malignant lymphomas, and 1·29 (1·23 to 1·35) for all tumour types.

Comment

We found no trend in risk of childhood leukaemia over the three defined birth cohorts. The low increase in risk seen for all tumour types combined since the 1940s has been described before[5] and is due mainly to continuous increases in risk of lymphomas in boys and of neuroblastomas in both sexes.

The comparison group in this study comprised children who were born before vitamin K was routinely given to the mother or child. Thus the trend in risk for childhood cancer, and for leukaemia in particular, seems to have been unaffected by the widespread use of intramuscular vitamin K since the early 1970s. Our findings are not consistent with those of Golding *et al.*[1] The findings would agree if the prevalance of a strong risk factor for leukaemia were decreasing in parallel to the increasing use of intramuscular vitamin K. This, however, seems unlikely.

We thank B Carstensen for help with statistical testing and A Bautz for help with computing.

1 Golding J, Greenwood R, Birmingham K, Mott M. Childhood cancer, intramuscular vitamin K, and pethidine given during labour. *BMJ* 1992;305: 341-6.
2 Draper GJ, Stiller CA. Intramuscular vitamin K and childhood cancer. *BMJ* 1992;305:709.
3 Ekelund H, Finnström O, Gunnarskog J, Källén B, Larsson Y. Administration of vitamin K to newborn infants and childhood cancer. *BMJ* 1993;307:89-91.
4 Klebanoff MA, Read JS, Mills JL, Shiono PH. The risk of childhood cancer after neonatal exposure to vitamin K. *N Engl J Med* 1993;329:905-8.
5 Brown PdN, Hertz H, Olsen JH, Yssing M, Scheibel E, Jensen OM. Incidence of childhood cancer in Denmark 1943-1984. *Int J Epidemiol* 1989;18:546-55.

(Accepted 7 February 1994)

Statistics Notes

Correlation, regression, and repeated data

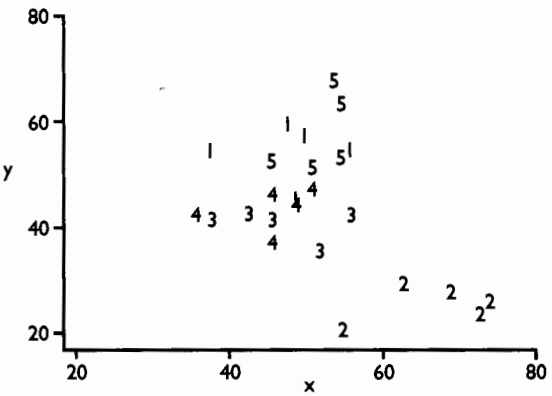
J Martin Bland, Douglas G Altman

This is the first in a series of occasional notes on medical statistics. Some notes will remind readers about the basics: others will keep them up to date with more complex techniques that are finding their way into medical studies. When we have a note on one of these more unusual statistical techniques we will aim to publish it in the same issue as a paper that uses the technique.

In clinical research we are often able to take several measurements on the same patient. The correct analysis of such data is more complex than if each patient were measured once. This is because the variability of measurements made on different subjects is usually much greater than the variability between measurements on the same subject, and we must take both kinds of variability into account. For example, we may want to investigate the relation between two variables and take several pairs of readings from each of a group of subjects. Such data violate the assumption of independence inherent in many analyses, such as *t* tests and regression.

Researchers sometimes put all the data together, as if they were one sample. Most statistics textbooks do not warn the researcher not to do this. It is so ingrained in statisticians that this is a bad idea that it never occurs to them that anyone would do it.

Consider the following example. The data were generated from random numbers, and there is no relation between *X* and *Y* at all. Firstly, values of *X* and *Y* were generated for each "subject," then a further random number was added to make the individual "observation." The data are shown in the table and figure. For each subject separately the correlation between *X* and *Y* is not significant. We have only five subjects and so only five points. Using each subject's mean values, we get the correlation coefficient $r=-0{\cdot}67$, df=3, P=0·22. However, if we put all 25 observations together we get $r=-0{\cdot}47$, df=23, P=0·02. Even though this correlation coefficient is smaller than that between means, because it is based on 25 pairs of observations rather than five it becomes significant. The calculation is performed as if we have 25 subjects, and so the number of degrees of freedom for the significance test is increased incorrectly and a spurious significant difference is produced. The extreme case would occur if we had only two subjects, with repeated pairs of observations on each. We would have two separate clusters of points centred at the subjects' means. We would get a high correlation coefficient, which would appear significant despite there being no relation whatsoever.



Simulated data for five pairs of measurement of two uncorrelated variables (X and Y) for five subjects

Simulated data showing five pairs of measurements of two uncorrelated variables for subjects 1, 2, 3, 4, and 5

	Subject 1		Subject 2		Subject 3		Subject 4		Subject 5	
	48	58	63	28	38	40	51	46	55	62
	56	53	74	24	56	41	46	36	51	50
	49	44	69	26	46	40	36	41	54	66
	38	53	55	19	43	41	49	43	46	51
	50	56	73	22	52	34	46	45	55	52
Subject mean	48·2	52·8	66·8	23·8	47·0	39·2	45·6	42·2	52·2	56·2
Correlation coefficient	$r=-0{\cdot}02$ P=0·97		$r=0{\cdot}32$ P=0·59		$r=-0{\cdot}30$ P=0·63		$r=0{\cdot}37$ P=0·55		$r=0{\cdot}55$ P=0·33	

The extreme case would occur if we had only two subjects, with repeated pairs of observations on each. We would have two separate clusters of points centred at the subjects' means. We would get a high correlation coefficient, which would appear significant despite there being no relation whatsoever.

There are two simple ways to approach these types of data. If we want to know whether subjects with a high value of *X* tend also to have a high value of *Y* we can use the subject means and find the correlation between them. For different numbers of observations for each subject, we can use a weighted analysis, weighting by the number of observations for the subject. If we want to know whether changes in one variable in the same subject are paralleled by changes in the other we can estimate the relation within subjects using multiple regression. In either case we should not mix observations from different subjects indiscriminately, whether using correlation or the closely related regression analysis.

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE
J Martin Bland, *reader in medical statistics*

Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX
Douglas G Altman, *head*

Correspondence to:
Dr Bland.

BMJ 1994;308:896