UNIT 1

Hypothesis testing and estimation

Aims

To understand how the methods of hypothesis testing and estimation complement one another when deciding whether there are important differences in summary statistics between two or more study groups.

Learning objectives

On completion of this unit, participants will be able to:

- understand and interpret *P* values;
- describe the meaning of type I and II errors;
- decide when to use a one-tailed or two-tailed test of significance;
- estimate and interpret 95% confidence intervals.

Background

In health care research, significance tests are usually conducted to assess whether there is evidence for a real difference in the summary statistics of two or more study groups. The summary statistic may be, for example, the mean of the outcome measurement or a frequency rate. When comparing two or more groups, the probability that the difference between the groups has occurred by chance, which is expressed as a *P* value, is used to describe the statistical significance of the findings. However, *P* values convey only part of the information and therefore they should be accompanied by an estimation of the effect size, that is the size of the difference between the study groups that was found.[1] The estimation, which may be a statistic such as the size of the mean difference between two groups, allows readers to assess whether the observed difference is important enough to warrant a change in current health care practice or to warrant further research. Reporting the effect size enables readers to judge whether a statistically significant result is also a clinically important finding.

Hypothesis testing and *P* values

Most medical statistics are based on the concept of hypothesis testing and therefore an associated *P* value is usually reported. In hypothesis testing, a 'null hypothesis' is first specified, that is a hypothesis stating that there is no difference in the summary statistics of the study groups. In essence, the null hypothesis assumes that the groups that are being compared are drawn from the same population. An alternative hypothesis, which states that there is a difference between groups, can also be specified. The *P* value, that is, the probability that the difference between the groups would have occurred assuming the null hypothesis was 'true', is then calculated. A *P* value is obtained by first calculating a test statistic, such as a *t*-statistic or a chi-square value, which is then compared to a known distribution. The known distribution is used to determine the probability that the observed test statistic value (or a more extreme value) would occur, if the null hypothesis were true. In the following units in this book, the calculation and interpretation of the most commonly used test statistics will be explored.

A *P* value of less than 0.05, that is a probability of less than 1 chance in 20, is usually accepted as being statistically significant. If a *P* value is less than 0.05, we accept that it is unlikely that a difference between groups has occurred by chance if the null hypothesis was true. In this situation, we reject the null hypothesis and accept the alternative hypothesis, and therefore conclude that there is a statistically significant difference between the groups. On the other hand, if the *P* value is greater than or equal to 0.05 and therefore the probability with which the test statistic occurs is greater than 1 chance in 20, we accept that the difference between groups has occurred by chance. In this case, we accept the null hypothesis and conclude that there is no difference between the study groups beyond variations that can be attributed to sampling.

In accepting or rejecting a null hypothesis, it is important to remember that the *P* value only provides a probability value and does not provide absolute proof that the null hypothesis is true or false. A *P* value obtained from a test of significance should only be interpreted as a measure of the strength of evidence against the null hypothesis.[2] The smaller the *P* value, the stronger the evidence provided by the data that the null hypothesis can be rejected. Thus, *P* values of 0.01 or lower are conventionally regarded as being highly significant because they indicate that it is highly unlikely that the difference between groups has occurred by chance. Although the cut-off point between statistical significance and non-significance is generally accepted as a *P* value of less than 0.05, it is important to remember that *P* values of 0.07 and 0.04 indicate very similar strengths of evidence even though a *P* value of 0.07 is conventionally regarded as being non-significant and a *P* value of 0.04 as being statistically significant. To convey the strength of evidence rather than using pre-conceived arbitrary categories, actual *P* values should be reported, for example $P = 0.04$ rather than $P < 0.05$ and $P = 0.63$ rather than NS (not significant).

In measuring between-group effects, the absolute magnitude of the difference between the groups and the direction of the effect are not conveyed by the *P* value. Thus, when *P* values alone are reported, the results can only be interpreted as probability values that indicate statistical significance with no regard for the clinical importance of the result. A *P* value that is larger than 0.05 does not necessarily mean that the treatments or the groups that are being compared are similar, because the *P* value depends on both the size of difference between the groups and on the sample size.[3] By only using *P* values, it is not possible to answer questions of how confident we are, given the study results, that a treatment is beneficial or has no effect, or how much better we expect patients to become if they receive a new treatment.[4, 5] For this, we need an estimation of the size of the effect in addition to significance tests.

Estimation

In health care research, rather than enrolling an entire population in a study, which would usually be impractical, a sample of the population is usually selected and then statistics are used to make inferences about the entire population. When using estimation, a summary statistic is calculated that describes the effect size in the sample, together with a margin of precision around the statistic that depends on the size of the sample that was enrolled. Estimation allows us to make judgements on the certainty, or uncertainty, of summary statistics calculated from a sample, and therefore to make inferences about the population from which the sample was drawn.

When comparing two study groups, estimation involves calculating the actual size of the difference between the groups in addition to a *P* value. A limitation in the interpretation of *P* values is that they are heavily influenced by the sample size. Although *P* values provide a measure of the strength of evidence, they convey only a small part of the total information about the effectiveness of a treatment in clinical research or about differences between population samples in epidemiological studies. In a clinical study, the outcome of interest may be, for example, a difference in mean lung function measurements or a per cent reduction in symptoms between groups receiving a new treatment compared to a standard treatment (control). These types of summary statistics indicate how much patients could expect their lung function to increase or their symptoms to improve if they received the new treatment compared to if they received the standard treatment. As such, the summary statistics quantify the actual effect of the new treatment in a way that complements the probability that the difference between groups arose by chance.

TAKE HOME LIST

- A *P* value indicates the strength of evidence against the null hypothesis.
- A *P* value of less than 0.05 indicates that there is a statistically significant difference between the study groups.
- Smaller *P* values provide stronger evidence that the null hypothesis is false.
- The actual *P* value, for example, $P = 0.04$ or $P = 0.56$, should be reported.
- A limitation of *P* values is that they only describe a probability and the statistical significance of a between-group difference.
- *P* values are strongly influenced by the sample size. The larger the sample size the more likely a difference between study groups will be statistically significant.
- Estimation provides an effect size between groups that complements the *P* value.

Confidence intervals

Confidence intervals are important in estimation in that they describe the precision around a summary statistic, such as a difference between study groups.[5] There is error in all estimates of effect because it is unlikely that the measured effect would be the same when a study is repeated in different random samples of the population. When different groups of people are sampled, variations in summary statistics occur simply because there is a large amount of inherent variation in human characteristics.

The confidence interval provides an estimated range of values that is likely to include the population value. The interpretation of a 95% confidence interval is that 95% of the confidence intervals calculated from many different samples would include the true value of the summary statistic that occurs in the population.[6] A simpler and perhaps more intuitive way to interpret a 95% confidence interval is that we can be roughly 95% certain, or confident, that the true value of the summary statistic in the population is within the 95% confidence interval calculated from a single study sample. Thus, confidence intervals provide an estimate of precision, or rather lack of precision, which can be attributed to sampling variation. In Unit 2, we explain how 95% confidence intervals for differences between study groups are calculated and show how these intervals can be used to make statistical inferences about differences between groups, sometimes without the need for computing a *P* value at all.

Confidence intervals are calculated from the standard error (SE), which is an estimate of the precision with which a summary statistic has been measured. The standard error can be used to calculate a 95% confidence interval as follows:

$$95\% \text{ confidence interval} = \text{summary statistic} \pm (1.96 \times \text{SE})$$

In this calculation, the summary statistic may be a value such as a mean value, a percentage or an odds ratio and the SE is the standard error around the summary value. A critical value of 1.96, which is derived from the normal population distribution of the summary statistic, is used to compute 95% intervals when the group or sample size is larger than 50 participants. If the sample size is smaller than 50, a larger critical value than 1.96 that can be derived from a statistical table should be used.[6]

It is important to remember that a 95% confidence interval only applies to populations with the same characteristics as the population from which the data were sampled.[6] However, a 95% confidence interval provides important information over and above the *P* value. This is especially important when the *P* value is greater than 0.05 because a judgement about the clinical importance of the difference that has been measured can be made by assessing the width of the 95% confidence interval. As might be expected, the *P* values and confidence intervals from any study are closely related to one another. In most cases, if the value of the null hypothesis, for example a value equal to 0, falls within the 95% confidence interval then the *P* value will be greater than 0.05.[6]

When critically appraising the literature, it is important to calculate 95% confidence intervals if they are not reported. Although 95% confidence intervals for mean values are calculated from the standard error, which describes the precision around the mean value, the only descriptor of variance that is often reported is the standard deviation (SD), which describes the distribution of the spread or the variation of the actual data points. In describing the error and spread around a mean value, the terms standard error and standard deviation have important distinctions[7] and for this reason they are explained in more detail in Unit 6. To calculate 95% confidence intervals, the standard deviation around a mean value can easily be converted into a standard error as follows:

$$\text{Standard error (SE)} = \text{SD}/\sqrt{n}$$

where n is the sample size of the group from which the mean and the standard deviation were estimated.

