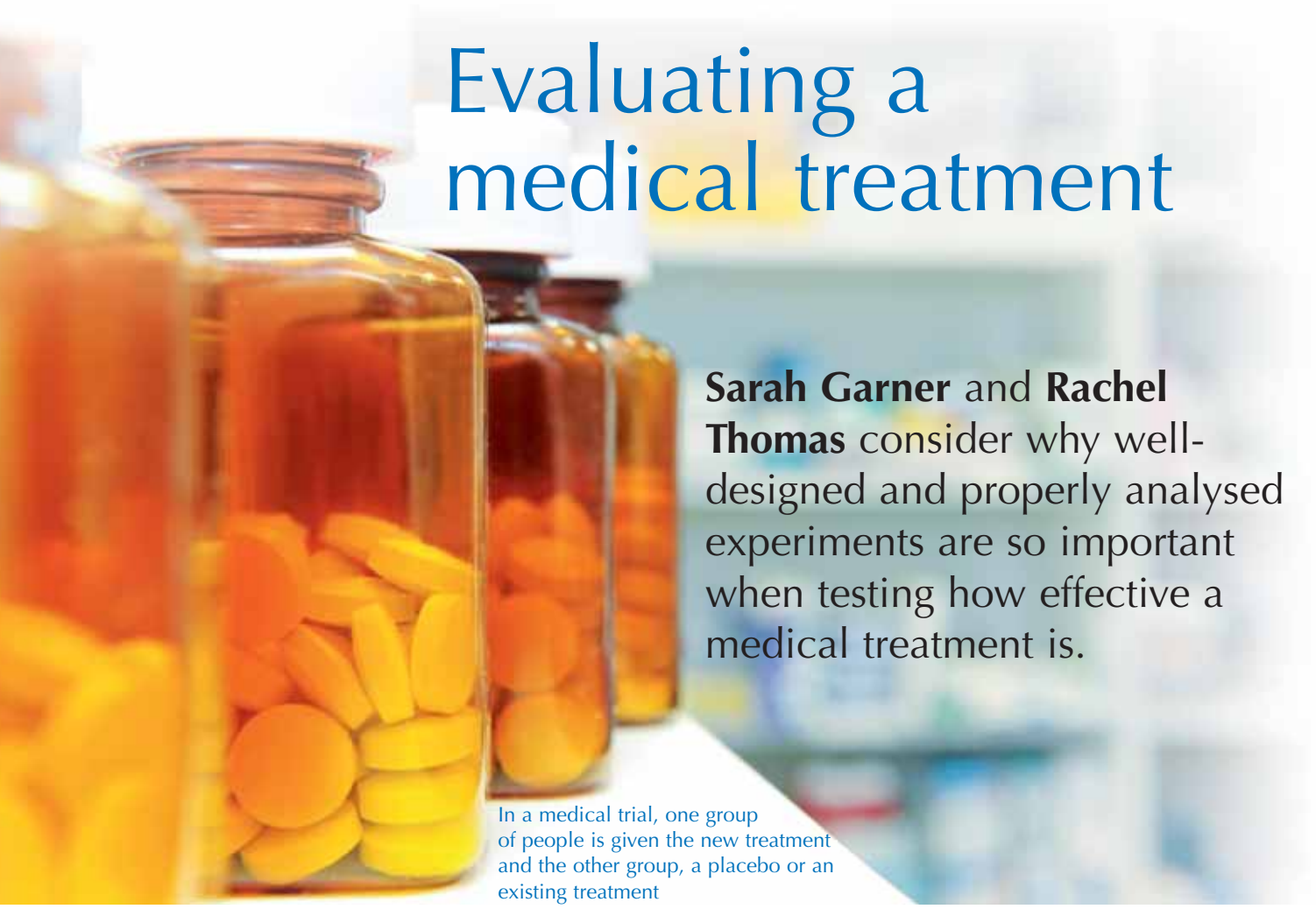
Evaluating a medical treatment

Sarah Garner and **Rachel Thomas** consider why well-designed and properly analysed experiments are so important when testing how effective a medical treatment is.



In a medical trial, one group of people is given the new treatment and the other group, a placebo or an existing treatment

Image courtesy of iStockphoto / gemphotography

Suppose a new medical treatment has been developed that may reduce high blood pressure. The treatment has been extensively tested in the laboratory and on a few volunteers, and the researchers believe that it will work on the general population. Now it is time to find out if they are right.

