 \times_1 \times_2$

$\times_1$ = β-blocker

$\times_2$ = calcium channel blocker

There is a strong difference in the total numbers of observations per center: between 4 and 81. This could lead to over-dispersion and the Pearson's goodness of fit test can be used to assess the presence of it. The calculation is given in Table 5. If we add up the other 3 treatment combination results to 10.0, we will end up with a χ^2 value of 10.0 +..... = 32. This χ^2

Table 3. Multiple logistic regression of an observational study of myocardial infarction in females treated with estrogen. The dependent variable is the myocardial infarction (yes/no), estrogen use (yes/no), and the other predictors below are included in the model.

Risk factors	Regression coefficient(b)	Standard error	P	Odds ratio
1.Estrogen	2.60	0.25	<0.0001	13.5
2.Cholesterol	0.81	0.21	0.0001	2.2
3.Obesity	0.50	0.25	0.04	1.6
4.Hypertension	0.42	0.21	0.05	1.5
5.Nicotine	0.53	0.53	ns	

Table 4. A hypothesized example of a 2×2 multi-center factorial clinical trial of the effect of a β-blocker and a calcium channel blocker on hypertension.

	Calcium channel blocker				Dummy calcium channel blocker			
	Dummy b-b		β-blocker		Dummy b-b		β-blocker	
Center	Resp	Total	Resp	Total	Resp	Total	Resp	Total
1	10	39	5	6	8	16	3	12
2	23	62	53	74	10	30	22	41
3	23	81	55	72	8	28	15	30
3	26	51	32	51	23	45	32	51
4	17	39	46	79	0	4	3	7
5			10	13				
Mean proportion per treatment combination								
	0.364		0.681		0.249		0.532	

b-b = β-blocker. Resp = number of responders (mean blood pressure <107 mm Hg). Total = total n. patients with specific treatment combination per center.

Table 5. The Pearson's goodness of fit test of the data of Table 4.

$\chi^2 = \Sigma$ (observed n. responders - expected n. responders)^2 expected n. responders

The calculation per treatment combination per center is as follows	
1. (10- 39×0.364) / 39× (10/39- (1-10/39))	=2.2
2.	=0.0
3.	=2.2
4.	=4.5
5.	=1.1 +
	=10.0

Table 6. The calculation of a standard error adjusted for over-dispersion.

		SE	P	SE_{adust}	P
a	-0.41	0.18	0.025	0.25	0.119
b_1	0.54	0.25	0.031	0.34	0.132
b_2	-0.15	0.22	0.513	0.30	0.638
b_3	0.78	0.31	0.011	0.42	0.080

SE adjust = SE adjusted for over-dispersion = $\sqrt{1.9} \times$ SE.

value should be approximately equal to its degrees of freedom for the logistic model to hold. We have, however, 21 (cells) – 4 (treatment combinations) = 17 degrees of freedom. This would mean that the data are over-dispersed. A solution recommended by Hojsgaard and Halekoh is used.[8] The magnitude of the dispersion can be estimated by the ratio:

χ^2 number / degrees of freedom = 32/17=1.9

The square root of this ratio (here, $\sqrt{1.9}$) , sometimes called the variance inflating factor can, subsequently, be used to adjust the standard errors in the study. ln odds of responding = $a + b_1 \times_1 + b_2 \times_2 + b_3 \times_1 \times_2$ (ln = natural logarithm). The calculation is given in Table 6.

The probability of responding to a dummy β-blocker and calcium channel blocker equals 0.36. This is unchanged after adjustment for dispersion. However, the 95% confidence interval changes from 0.31-0.42 to 0.28-0.45. In conclusion, with over-dispersion the parameter estimates are not affected but their standard errors are likely to be underestimated and should be adjusted to compensate for that flaw.