As can be seen from the formula, the standard error is inversely related to the square root of the sample size. Thus, the standard error becomes smaller as the sample size increases. As the sample size becomes larger, the width of the 95% confidence interval for the same effect becomes smaller, indicating greater certainty in the precision of the result. On the other hand, as the sample size becomes smaller, the standard error becomes larger and thus the width of the 95% confidence interval becomes wider, indicating less certainty in the precision of the result. The above methods for estimating and calculating 95% confidence intervals apply to all summary statistics. The calculation of standard errors and the 95% confidence interval for proportions, for example incidence and prevalenvce rates, and for odds ratios are discussed in the following units.

Glossary

Term	Definition
Null hypothesis	A hypothesis stating that there is no difference between the study groups.
P value	Probability that a difference between study groups would have occurred if the null hypothesis was true.
95% confidence interval	Range in which we can be approximately 95% certain that the true population value lies.

Type I and II errors

Confidence intervals clearly show the lack of precision around an estimate but, when only a *P* value is calculated, the degree of uncertainty about whether the null hypothesis should be accepted or rejected is easily overlooked. Obviously if a *P* value is very small, say less than 0.01, then the probability that the groups have been sampled from the same population is quite unlikely and we can be confident that there is a real difference. Similarly if the *P* value is large, say over 0.1, then we can be confident that there is no difference between the groups beyond sampling variation.

When testing between-group differences, the *P* value is closely related to the sample size. Thus, the larger the sample size, the smaller the *P* value will be for the same summary statistic, such as a mean difference between groups. The *P* value is smaller when the sample size is large because the summary statistic represents a more accurate estimate of the true value in the population from which the sample is drawn. Thus, the *P* value depends on both the size of the summary statistic and on the sample size. Therefore it is important to consider how the clinical importance of a difference (that is, the actual magnitude of the difference between groups) compares with the statistical significance (that is, the *P* value which is dependent on sample size). The decision about the size of difference between groups that is considered clinically important depends solely on expert knowledge and can only be made by health care practitioners and researchers with experience in the field.

When accepting or rejecting a null hypothesis it is possible that a type I or type II error has occurred. A 'type I error' occurs when the null hypothesis is incorrectly rejected. That is, it is concluded that there is a statistically significant difference between groups when no clinically important difference exists. The probability of a type I error occurring is reported as the *P* value. With a *P* value of 0.05, there is a chance of 5 in 100 or 1 in 20 that the significant results occurred by chance alone. So for every 20 statistical tests that are conducted, one test will be significant by chance alone. Type I errors frequently occur when data is repeatedly analysed, when there are multiple comparisons or multiple outcomes.

A 'type II error' occurs when the null hypothesis is incorrectly accepted. That is, it is concluded that there is no statistically significant difference between groups when a clinically important difference exists. The probability of avoiding a type II error is referred to as the power of the study, that is, the probability of correctly rejecting the null hypothesis. Type II errors typically occur when the sample size is too small for a clinically important difference to reach statistical significance. Because both type I and II errors are a product of the sample size, the risk of a type I error is reduced when the sample size becomes smaller but the risk of a type II error increases.

Although the occurrence of type I and II errors is usually related to the sample size, the consequences of these two types of errors are very different. For example, if a type I error occurs in a clinical trial then a new treatment will be incorrectly judged to be more effective than the control group treatment. If the new treatment is more expensive or has more severe side effects, recommendation of the new treatment will not confer benefit on average but will have an adverse impact on people to whom it is recommended. On the other hand, if a type II error occurs, the new treatment will be incorrectly judged to have no advantage over the control treatment even though many people who receive the new treatment will experience beneficial effects. Type I and II errors not only have clinical implications for interpretation of summary statistics but also have ethical implications. If the sample size is too small, the study may be unethical because too few participants are enrolled than are needed to test the study hypothesis, and therefore research resources will also be wasted. If the sample size is too large, the study may be unethical because more participants are enrolled than are needed to test the study hypothesis and research resources will also be wasted. For these reasons, ethics committees often request that a statistician is consulted when a study is being designed to ensure that the probability of type I and II errors is minimised.

Glossary

Term	Definition
Type I error	When the null hypothesis is incorrectly rejected. That is, a difference between groups is statistically significant although a clinically important difference does not exist.
Type II error	When the null hypothesis is incorrectly accepted. That is, a difference between groups is not statistically significant although a clinically important difference exists.

TAKE HOME LIST

- As the sample size increases, the width of the 95% confidence interval becomes smaller, indicating greater certainty in the precision of the result.
- Summary statistics and their 95% confidence intervals should be reported, together with *P* values, to indicate the absolute size of the difference between the groups and the direction of effect.
- Type I and II error rates are inversely related because both are influenced by sample size – when the risk of a type I error is reduced, the risk of a type II error is increased.

One-tailed and two-tailed tests of significance

The calculation of a *P* value is influenced by the expected direction of difference between study groups, which is generally specified as the alternative hypothesis. When the difference between two study groups is expected to occur in one direction only, for example when a group of people receiving one treatment could only show greater improvements than a group receiving another treatment, a one-tailed (or one-sided) test of significance is used. For one-tailed *t*-tests, the probability of the test statistic value or one more extreme occurring in only one direction, such as occurring in only the upper tail of the distribution, is calculated.

When the difference between two study groups is expected to occur in either direction, for example when a group of people receiving one treatment could show a larger or smaller improvement than a group receiving another treatment, a two-tailed (or two-sided) test of significance is used. For two-tailed tests, the probability of the test statistic occurring in either the upper or lower tail of the distribution is calculated. Since one-tailed tests involve calculating the probability using only one tail of the distribution of the test statistic, the *P* value is reduced by half so that it is more significant than when both tails are used in a two-sided test.

In most health care research studies, the use of one-tailed tests is rarely justified because we should expect that a result could be in either direction. It is most unusual for researchers to be certain about the direction of effect before the study is conducted and, if they were, the study would probably not need to be conducted at all.[7] For this reason, one-tailed statistics are rarely used. A search of the abstracts published in the *British Medical Journal* between 1994 and 2006 found only one study in thirteen years in which the results were described using a one-tailed significance test. If a one-tailed *P* value is reported, the *P* value can easily be converted into a two-tailed (or two-sided) value by doubling its numerical value.

In the vast majority of studies, two-tailed tests of significance are used unless there is a very good reason for not doing so.[9] In health care research, it is almost always important to allow for the possibility that extreme results could occur by chance and could occur equally often in either direction, which in clinical trials would mean towards a beneficial or towards an adverse effect. Two-tailed tests provide a more conservative result than one-tailed tests in that the *P* value is higher, that is, less significant. In this way, two-tailed tests reduce the chance that a between-group difference is declared statistically significant in error, and thus that a new treatment is incorrectly accepted as being more effective than an existing treatment. A conservative approach is essential because no health care practice should be modified on the basis of results that have arisen entirely by chance.

Reading and questions

Reprint

Berry G. Statistical significance and confidence intervals. Med J Aust 1986;144:618–619. (See p. 7.)

After reading Unit 1 and the reprint by Berry (1986) answer the following questions:

1. Can 95% confidence intervals be used to infer *P* values and vice versa?
2. When might a significance test fail to detect a real effect?
3. When is the null hypothesis value outside the 95% confidence interval?
4. What type of error occurs when a difference between groups is not statistically significant but is large enough to be thought clinically important?
5. Who decides what size of difference between groups is clinically important?

Worked example

Set article

Logroscino G, Kang, JH, Grodstein F. Prospective study of type 2 diabetes and cognitive decline in women aged 70–81 years. BMJ (Published 23 February 2004). (See p. 10.)

In the set article by Logroscino *et al.* (2004) the authors refer to Table 2 and state that "On every cognitive test, means baseline scores were lower for women with diabetes". Review this table and decide how this conclusion was reached.

- What statistical test was used?
- What do the authors mean by "lower"?
- Have the authors used hypothesis testing or estimation to reach this conclusion?
- What is the size of the difference between groups and is it clinically important?
- Was there a type I or type II error?
- Would you reach the same conclusion?

