BMJ

BMJ 2014;348:g3557 doi: 10.1136/bmj.g3557 (Published 30 May 2014)

ENDGAMES

STATISTICAL QUESTION

Understanding statistical hypothesis testing

Philip Sedgwick *reader in medical statistics and medical education*

Centre for Medical and Healthcare Education, St George's, University of London, London, UK

Researchers assessed the efficacy of varenicline (a licensed cigarette smoking cessation aid) in helping users of smokeless tobacco to quit. A double blind, placebo controlled, parallel group, randomised controlled trial study design was used. The intervention was varenicline 1 mg twice daily. Treatment lasted for 12 weeks, with 14 weeks' follow-up. All participants were aged 18 years or more. They had been using smokeless tobacco for at least one year before recruitment, with no abstinence from smoking of longer than three months, but wished to quit. A total of 431 participants were recruited and randomised to varenicline (n=213) or placebo (n=218). All participants were offered brief behavioural support or counselling at the discretion of the investigators.[1]

The primary endpoint was continuous abstinence from smoking for four weeks at the end of treatment (weeks 9-12), confirmed by cotinine concentration. Statistical hypothesis testing was two sided, with a critical level of significance of 0.05 (5%). The rate of abstinence in the varenicline group was significantly higher than in the placebo group (59% *v* 39%; relative risk 1.6, 95% confidence interval 1.32 to 1.87; P<0.001).

Which of the following statements, if any, are true?

a) The alternative hypothesis states that, in the population sampled, treatment with varenicline is inferior or superior to placebo with regard to the primary endpoint

b) The research hypothesis states that, in the population sampled, treatment with varenicline is superior to placebo with regard to the primary endpoint

c) It can be inferred that the null hypothesis was not true

Answers

Statements *a* and *b* are true, whereas *c* is false.

The aim of the trial was to assess the efficacy of varenicline (a licensed cigarette smoking cessation aid) in helping users of smokeless tobacco to quit. Smokeless tobacco is often used by smokers trying to quit because it is considered less harmful than smoking. A randomised placebo controlled trial study design was used.

Sample estimates of percentage continuous abstinence from smoking were collected to estimate the effectiveness of varenicline versus placebo in the population. In statistics, the population is the entire group of people that the study aims to investigate. For the above trial, this would have been all users of smokeless tobacco who met the inclusion criteria. The treatment groups were compared with regard to the primary endpoint using traditional statistical hypothesis testing, which quantifies our belief that the collected data support a specified hypothesis about the population.

Statistical hypothesis testing involves the statement of the statistical null and alternative hypotheses. The researchers would have done this conceptually before the trial was started. Traditional statistical hypothesis testing starts at the position of equipoise as specified by the null hypothesis. For the trial above, the null hypothesis states that, in the population of users of smokeless tobacco from which the sample was obtained, no difference exists between treatment with varenicline and placebo in the percentage of continuous abstinence from smoking (for four weeks at the end of 12 weeks' treatment). The aim was to establish whether the sample data supported this position or provided evidence of a difference between treatment groups, as specified by the alternative hypothesis.

The alternative hypothesis states that a difference exists. In other words, it states that in the population sampled, the percentage of continuous abstinence from smoking for those treated with varenicline is not the same as in those taking a placebo. No direction is specified—the alternative hypothesis is two sided—treatment with varenicline could be inferior or superior to placebo in the primary endpoint (percentage of continuous abstinence for treatment; *a* is true).

It is important to distinguish between the research hypothesis and the statistical hypotheses. The researchers would have stated the research hypothesis, which predicts the study results, before starting the trial. The research hypothesis would have been that the outcome would be superior with varenicline compared with placebo (*b* is true); this would have been based on anecdotal evidence or perhaps on a pilot or exploratory study. The expectation that varenicline would increase the proportion of participants who continuously abstained from smoking provided the basis for undertaking a placebo controlled trial. The trial was necessary to obtain evidence that the intervention was

p.sedgwick@sgul.ac.uk

effective. Although the research hypothesis predicted that varenicline was superior to placebo in outcome, results are sometimes unexpected, so it was important that statistical hypothesis testing allowed for the possibility of placebo being superior. It is for this reason that a two sided statistical alternative hypothesis was used to compare treatment groups in the outcome measure.

The P value (P<0.001) in the above trial resulted from a statistical hypothesis test and was used to establish whether the sample data supported the null hypothesis or provided evidence of a difference, as specified by the alternative hypothesis. The P value is a probability and indicates how likely it is that an event will occur. It was derived using the sample data, and it represents the strength of evidence in support of the null hypothesis. A large P value suggests that the sample data support the null hypothesis, whereas a small P value suggests they do not. The cut off between a large and a small P value is conventionally set at 0.05 (5%), which is termed the critical level of significance. The P value for the statistical test of continued abstinence was P<0.001, which is less than 0.05. Therefore, there was little evidence to support the null hypothesis, and it was rejected in favour of the alternative hypothesis. There was a statistically significant difference in continued abstinence at the 0.05 level of significance—observation of the sample data shows that treatment with varenicline resulted in a greater proportion of continued abstinence from smoking than did treatment with placebo.

It is not possible to infer from the P value for the statistical test of continued abstinence that the null or alternative hypothesis is true or false (*c* is false). Sample data only ever provide evidence in support of the null or alternative hypothesis, in turn permitting inferences to be made about the population. This is because a further study, with a different sample of smokeless tobacco users, may give different results.

Care is needed when interpreting significance on the basis of a P value. It is important to consider the size of the difference between treatment groups in the outcome measure and its associated confidence interval. The size of the P value will depend on, among other factors, the sample size. Generally, trials with larger sample sizes tend to result in smaller P values and therefore show a statistically significant difference. However, a disadvantage of increasing the sample size is that, although differences between treatment groups in outcome measures are more likely to be statistically significant, they may not be clinically significant. Equally, trials with small sample sizes may result in differences between treatment groups in the outcome measure that are clinically significant but not statistically significant. The concepts of statistical significance and clinical significance have been described in a previous question.[2] Ensuring that a trial has a large enough sample size for a clinically significant difference to show as statistically significant underlies the concept of statistical power.[3] For the above trial the researchers will have considered the optimal sample size needed for a clinically significant difference between treatments, if it existed in the population, to show as statistically significant.

Competing interests: None declared.

1 Fagerström K, Gilljam H, Metcalfe M, Tonstad S, Messig M. Stopping smokeless tobacco with varenicline: randomised double blind placebo controlled trial. *BMJ* 2010;341:c6549.
2 Sedgwick P. Clinical significance versus statistical significance. *BMJ* 2014;348:g2130.
3 Sedgwick P. The importance of statistical power. *BMJ* 2013;347:f6282.

Cite this as: *BMJ* 2014;348:g3557