Discussion

This review is far from complete, many more examples can be given. Data without measure of dispersion also include pharmacokinetic/pharmacodynamic parameters in simulated and real-data drug trials, diagnostic odds ratios in diagnostic meta-analyses,[9] node impurities with binary partitioning,[10] propensity scores for data matching.[11] Also data with over-dispersion are very common with current multicenter and international cardiovascular trials, though rarely assessed for that purpose.[2]

Conclusions from data without measure of dispersion should be interpreted with caution because statistically insignificant differences may be interpreted as real differences while they are just a result of random fluctuations. Random fluctuations should never be the basis for new treatments. The STARD working party recently recommended "to include in the estimates of diagnostic accuracy adequate measures of uncertainty, e.g., 95%-confidence intervals",[1] and rightly so, because the problem is not sporadically encountered but can be almost routinely observed in research reports. For example, even in a journal like the *Journal of the International Federation of Clinical Chemistry and Laboratory Medicine* out of 17 original papers addressing novel chemistry methods, 16 communicated the above-mentioned flawed reproducibility assessments while the correct methods were used in only one.[12]

What solutions can be given? First, calculating standard errors or confidence intervals is often possible. If not, alternative confidence intervals may be a possibility, for example, those based on Monte Carlo methods like bootstrap confidence intervals. Second, sometimes the choice is deliberately made not to use the data fully but to skip the standard errors and to use the summary measures only. Number needed to treat can be considered as such a summary measure. The problem with this approach is that without accounting for the uncertainty of the summary measure the overall results may produce inflated results because the dispersion in the data is artificially minimized by removing this uncertainty. This limitation should be recognized in research reports.

Conclusions from data with over-dispersion should also be interpreted with caution because the calculated confidence intervals and P values are too small and the conclusion of a significant effect may be erroneously made. The presence of over-dispersion should be assessed particularly if a strong difference in numbers of responders or magnitudes of responses is in the data. Goodness of fit tests are available for that purpose. The advantage of the Pearson's χ^2 goodness of fit test is that, in addition to detecting over-dispersion, it enables adjustment for it. The adjusted mean of the data remains unchanged while the measures of dispersion in the data, including variances and co-variances, log - likelihoods, Wald – intervals, etc. are simply multiplied by the square root of the ratio of the χ^2 value and its degrees of freedom (variance inflating factor = χ^2 / degrees of freedom).

In conclusion, we recommend that analytical methods in clinical research should always try to include a measure of dispersion in the data. Often standard errors or 95% confidence intervals can be used for the purpose. With large differences in the data, the presence of over-dispersion should be assessed and appropriate adjustments made.

References

1. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. BMJ 2003;326:41-4.
2. Tan M. Describing data, variability and over-dispersion in medical research. In: Advanced Medical Statisics. Eds: Lu Y, Fang J. World Scientific, New Jersey, USA, 2003, pp 319-32.
3. SAS Statistical Software, SAS Intitute, Cary, NC, USA. www.sas.com
4. Gardner MJ. Confidence Interval Analysis. BMJ Productions, London, UK, 1989.
5. Levin MD, Van de Bos E, Van Ouwerkerk BM, et al. Uncertainty of diagnostic tests. Perfusion 2008;21:42-8.

6. Cleophas TJ, Zwinderman AH. Markow modeling. In: Statistics applied to clinical trials, 4th edition. Springer, Dordrecht, Netherlands, 2009, pp 212-3.
7. Cleophas TJ, Zwinderman AH. Testing clinical trials for randomness. In: Statistics applied to clinical trials, 4th edition. Springer, Dordrecht, Netherlands, 2009, pp 355-66.
8. Hojsgaard S, Halekoh U. Overdispersion. Danish Institute of Agricultural Sciences, Copenhagen, June 1, 2005. http://gbi.agrsci.dk/statistics/courses
9. Moses LE, Shapiro D, Littenberg B. Combining independent studies of a diagnostic test into a summary ROC curve: data-analytic approaches and some additional considerations. Stat Med 1993;12:1293-316.
10. Lesterhuis W, Cleophas TJ. Cardiovascular research: decision analysis using binary partitioning. Perfusion 2009;22:88-91.
11. Wasson JH, Sox HC, Neff RK, Goldman L. Clinical prediction rules. Applications and methodological standards. N Engl J Med 1985;313:793-9.
12. Imbert-Bismut F, Messous D, Thibauet V, et al. Intra-laboratory analytical variability of biochemical markers of fibrosis and activity and reference ranges in healthy blood donors. Clin Chem Lab Med 2004;42:323-33.