Exercise

The standard deviation around each estimate in Table 2 from Logroscino *et al.* (2004) is easily converted first into an SE and then to a 95% confidence interval. In Table 1.1, calculate the SE and 95% confidence intervals for the participants with diabetes.

After completing the new estimations in Table 1.1 decide:

- What factors influence the 95% confidence intervals and in what way?
- Why are the confidence intervals so narrow?

Table 1.1 Mean and 95% CI cognitive scores at baseline in 1394 women with type 2 diabetes

	N	Mean	SD	SE	Lower 95% CI	Upper 95% CI
TICS (8–41 points)	1394	33.2	2.9	0.08	33.1	33.4
TICS 10 word list	1394	2.0	1.9			
East Boston memory test – immediate recall	1394	9.3	1.8			
East Boston memory test – delayed recall	1394	8.9	2.1			

Table 1.2 Mean and 95% CI cognitive scores at baseline in 50 women with type 2 diabetes

	N	Mean	SD	SE	Lower 95% CI	Upper 95% CI
TICS (8–41 points)	50	33.2	2.9	0.41	32.4	34.0
TICS 10 word list	50	2.0	1.9			
East Boston memory test – immediate recall	50	9.3	1.8			
East Boston memory test – delayed recall	50	8.9	2.1			

Next calculate the SE and 95% confidence intervals if the sample comprised only 50 participants, rather than the enrolled number of 1394.
After completing the new estimations in Table 1.2 decide:

- What happens to the 95% confidence intervals when the sample size is smaller?
- Why does this happen?

Quick quiz

Tick the correct answer for each of the following questions.

1 A 95% confidence interval is:
(a) the range in which a mean value falls approximately 95% of the time;
(b) the range in which 95% of the study observations can be expected to lie;
(c) the range in which we are 95% certain that the true population value lies;
(d) the range calculated as the mean ± 1.96 standard deviations and which excludes 5% of the sample.

2 A type II error occurs when:
(a) a statistician makes an error in calculating a *P* value;
(b) an important difference between groups has a *P* value that is larger than 0.05;
(c) a clinically important effect is unlikely to have occurred by chance;
(d) a new treatment proves more effective than was thought when the sample size was calculated.

3 Two-tailed tests of significance are used because:
(a) that is what statisticians recommend as standard practice;
(b) statisticians are often unsure of what the study results will show;
(c) all studies have some degree of sampling variation that affects the results;
(d) a new treatment could turn out to be better or worse than the control treatment.

4 An estimation of the difference between study groups provides important information that is additional to a *P* value because:
(a) it conveys the size of the difference of effect between the groups;
(b) it provides a more reliable summary statistic;
(c) it conveys how well the new treatment works;
(d) it is an essential component of evidence-based practice.

References

1. Altman DG. Estimation or hypothesis testing? In: Practical statistics for medical research. London: Chapman & Hall, 1996; pp 174–175.
2. Bland JM. Principles of significance tests. In: An introduction to medical statistics. Oxford: Oxford University Press, 1996; p 136.
3. Freiman JA, Chalmers TC, Smith H, Keubler RR. The importance of the beta, the type II error and sample size in the design and interpretation of the randomised controlled trial. Survey of 71 "negative" trials. N Engl J Med 1978;299:690–694.
4. Shakespeare TP, Gebski VJ, Veness MJ, Simes J. Improving interpretation of clinical studies by use of confidence levels, clinical significance curves, and risk-benefit contours. Lancet 2001;357:1349–1353.
5. Gardner MJ, Altman DG. Confidence intervals rather than *P* values: estimation rather than hypothesis testing. BMJ (Clin Res Ed) 1986;292:746–750.
6. Bland JM, Peacock J. Interpreting statistics with confidence. TOG 2002;4:176–180.
7. Altman DG, Bland JM. Standard deviations and standard errors. BMJ 2005;331:903.
8. Altman DG. Two-sided or one-sided *P* values? In: Practical statistics for medical research. London: Chapman & Hall, 1996; pp 170–171.
9. Bland JM, Altman DG. One and two sided tests of significance. BMJ 1994;309:248.

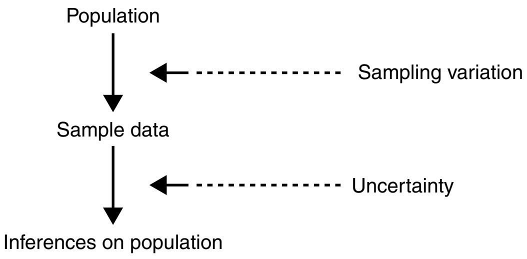
Statistical significance and confidence intervals

Geoffrey Berry

Many papers in the Journal use statistical methods and one of the aims of the review process is to try to ensure that appropriate methods have been used. Often papers report results of comparative studies that are designed to answer questions such as whether one treatment is superior to another for a particular disease, or whether there is an association between some form of behaviour (for example, taking regular exercise or smoking) and the occurrence of some disease. Comparative studies are almost invariably carried out on a *sample* of individuals who are chosen from the *population* of individuals to whom it is intended to generalize the results. Data are collected on the sample in order to make inferences on the population. Valid inferences can only be drawn if the sample is chosen in such a way that it is representative of the population. Otherwise a bias could occur; epidemiological methods are designed to eliminate such biases.

Since the aim of a statistical analysis is to make inferences, it is paramount to express whatever inferences that can be drawn in the most informative way. There are several methods of statistical inference, but the two that are most commonly used are significance testing and confidence interval estimation. The former is well known and is featured by quoting *P* values. Many authors appear to be under the impression that a profusion of *P* values is necessary; regrettably this impression has been bolstered in the past by editors of biological journals. Significance testing has its place but, as mentioned by Healy in 1978,[1] "it is widely agreed among statisticians (if less so among the more naive users of statistics) that significance testing is not the be-all and end-all of the subject". In this leading article I would like to discuss the characteristics of both methods of inference, show that a confidence interval contains the result of a significance test, but not vice versa, and suggest that confidence intervals are the answers to the more interesting questions that data can be used to answer.

Any particular study is based on a particular sample; however, it is useful to imagine that the study is repeated with a different sample being selected each time. These hypothetical studies will give different results because they contain different individuals, and individuals vary in any characteristic because of biological variability. The differences are termed *sampling variability.* It follows then that the results that are obtained from a particular sample can only be taken as an approximation to the actual situation in the whole population. Statistical methods are concerned with assessing the degree of approximation and what may be reasonably inferred, given that a different sample would have produced a different result.



The methods are based on the assumption that it is a matter of chance which particular subjects are in the sample that is being studied, and the sampling variability is thus random variation which is determined by the laws of probability. Therefore, the inferences are expressed in terms of probability. The situation is illustrated below.

Taking a sample from the population involves sampling variation. As a consequence of this, inferences from the sample data back to the population involve uncertainty.

A statistical analysis may be thought of as asking questions of the data. In an investigation that compares two groups for the mean value of, for example, blood pressure or the prevalence of some disease, three questions may be posed: Is there a difference between the groups?; How large is the difference?; and How accurately is the size of the difference known?

As expressed, the first question expects the answer "yes" or "no"; although the answer cannot be given in precisely these terms, it is often reduced to two possibilities. The appropriate methodology is the *significance test.* The second question expects a numerical value to be the answer. This is an estimate and, as it is a single value, is referred to as a *point estimate.* In effect, the third question asks how reliable this point estimate is; the answer is a range of values which is referred to as an *interval estimate* or a *confidence interval.*

These questions represent two approaches to inference: hypothesis testing and estimation. Although at first sight they appear to be quite different, in concept they have much in common. Both make inferential statements about the value of a parameter. (A parameter is an unknown quantity which partly or wholly characterizes a population, for example, a mean or a measure of association.)

Associate Professor of Biostatistics, School of
Public Health and Tropical Medicine
The University of Sydney

Originally published in *MJA* 1986; **144**: 618–19. Reproduced with permission.

The significance test is an appropriate technique when there is an a priori hypothesis to test. For the purpose of the statistical test this hypothesis is expressed in *null* form — such as when no difference exists between groups — and the test evaluates whether the data are consistent with the null hypothesis. If the data differ markedly from those which would be expected under the null hypothesis, to the extent that the probability of such an extreme result is low, then it is said that the result is statistically significant. Probability is measured on a continuum between 0 and 1, but in significance testing a probability is considered low if it is less than conventional values such as 0.05 (5%) or 0.01 (1%). A significant result is equated with the rejection of the null hypothesis or the claim of a real effect. By definition, when the null hypothesis is true, significant results will occur by chance with the same relative frequency as the significance probability. That is, real effects will be claimed when the null hypothesis is true; however, the probability of this error (type 1) is determined in the data analysis.

One disadvantage of a significance test is that it may fail to detect a real effect; that is, although the null hypothesis is false, the evidence is not strong enough to reject it. The probability of this error (type II) can be controlled at the design stage only, by appropriate selection of the sample size, and may be quite large. Thus, the trap of equating non-significance with no effect must be avoided; failure to reject the null hypothesis is not the same as accepting it.

In the approach of confidence interval estimation no particular hypothesis is considered; rather, the emphasis is on estimating those values of the parameter with which the data are consistent. These values form a range — the confidence interval. The range is calculated so that there is a high probability — conventionally 95% or 99% — that it contains the true value of the parameter.

A significance test is essentially a test of whether the data are consistent with a specified parameter value, and the confidence interval contains those parameter values with which the data are consistent. Therefore, a 5% significance test and a 95% confidence interval contain some information in common: significance implies that the null hypothesis value is outside the confidence interval; non-significance implies that the null hypothesis value is within the confidence interval. However, the confidence interval contains more information because it is equivalent to performing a significance test for all values of the parameter, not just a single value. A confidence interval enables a reader to see how large the effect may be, not simply whether it is different from zero.

The limitations of the interpretations that are provided by a significance test may now be considered.

The difference is significant. This means that there is a difference or, in other words, the size of the difference is not zero. We know no more than this. The difference may be large and of great importance or it may be small and of no practical importance. It is unsatisfactory that the test provides no way of distinguishing between these quite different possibilities.

The difference is not significant. This means that there is insufficient evidence to enable us to conclude that there is a difference. So the difference may well be zero. But this is not the same as saying that it is zero. The true difference may be quite large. Again, it is unsatisfactory that this possibility is not addressed.

The conclusions that may be drawn from a significance test are considered to be incomplete because it is rarely that one is interested solely in whether a null hypothesis is or is not true; indeed in many cases it may be recognized at the outset that the null hypothesis is unlikely to be true. Rather, the question is how large is the difference and is it possibly large enough to be important? The emphasis is on measuring rather than on testing. The addition of the concept of an important difference to that of a null hypothesis means that there are four possible interpretations to an analysis: (a) the difference is significant and large enough to be of practical importance; (b) the difference is significant but too small to be of practical importance; (c) the difference is not significant but may be large enough to be important; and (d) the difference is not significant and also not large enough to be of practical importance.

The size of difference that is considered to be large enough to be important is a matter for debate, and genuine differences of opinion may arise. It is a medical, not a statistical, question, although a medical statistician who is experienced in the subject area could contribute to setting a value. The fact that agreement on a unique value may be impossible in no way detracts from the argument. In fact, expressing the results as a confidence interval enables interpretations to be made for any particular value that is considered appropriate.

These possibilities are illustrated in the Figure where the confidence intervals are shown. The significant and

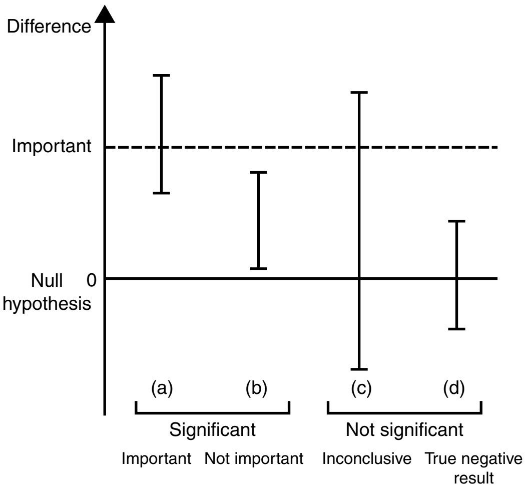


Figure Confidence intervals showing four possible conclusions in terms of statistical significance and practical importance.

Originally published in *MJA* 1986; **144**: 618–19. Reproduced with permission.

non-significant cases are distinguished by the confidence intervals that exclude or include zero respectively. The main point is that in each case the confidence interval gives the range of possible values for the true difference. Of particular concern is (c). Here there may be no true difference or there may be a large, important difference. In other words the study is completely inconclusive. Such a possibility is missed by the simple expression "not significant" with its lure of equating this falsely with "no effect". This situation will arise with a study that is carried out on too small a sample and this is why good study design demands attention to sample size to try to prevent the occurrence of an inconclusive result. Altman found that it was common for undue emphasis to be placed on "negative" findings from small studies,[2] while Freimen et al. noted that "negative" trials were often too small to constitute a fair test of therapies.[3] Similarly, a significance test will contrast (b) as significant and (d) as not significant but fails to recognize that they give essentially the same conclusion — that any difference is too small to be important.

As an example, consider some results which were obtained by Garraway et al. from a clinical trial for the management of acute stroke in the elderly.[4] Of 155 patients who were managed in a stroke unit, 78 were assessed as independent when they were discharged from the unit compared with 49 of 152 who were managed in a medical unit. The simplest analysis shows that the difference between the success rates of the two units is significant at the 1% level. Therefore, a genuine effect has been established. To appreciate the importance of this effect the advantage of the stroke unit may be measured by the difference between the two units in the percentage of subjects who were discharged as independent: 50.3% − 32.2% = 18.1%. This is the point estimate. The accuracy of this estimate is given by its standard error (5.5) and the 95% confidence limits (7.3% and 28.9%). Thus, the gain could be as large as 29% or as small as 7%.

Recently, Gardner and Altman have argued against the excessive use of hypothesis testing and urged a greater use of confidence intervals.[5] In an appendix to their paper they give methods to calculate confidence intervals for the commonly occurring two-sample comparisons.

In presenting the main results of a study it is good practice to provide confidence intervals rather than to restrict the analysis to significance tests. Only by so doing can authors give readers sufficient information for a proper conclusion to be drawn; otherwise readers have to rely upon the authors' own interpretation.[2] Therefore, intending authors are urged to express their main conclusions in confidence interval form (possibly with the addition of a significance test, although strictly that would provide no extra information). One of the aims of the Journal's statistical review process will be to ensure that where possible this is done.

References

1 Healy MJR. Is statistics a science? *J R Statist Soc A* 1978; 141: 385–393.

2 Altman DG. Statistics in medical journals. *Stat Med* 1982; 1: 59–71.

3 Freiman JA, Chalmers TC, Smith H Jr, Kuebler RR. The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial. *N Engl J Med* 1978; 299: 690–694.

4 Garraway WM, Akhtar AJ, Prescott RJ, Hockey L. Management of acute stroke in the elderly: preliminary results of a controlled trial. *Br Med J* 1980; 280: 1040–1043.

5 Gardner MJ, Altman DG. Confidence intervals rather than P values: estimation rather than hypothesis testing. *Br Med J* 1986; 292: 746–750.

Originally published in *MJA* 1986; **144**: 618–19. Reproduced with permission.

Prospective study of type 2 diabetes and cognitive decline in women aged 70–81 years

Giancarlo Logroscino, Jae Hee Kang, Francine Grodstein

Abstract

Objective To examine the association of type 2 diabetes with baseline cognitive function and cognitive decline over two years of follow up, focusing on women living in the community and on the effects of treatments for diabetes.

Design Nurses' health study in the United States. Two cognitive interviews were carried out by telephone during 1995–2003.

Participants 18 999 women aged 70–81 years who had been registered nurses completed the baseline interview; to date, 16 596 participants have completed follow up interviews after two years.

Main outcome measures Cognitive assessments included telephone interview of cognitive status, immediate and delayed recalls of the East Boston memory test, test of verbal fluency, delayed recall of 10 word list, and digit span backwards. Global scores were calculated by averaging the results of all tests with *z* scores.

Results After multivariate adjustment, women with type 2 diabetes performed worse on all cognitive tests than women without diabetes at baseline. For example, women with diabetes were at 25–35% increased odds of poor baseline score (defined as bottom 10% of the distribution) compared with women without diabetes on the telephone interview of cognitive status and the global composite score (odds ratios 1.34, 95% confidence interval 1.14 to 1.57, and 1.26, 1.06 to 1.51, respectively). Odds of poor cognition were particularly high for women who had had diabetes for a long time (1.52, 1.15 to 1.99, and 1.49, 1.11 to 2.00, respectively, for ≥15 years' duration). In contrast, women with diabetes who were on oral hypoglycaemic agents performed similarly to women without diabetes (1.06 and 0.99), while women not using any medication had the greatest odds of poor performance (1.71, 1.28 to 2.281, and 1.45, 1.04 to 2.02) compared with women without diabetes. There was also a modest increase in odds of poor cognition among women using insulin treatment. All findings were similar when cognitive decline was examined over time.

Conclusions Women with type 2 diabetes had increased odds of poor cognitive function and substantial cognitive decline. Use of oral hypoglycaemic therapy, however, may ameliorate risk.

Introduction

Several population based studies have shown that type 2 diabetes increases the risk of dementia.[1–5] Cognitive decline is an intermediate stage between normal ageing and dementia.[6] As dementia may be most effectively delayed in its initial stages, identifying diabetes as a modifiable risk factor for early cognitive decline could be of major public health importance. Estimates in the United States indicate that delaying onset of dementia by one year could lead to 800 000 fewer cases after 50 years.[7]

Though many investigations have examined diabetes in relation to early cognitive decline,[5,8–19] only one large prospective study has focused on women.[8] Type 2 diabetes disproportionately affects older women and is a stronger risk factor for cardiovascular disease in women than in men.[20] As cardiovascular disease is an independent risk factor for cognitive decline, we need to determine the impact of diabetes on cognition in women.[20] Moreover, few studies have evaluated the influence of different treatments for diabetes on the association between type 2 diabetes and cognition.

We assessed the associations between type 2 diabetes, different treatments for diabetes, and cognitive function in more than 16 000 women.

doi 10.1136/bmj.37977.495729.EE

Department of Epidemiology, Harvard School of Public Health, Boston, MA 02115, USA

Giancarlo Logroscino *associate professor of neuroepidemiology*

Channing Lab, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston

Francine Grodstein *associate professor of medicine*
Jae Hee Kang *instructor of medicine*

Correspondence to: G Logroscino
(email: glogrosc@hsph.harvard.edu)

(*Accepted 27 November 2003*)

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

Methods

The nurses' health study is a prospective cohort of 121 700 US female registered nurses, who were aged 30–55 years in 1976, when the study began. Participants' health information has been updated with biennial mailed questionnaires. Over 90% of the original cohort have been followed up to date.

From 1995–2001, participants aged 70 years and older who had not had a stroke were given baseline cognitive assessments by telephone. Overall, 93% completed the interview. Interviewers were blinded to participants' health status (including diabetes). For the baseline analyses of cognitive function, we included 18 999 women with complete information on education and without type 1 diabetes, gestational diabetes, or unconfirmed diabetes (see below).

The follow up cognitive assessment began about two years after the baseline interview. After the exclusion of the 3% who died, calls have been attempted for 98% to date. Of these, 92% (n = 16 596) completed the interview, 5% (n = 967) refused, 3% (n = 526) were unreachable. For analyses of cognitive decline, we included 16 596 participants who completed both assessments and excluded women in whom diabetes had been newly diagnosed between the baseline and second interviews.

Assessment of cognitive function

Our cognitive assessment has been previously described.[21] Briefly, we initially administered only the telephone interview for cognitive status (TICS) (n = 18 999)[22] but gradually added more tests: immediate (n = 18 295) and delayed recalls of the East Boston memory test (n = 18 268), test of verbal fluency (naming animals, n = 18 285), digit span backwards (n = 16 591), and delayed recall of a 10 word list (n = 16 582). To summarise performance, we calculated a global score averaging results of the six tests using z scores (16 563 women completed all six tests).

We have established high validity ($r = 0.81$ comparing the global score from our telephone interview to an in-person exam) and high reliability ($r = 0.70$ for two administrations of the TICS, 31 days apart)[21] for these telephone interviews in highly educated women.

Ascertainment of type 2 diabetes

We identified women who reported that diabetes had been diagnosed by a physician before the baseline cognitive interview. We then confirmed reports based on responses to a supplementary questionnaire including complications, diagnostic tests, and treatment; confirmations conformed to guidelines of the National Diabetes Data Group[23] until 1997, and revised criteria of the American Diabetes Association from 1998.[24] Validation studies found 98% concordance of our nurse participants' reports of type 2 diabetes with medical records.[25] We estimated duration of diabetes by subtracting date of diagnosis from date of baseline cognitive interview. We obtained information on recent drug treatment for diabetes from the biennial questionnaire before the baseline interview.

Statistical analyses

Baseline analyses—We examined the relation between type 2 diabetes and cognitive performance by comparing "poor scorers" to remaining women. "Poor scorers" on the TICS were those who scored <31 points (a pre-established cut off point[21]); on other tests, we defined poor scorers as those below the lowest 10th centile (≤7 for immediate recall and ≤6 for delayed recall on Boston memory test, ≤11 for verbal fluency test, ≤0 for delayed recall of the TICS 10 words list, and ≤3 for digit span backwards). Multivariate adjusted odds ratios of a poor score and 95% confidence intervals were calculated with logistic regression models. We also analysed scores continuously using multiple linear regression to obtain adjusted differences in mean score between women with and without diabetes.

Analyses of cognitive decline—We used logistic regression to calculate odds ratios of "substantial decline," defined as the worst 10% of the distribution of change from the baseline to the second interview (with cut off points for decline of ≥4 on the TICS, ≥6 on the category fluency test, and ≥3 on the other tests). We also used linear regression to estimate adjusted mean differences in decline by diabetes status.

Potential confounding factors—Data on potential confounders were identified from information provided as of the questionnaire immediately before the baseline cognitive assessment. All potential confounding variables were selected a priori based on risk factors for cognitive function in the existing literature (see tables 3 and 4). In analyses of cognitive decline, we adjusted for baseline performance.[26]

Results

At baseline interview 7.3% (n = 1394) of the women had type 2 diabetes, with a mean duration of 12 years since diagnosis. Of the 1248 women with diabetes who completed the most recent questionnaire, 901 reported recent medication for management of diabetes (294 (33%) insulin, 607 (67%) oral hypoglycaemic agents). As expected, women with diabetes had higher prevalence of several comorbid conditions (hypertension, high cholesterol, heart disease, obesity, depression) than women without diabetes (table 1), and used hormone therapy less and drank less alcohol. On every cognitive test, mean baseline scores were lower for women with diabetes (table 2).

We focused analyses on two measures of general cognitive function: the TICS and the global score (table 3). After we adjusted for potential confounding factors, women with diabetes were at 25–35% increased odds of poor baseline score compared with women without diabetes (odds ratio 1.34, 95% confidence interval 1.14 to 1.57, for TICS and 1.26, 1.06 to 1.51, for global score). Findings were consistent when we examined mean differences in scores; the mean score for women with diabetes was lower by −0.42 points, −0.58 to −0.27 points, on the TICS and by −0.09 units, −0.12 to −0.05 units, on the global score compared with women without diabetes. Associations became stronger with longer duration of diabetes. For those with diabetes for ≥ 15 years the odds

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

Table 1 Characteristics of women aged 70–81 years, according to type 2 diabetes. Figures are percentage of respondents unless stated otherwise*

	Without diabetes	With diabetes
No of participants	17605	1394
Mean age (years)	74.2	74.2
Masters or doctorate degree	5.8	5.0
History of hypertension	53.2	78.1
History of hypercholesterolaemia	64.0	75.5
History of heart disease	5.2	15.2
Obesity (body mass index ≥30 kg/m^2)	15.3	38.8
Self perceived low energy (<55 in SF-36 energy-fatigue index)	13.4	24.7
Self perceived depression (<52 in SF-36 mental health index)	2.6	5.0
Current antidepressant use	5.3	7.9
Current regular aspirin use	37.8	42.0
Current regular use of other non-steroidal inflammatory drugs	17.1	18.2
Current use of vitamin E	41.9	37.2
Current use of postmenopausal hormone	32.6	22.0
Mean (SD) age at menopause in years	48.3 (6.4)	47.7 (6.8)
Median physical activity in metabolic equivalents/week (25th–75th centile)	9.8 (3.2–21.9)	4.3 (1.0–14.0)
Current smoking	8.7	6.0
Median alcohol intake in g/day (25th–75th centile)	1.0 (0.0–6.4)	0.0 (0.0–0.9)

*Characteristics from questionnaire immediately before baseline cognitive test. Type 2 diabetes defined as diagnosis at any time before baseline cognitive test.

Table 2 Mean cognitive test scores at baseline in women aged 70–81, according to type 2 diabetes. Figures are means (SD)

Test (range of scores)	Without diabetes	With diabetes
TICS (8–41 points)	33.8 (2.8)	33.2 (2.9)
TICS 10 word list—delayed (0–10 points)	2.3 (2.0)	2.0 (1.9)
Global score (–4–2 standard units)	0.005 (0.6)	–0.1 (0.6)
East Boston memory test—immediate recall (0–12 points)	9.4 (1.7)	9.3 (1.8)
East Boston memory test—delayed (0–12 points)	9.0 (2.0)	8.9 (2.1)
Verbal fluency test (0–38 points)	16.9 (4.7)	16.3 (4.6)
Digit span backwards (0–12)	6.7 (2.4)	6.4 (2.4)

TICS = telephone interview of cognitive status.

of poor cognitive performance was 50% higher than for women without diabetes (1.52, 1.15 to 1.99, and 1.49, 1.11 to 2.00, respectively).

Odds of poor performance also seemed to differ across treatment groups (table 3). Compared with women without diabetes, we found high odds of poor performance for women with diabetes who did not report pharmaceutical treatment (1.71, 1.28 to 2.28, and 1.45, 1.04 to 2.02, respectively). Those taking insulin also had modestly increased odds of poor cognition (1.20, 0.85 to 1.70, and 1.38, 0.97 to 1.95, respectively). In the more powerful analyses of mean differences, the worst performance was among women using

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

insulin (mean differences −0.40, 0.72 to −0.09, and −0.11, −0.18 to −0.03, respectively). In contrast, those taking oral medications had similar odds of poor cognitive performance as those without diabetes (odds ratios 1.06, 0.81 to 1.37, and 0.99, 0.74 to 1.33, respectively) and had the smallest mean difference in score (mean differences −0.35, −0.58 to −0.13, and −0.06, −0.11 to −0.01, respectively).

As cognitive impairment may be a cause rather than a consequence of not taking medications, we also examined use of medication at time of diagnosis (average of 12 years before cognitive assessment). However, results were similar: the odds ratios for poor score were 1.61, 1.19 to 2.16, and 1.43, 1.02 to 2.00, respectively, for women with diabetes who were not taking medication at diagnosis compared with women without diabetes.

In addition, as duration of diabetes, medication use, and level of control are correlated we conducted additional analyses to try to assess their independent effects. The results for duration of diabetes were largely similar after we adjusted for medication use, and results for medication use were largely unchanged after we included a term for duration in the model or stratified by duration of diabetes. For example, among women with diabetes, those not taking medication had a higher risk of poor cognitive performance on the TICS compared with those taking oral medication both in the group with duration of diabetes <10 years (1.73, 1.01 to 2.98) and

Table 3 Diabetes, duration of diabetes, and use of medication for diabetes in women aged 70–81 in relation to baseline cognitive function

		Odds ratio of poor cognitive performance (95% CI)		Mean difference in cognitive performance (95% CI)	
	% of women	TICS (n = 18 999)	Global score* (n = 16563)	TICS (n = 18999)	Global score* (n = 16563)
Diagnosis					
No diabetes	92.7	1.00	1.00	0	0
Diabetes:					
Adjusted for age and education	7.3	1.44 (1.24 to 1.69)	1.37 (1.16 to 1.63)	−0.55 (−0.70 to −0.41)	−0.11 (−0.15 to −0.08)
Multivariate adjusted†	7.3	1.34 (1.14 to 1.57)	1.26 (1.06 to 1.51)	−0.42 (−0.58 to −0.27)	−0.09 (−0.12 to −0.05)
Duration of diabetes (years)					
No diabetes	92.7	1.00	1.00	0	0
Adjusted for age and education:					
≤4	1.5	1.35 (0.97 to 1.88)	1.53 (1.08 to 2.18)	−0.37 (−0.69 to −0.06)	−0.10 (−0.17 to −0.03)
5–9	2.1	1.16 (0.86 to 1.58)	0.91 (0.64 to 1.31)	−0.51 (−0.79 to −0.24)	−0.09 (−0.15 to −0.03)
10–14	1.6	1.59 (1.17 to 2.16)	1.44 (1.03 to 2.02)	−0.68 (−1.00 to −0.37)	−0.12 (−0.19 to −0.05)
≥15	2.1	1.69 (1.30 to 2.21)	1.68 (1.27 to 2.24)	−0.63 (−0.91 to −0.36)	−0.14 (−0.21 to −0.08)
P for trend		<0.0001	<0.0001	<0.0001	<0.0001
Multivariate adjusted†:					
≤4	1.5	1.27 (0.91 to 1.79)	1.48 (1.03 to 2.11)	−0.27 (−0.59 to 0.04)	−0.08 (−0.16 to −0.01)
5–9	2.1	1.10 (0.81 to 1.50)	0.86 (0.60 to 1.25)	−0.41 (−0.69 to −0.14)	−0.07 (−0.13 to −0.01)
10–14	1.6	1.48 (1.08 to 2.02)	1.31 (0.93 to 1.85)	−0.53 (−0.84 to −0.22)	−0.09 (−0.16 to −0.02)
≥15	2.1	1.52 (1.15 to 1.99)	1.49 (1.11 to 2.00)	−0.46 (−0.73 to −0.18)	−0.11 (−0.17 to −0.04)
P for trend		0.0002	0.007	<0.0001	<0.0001

Continued

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

Table 3 Continued

		Odds ratio of poor cognitive performance (95% CI)		Mean difference in cognitive performance (95% CI)	
	% of women	TICS (n = 18 999)	Global score* (n = 16 563)	TICS (n = 18 999)	Global score* (n = 16 563)
Medication‡					
No diabetes	92.7	1.00	1.00	0	0
Adjusted for age and education:					
Insulin	1.5	1.27 (0.91 to 1.78)	1.48 (1.06 to 2.08)	−0.55 (−0.86 to −0.23)	−0.14 (−0.20 to −0.07)
Oral medication	3.2	1.05 (0.82 to 1.36)	0.99 (0.74 to 1.31)	−0.40 (−0.62 to −0.18)	−0.06 (−0.11 to −0.01)
No reported treatment	1.8	1.70 (1.28 to 2.26)	1.43 (1.03 to 1.98)	−0.42 (−0.71 to −0.13)	−0.09 (−0.16 to −0.02)
Multivariate adjusted†:					
Insulin	1.5	1.20 (0.85 to 1.70)	1.38 (0.97 to 1.95)	−0.40 (−0.72 to −0.09)	−0.11 (−0.18 to −0.03)
Oral medication	3.2	1.06 (0.81 to 1.37)	0.99 (0.74 to 1.33)	−0.35 (−0.58 to −0.13)	−0.06 (−0.11 to −0.01)
No reported treatment	1.8	1.71 (1.28 to 2.28)	1.45 (1.04 to 2.02)	−0.38 (−0.67 to −0.09)	−0.08 (−0.15 to −0.01)

TICS = telephone interview of cognitive status.

*Global score combines TICS, test of verbal fluency, delayed recall of TICS 10 word list, digit backwards test, immediate and delayed recalls of East Boston memory test.

†Adjusted for age at interview (years), highest attained education (registered nurse diploma, Bachelor's degree, Master's or Doctoral degree), history of high cholesterol (yes, no), history of high blood pressure (yes, no), use of vitamin E supplement (currently yes, no), age at menopause (<50, 50–52, ≥53 years), body mass index (<22, 22–24.9, 25–29.9, ≥30 kg/m^2), cigarette smoking (current, past, never), antidepressant use (yes, no), alcohol intake (0, 1–4, 5–14, ≥15 g/day), use of aspirin (current use 1–5 times/week, use ≥6 times/week, no), use of other NSAID (current use, no), postmenopausal hormone use (currently yes, no), mental health index (0–52, 52–100), and energy-fatigue index (0–54, 55–100) from SF-36.

‡Data on medication use from questionnaire immediately before baseline cognitive assessment. Percentages do not total 100% as 0.8% who did not respond to medication question are not presented.

≥10 years (1.90, 1.04 to 3.48). Furthermore, although we did not have detailed information on level of control (for example, data on haemoglobin A_{1c} concentration), all results were generally unchanged when we excluded data from women with metabolic complications (for instance, those with severely uncontrolled disease).

Finally, we restricted analyses to participants who did not report any difficulty with hearing (n = 12 099) to reduce confounding by hearing status. The results were similar when we compared women with and without diabetes (1.45, 1.18 to 1.78, and 1.37, 1.10 to 1.71, respectively).

Prospective analyses of decline

Although cognitive decline was measured over just a two year period, we observed a significantly increased odds of substantial decline on the TICS (1.26, 1.03 to 1.54) for women compared with women without type 2 diabetes (table 4). However, we observed little overall relation between diabetes and decline on the global score (1.11, 0.90 to 1.37). Similarly, mean decline was greater among women with diabetes by −0.17 points (−0.33 to −0.01) on the TICS but was comparable in the two groups on the global score (mean difference in decline −0.01, −0.04 to 0.03). In addition, qualitative relations with longer duration diabetes and use of medication were generally similar to those observed with baseline cognitive function.

Discussion

In this large prospective study of women aged 70–81 years with type 2 diabetes who were living in the community we found that they had marginally worse baseline cognitive performance and greater cognitive decline than women without diabetes. Longer duration of diabetes resulted in larger associations. However, women who said they were on hypoglycaemic treatment seemed to have a similar likelihood of poor cognition as women without diabetes, while women not taking medication for diabetes or those taking insulin had worse performance.

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

A major strength of our study is the large sample size for assessing the relations between type 2 diabetes, duration, treatment, and cognition. Other strengths are the prospective assessment of diabetes and potential confounders over 25 years of follow up and the relative homogeneity of the sample in terms of education and access to health care, which should minimise confounding.

Limitations

Limitations should be considered. Firstly, as we relied on the women reporting their own diabetes status, we may have included some women with undiagnosed diabetes in the reference group, which could have led to underestimation of the true associations. However, undiagnosed diabetes was probably rare in these nurses. Among a random sample of those with no reported diabetes, plasma samples indicated just 2% had diagnostic signs of type 2 diabetes. Secondly, as in all studies of cognitive decline, there is regression to the mean on the repeat cognitive assessment. As women with type 2 diabetes had worse cognitive performance at baseline, regression to the mean would probably have attenuated the true magnitude of cognitive decline associated with diabetes.

In addition, there are important issues to consider in interpreting our findings regarding pharmaceutical treatment of diabetes. Participants who were not taking any treatment for diabetes probably included a heterogeneous group of women with untreated diabetes and diabetes controlled through

Table 4 Diabetes, duration of diabetes, use of medication for diabetes in women aged 70–81 in relation to cognitive decline over two years

		Odds ratio of substantial decline (95% CI)		Mean difference in cognitive decline (95% CI)	
	%	TICS (n = 16596)	Global score* (n = 14470)	TICS (n = 16596)	Global score* (n = 14470)
Diagnosis					
No diabetes	92.9	1.00	1.00	0	0
Diabetes:					
Adjusted for age and education	7.1	1.36 (1.12 to 1.65)	1.20 (0.97 to 1.47)	–0.29 (–0.44 to –0.13)	–0.03 (–0.06 to 0.00)
Multivariate adjusted†	7.1	1.26 (1.03 to 1.54)	1.10 (0.89 to 1.37)	–0.17 (–0.33 to –0.01)	–0.01 (–0.04 to 0.02)
Duration of diabetes (years)					
No diabetes	92.9	1.00	1.00	0	0
Adjusted for age and education:					
≤4	1.6	1.25 (0.83 to 1.88)	0.68 (0.40 to 1.17)	0.04 (–0.28 to 0.35)	0.05 (–0.01 to 0.12)
5–9	2.0	1.08 (0.74 to 1.59)	1.08 (0.73 to 1.59)	–0.10 (–0.38 to 0.18)	0.01 (–0.05 to 0.06)
10–14	1.6	1.35 (0.90 to 2.02)	1.53 (1.03 to 2.27)	–0.36 (–0.67 to –0.04)	–0.09 (–0.15 to –0.03)
≥15	1.9	1.77 (1.27 to 2.47)	1.51 (1.05 to 2.15)	–0.68 (–0.97 to –0.40)	–0.08 (–0.13 to –0.02)
P for trend		0.0004	0.005	<0.0001	0.001
Multivariate adjusted:					
≤4	1.6	1.15 (0.76 to 1.74)	0.65 (0.38 to 1.12)	0.14 (–0.18 to 0.46)	0.07 (0.01 to 0.13)
5–9	2.0	1.00 (0.68 to 1.47)	1.01 (0.68 to 1.49)	–0.01 (–0.29 to 0.27)	0.02 (–0.04 to 0.07)
10–14	1.6	1.26 (0.83 to 1.90)	1.40 (0.94 to 2.09)	–0.23 (–0.55 to 0.09)	–0.07 (–0.13 to 0.00)
≥15	1.9	1.64 (1.17 to 2.30)	1.35 (0.93 to 1.94)	–0.54 (–0.83 to –0.25)	–0.05 (–0.11 to 0.01)
P for trend		0.005	0.05	0.0004	0.05

Continued

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

Table 4 Continued

	%	Odds ratio of substantial decline (95% CI)		Mean difference in cognitive decline (95% CI)	
		TICS (n = 16 596)	Global score* (n = 14 470)	TICS (n = 16 596)	Global score* (n = 14 470)
Medication‡					
No diabetes	92.9	1.00	1.00	0	0
Adjusted for age and education:					
Insulin	1.5	1.49 (0.99 to 2.25)	1.22 (0.79 to 1.89)	–0.59 (–0.92 to –0.26)	–0.08 (–0.15 to –0.01)
Oral medication	3.1	1.12 (0.82 to 1.51)	0.82 (0.58 to 1.14)	0.00 (–0.22 to 0.23)	0.02 (–0.03 to 0.06)
No reported treatment	1.8	1.35 (0.93 to 1.95)	1.67 (1.18 to 2.37)	–0.27 (–0.56 to –0.03)	–0.02 (–0.08 to 0.04)
Multivariate adjusted:					
Insulin	1.5	1.39 (0.91 to 2.10)	1.08 (0.69 to 1.68)	–0.44 (–0.77 to –0.11)	–0.05 (–0.12 to 0.02)
Oral medication	3.1	1.09 (0.80 to 1.48)	0.77 (0.54 to 1.08)	0.07 (–0.16 to 0.30)	0.03 (–0.02 to 0.08)
No reported treatment	1.8	1.31 (0.90 to 1.90)	1.62 (1.13 to 2.30)	–0.23 (–0.53 to 0.06)	–0.02 (–0.08 to 0.05)

TICS = telephone interview of cognitive status.

*Global score combines TICS, test of verbal fluency, delayed recall of TICS 10 word list, digit backwards test, immediate and delayed recalls of East Boston memory test.

†Adjusted for age at interview (years), highest attained education (registered nurse diploma, Bachelor's degree, Master's or Doctoral degree), history of high cholesterol (yes, no), history of high blood pressure (yes, no), use of vitamin E supplement (currently yes, no), age at menopause (<50, 50–52, ≥53 years), body mass index (<22, 22–24.9, 25–29.9, ≥30 kg/m^2), cigarette smoking (current, past, never), antidepressant use (yes, no), alcohol intake (0, 1–4, 5–14, ≥15 g/day), use of aspirin (current use 1–5 times/week, use ≥6 times/week, no), use of other NSAID (current use, no), postmenopausal hormone use (currently yes, no), mental health index (0–52, 52–100), and energy-fatigue index (0–54, 55–100) from SF-36.

‡Data on medication use from questionnaire immediately before baseline cognitive assessment. Percentages do not total 100% as 0.8% who did not respond to medication question are not presented.

diet. Diabetes that can be controlled through diet may not be associated with poor cognition.[14] Thus, we have probably underestimated the effect of untreated diabetes. However, the increased odds of poor cognition associated with no treatment was similar across those with shorter and longer duration of diabetes (and duration is probably a good indicator of prevalence of dietary control), suggesting that our underestimate may be minimal.

What is already known on this topic

Many epidemiological studies have shown that type 2 diabetes increases the risk of cognitive decline, though most studies have been in men

Type 2 diabetes is associated with greater risk of cardiovascular disease in women than in men, and cardiovascular disease may increase the risk of cognitive decline

What this study adds

Women with type 2 diabetes have about 30% greater odds of poor cognitive function than those without diabetes, with a 50% increase after 15 years' of diabetes

Women with diabetes who did not report medical treatment had the highest risk of poor cognitive function and substantial cognitive decline

Women with diabetes who reported taking oral medication had a similar risk of cognitive decline as women without diabetes

Though our finding that insulin treatment was associated with poor cognitive performance is consistent with results of other studies of cognition,[8,14] it is difficult to draw conclusions; people with diabetes who use insulin all have longer

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

duration of diabetes, worse control, and higher prevalence of hypoglycaemic attacks, rendering it hard to adjust appropriately for confounding. None the less, there is growing evidence directly linking insulin to cognitive impairment: chronic hyperinsulinaemia[10] and incremental increases in serum insulin concentration after a glucose load[13] predict diminished cognition in the absence of diabetes or glucose intolerance. Moreover, insulin degrading enzyme regulates concentrations of both insulin and amyloid β in the brain[27] and infusion of insulin into healthy humans increases amyloid β concentrations in the cerebrospinal fluid,[28] further supporting a direct association between insulin and cognition.

Finally, consistent with our findings of similar cognitive performance among women taking oral medication and those without diabetes, in a controlled trial of participants with type 2 diabetes, Testa and Simonson noted that improved glucose control with oral medications resulted in better cognitive acuity, memory, and orientation.[29] In addition, an observational study of Mexican-Americans with diabetes reported significantly less cognitive decline in those with medical treatment than without.[30] Thus, although physicians may avoid prescribing oral therapy for diabetes in older people, it may be important to their cognitive health.

Conclusions

In conclusion, we found worse cognitive function and accelerated cognitive decline among women with type 2 diabetes, which seemed to be ameliorated with oral hypoglycaemic treatment. Studies have established that, in apparently healthy people, even modest differences in cognition result in substantially increased risks of dementia over several years.[6] Prevention and control of type 2 diabetes in women could have critically important public health consequences.

Contributors: GL and FG led the study design; GL, JHK, and FG contributed to the interpretation and the analysis of the data; JHK conducted the analysis of the data. FG was responsible for obtaining funding for the study. All authors contributed to writing the manuscript and are joint guarantors.

Funding: Grants AG15424 and CA87969 from the National Institutes of Health. FG is partially supported by a New Scholars in Aging award from the Ellison Medical Foundation.

Competing interests: During the last five years GL has received honorariums for lectures from Pfeizer and Lilly Pharmaceutical. During the past five years FG has received honorariums or temporary consulting fees from Novo Nordisk, Schering-Plough, Novartis, Orion Pharma, and Wyeth Ayerst.

Ethical approval: This study was approved by the Institutional Review Board of Brigham and Women's Hospital, Boston, MA.

References

1 Ott A, Stolk RP, Van Harskamp F, Pols HA, Hofman A, Breteler MM. Diabetes mellitus and the risk of dementia: the Rotterdam study. *Neurology* 1999;53:1937–42.

2 Leibson CL, Rocca, WA, Hanson VA, Cha R, Kokmen E, O'Brien PC, et al. Risk of dementia among persons with diabetes mellitus: a population-based cohort study. *Am J Epidemiol* 1997;145:301–8.

3 Curb JD, Rodriguez BL, Abbott RD, Petrovitch H, Ross GW, Masaki KH, et al. Longitudinal association of vascular and Alzheimer's dementias, diabetes, and glucose tolerance. *Neurology* 1999;52:971–5.

4 Luchsinger JA, Tang MX, Stern Y, Shea S, Mayeaux R. Diabetes mellitus and risk of Alzheimer's disease and dementia with stroke in a multiethnic cohort. *Am J Epidemiol* 2001;154:635–41.

5 MacKnight C, Rockwood K, Awalt E, McDowell I. Diabetes mellitus and the risk of dementia, Alzheimer's disease and vascular cognitive impairment in the Canadian study of health and aging. *Dement Geriatr Cogn Disord* 2002;14:77–83.

6 Bozoki A, Giordani B, Heidebrink JL, Berent S, Foster NL. Mild cognitive impairments predict dementia in nondemented elderly patients with memory loss. *Arch Neurol* 2001;58:411–6.

7 Brookmeyer R, Gray S, Kawas C. Projections of Alzheimer's disease in the United States and the public health impact of delaying disease onset. *Am J Pub Health* 1998;88:1337–42.

8 Gregg EW, Yaffe K, Cauley JA, Rolka DB, Blackwell TL, Narayan KM, et al. Is diabetes associated with cognitive impairment and cognitive decline among older women? Study of Osteoporotic Fractures Research Group. *Arch Intern Med* 2000;160:174–80.

9 Knopman D, Boland LL, Mosley T, Howard G, Liao D, Szklo M, et al. Cardiovascular risk factors and cognitive decline in middle-aged adults. *Neurology* 2001;56:42–8.

10 Kalmijn S, Feskens EJ, Launer LJ, Stijnen T, Kromhout D. Glucose intolerance, hyperinsulinaemia and cognitive function in a general population of elderly men. *Diabetologia* 1995;38:1096–102.

11 Scott RD, Kritz-Silverstein D, Barrett-Connor E, Wiederholt WC. The association of non-insulin-dependent diabetes mellitus and cognitive function in an older cohort. *J Am Geriatr Soc* 1998;46:1217–22.

12 Fontbonne A, Berr C, Ducimetiere P, Alperovitch A. Changes in cognitive abilities over a 4-year period are unfavorably affected in elderly diabetic subjects: results of the epidemiology of vascular aging study. *Diabetes Care* 2001;24:366–70.

13 Stolk RP, Breteler MM, Ott A, Pols HA, Lamberts, SW, Grobbee DE, et al. Insulin and cognitive function in an elderly population. The Rotterdam study. *Diabetes Care* 1997;20:792–5.

14 Elias PK, Elias MF, D'Agostino RB, Cupples LA, Wilson PW, Silbershatz H, et al. NIDDM and blood pressure as risk factors for poor cognitive performance. The Framingham study. *Diabetes Care* 1997;20:1388–95.

15 Rodriguez-Saldana J, Morley JE, Reynoso MT, Medina CA, Salazar P, Cruz E, et al. Diabetes mellitus in a subgroup of older Mexicans: prevalence, association with cardiovascular risk factors, functional and cognitive impairment, and mortality. *J Am Geriatr Soc* 2002;50:111–6.

16 Nguyen HT, Black SA, Ray LA, Espino DV, Markides KS. Predictors of decline in MMSE scores among older Mexican Americans. *J Gerontol A Biol Sci Med Sci* 2002;57:M181–5.

17 Wu JH, Haan MN, Liang J, Ghosh D, Gonzalez HM, Herman WH. Impact of diabetes on cognitive function among older Latinos: a population-based cohort study. *J Clin Epidemiol* 2003;56:686–93.

18 Vanhanen M, Kuusisto J, Koivisto K, Mykkanen L, Helkala EL, Hanninen T, et al. Type-2 diabetes and cognitive function in a non-demented population. *Acta Neurol Scand* 1999;100:97–101.

19 Lindeman RD, Romero LJ, LaRue A, Yau CL, Schade DS, Koehler KM, et al. A biethnic community survey of cognition in

Originally published in *BMJ* 2004; **328**. Reproduced with permission.

participants with type 2 diabetes, impaired glucose tolerance, and normal glucose tolerance: the New Mexico elder health survey. *Diabetes Care* 2001;24:1567–72.

20 Coker LH, Shumaker SA. Type 2 diabetes mellitus and cognition: an understudied issue in women's health. *J Psychosom Research* 2003;54:129–39.

21 Kang JH, Grodstein F. Regular use of nonsteroidal anti-inflammatory drugs and cognitive function in aging women. *Neurology* 2003;60:1591–7.

22 Brandt J, Spencer M, Folstein M. The telephone interview for cognitive status. *Neuropsychiatry Neuropsychol Behav Neurol* 1988;1:111–7.

23 National Diabetes Data Group. Classification and diagnosis of diabetes mellitus and other categories of glucose intolerance. *Diabetes* 1979;28:1039–57.

24 Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the expert committee on the diagnosis and classification of diabetes mellitus. *Diabetes Care* 2000;suppl 1:S4–19.

25 Manson JE, Rimm EB, Stampfer MJ, Colditz GA, Willett WC, Krolewski AS, et al. Physical activity and incidence of non-insulin-dependent diabetes mellitus in women. *Lancet* 1991;338:774–8.

26 Vickers A, Altman D. Analysing controlled trials with baseline and follow-up measurements. *BMJ* 2001;323:1123–4.

27 Farris W, Mansourian S, Chang Y, Lindsley L, Eckman EA, Frosch MP, et al. Insulin-degrading enzyme regulates the levels of insulin, amyloid beta-protein, and the beta-amyloid precursor protein intracellular domain in vivo. *Proc Natl Acad Sci USA* 2003;100:4162–7.

28 Watson GS, Peskind ER, Asthana S, Purganan K, Wait C, Chapman D, et al. Insulin increases CSF Abeta42 levels in normal older adults. *Neurology* 2003;60:1899–903.

29 Testa MA, Simonson DC. Health economic benefits and quality of life during improved glycemic control in patients with type 2 diabetes mellitus: a randomized, controlled, double-blind trial. *JAMA* 1998;280:1490–6.

30 Wu JH, Haan MN, Liang J, Ghosh D, Gonzalez HM, Herman WH. Impact of antidiabetic medications on physical and cognitive functioning of older Mexican Americans with diabetes mellitus: a population-based cohort study. *Ann Epidemiol* 2003;13:369–76.

Originally published in *BMJ* 2004; **328**. Reproduced with permission.