Historically, doctors found out whether a treatment worked in practice by using it on their patients. They could then compare the patients' responses to the new treatment and to previous treatments for the same illness, and also compare how responses to the new treatment varied between patients. However, if patients did indeed recover from their condition, there was no way of telling whether it was due to the treatment or to something else.

Image courtesy of iStockphoto / thelinke

How do we test whether a new drug for reducing blood pressure really works?

There are many other factors that could have caused the patients to recover: for example, they may have felt better simply because they were being treated by a doctor (a reaction known as the placebo effect); they may have recovered anyway, regardless of the treatment; or perhaps their recovery was due to changes in their personal circumstances or lifestyle. Without taking these and other factors into account, it could be easy to conclude incorrectly that the treatment worked. Doctors would then incorporate it into their everyday practice, mistakenly believing it to be effective.

The development of the randomised–controlled trial

In the 19th century, scientists proposed a method of controlling exactly what was happening and recording any changes in the patients' condition. In these controlled experiments, there were two groups of patients – the study group, which received the new treatment, and the control group, which received a placebo (an inert medication) or an established treatment. The patients were then observed, and the outcomes of the two groups (such as whether each patient lived or died) were recorded and compared.

Some time later, in 1917, the process of 'blinding' improved the scientific method even further. If neither the patient nor the researcher knows which treatment the patient is receiving, then the results cannot be interfered with either intentionally or unintentionally. This is known as a double-blind trial (in a single-blind trial, either the patient or researcher knows which treatment is being received).

However, the results could still be deliberately biased to prove that a treatment worked, by including sicker patients in the study group than in the control group. The solution to this, first used by the UK's Medical Research Council in the 1940s for its study of whooping cough vaccines, is to randomly choose which patients will get the new treatment, and which will get the control treatment.

Controlled trials with random allocation to the two groups became known as randomised–controlled tri-

REVIEW

- **Biology**
- **Mathematics**
- **Health education**
- **Statistics**
- **Ethics**
- **Sociology**
- **Ages 14+**

The article gives an insight into modern, evidence-based medicine, covering the often overlooked and seldom understood route from drug development to successful medical treatment. Statistical methods and their problems are discussed, providing interdisciplinary opportunities for teaching students aged 14 and over.

It is full of hot topics to be discussed with older students and teachers of different subjects. For example:

- 'Knowledge' is not static: as new side effects are reported or other new evidence is accumulated, the accepted knowledge can change.
- Often the clinical trials that are necessary before a drug can be marketed require more time than some very ill people have to live; who should be allowed to take part in a clinical trial – and which of those patients should get the control treatment and which the new drug?
- Why might the reporting of new treatments be biased?
- Do statistics give a false sense of security?

Picking up on the example of blood pressure and how variable this is, the class could measure their blood pressure and see how it varies from student to student. They could then run up and down the stairs a few times and see how one person's blood pressure can vary. Against this background level of variation, how do researchers determine the effect of drugs to lower blood pressure?

The article could also be used to spark off larger activities. For example, the students could be given newspaper articles about a piece of clinical research or a 'wonder drug', perhaps related to conditions they are familiar with, such as migraine, glucose intolerance or allergies. In groups, the students could use textbooks, the Internet and other sources of information to research:

- The disease being treated;
- Which treatments are available so far for that disease;
- Whether the new treatment has been tested on animals;
- Whether previous clinical trials on that treatment have been published;
- How the current clinical trial was designed and what statistical analysis was done;
- How they think the trial could have been improved.

On the basis of this research, each group of students could then write their own newspaper article about the clinical research. Do they think the original newspaper article was accurate? If not, why not?

For most teachers, the article will also be a valuable source of information on the history of medical research and randomised–controlled trials.

Friedlinde Krotscheck, Austria

als or RCTs. By randomising, you not only end up with a random distribution of sicker and healthier patients between the two groups, but also achieve a random distribution of things you do not know about (but which may also affect the patient's health and therefore the outcome of the treatment). Then – because, in theory, the only difference between the two groups is whether they received the treatment being tested – you can assume that any differences in outcome are most likely due to the treatment and nothing else.

RCTs are now universally used in clinical research to evaluate new treatments.

Designing and analysing RCTs

More people, more power

If you are planning to test your blood-pressure treatment with an RCT, you need to design it carefully. One important question is: How many patients should you include in the trial? This depends on how big an effect the new treatment has: the bigger the effect, the smaller the number of patients you need to distinguish it from the random fluctuations that happen by chance.

Of course, the effect of the treatment is exactly what you want to find out with your RCT. Before you start an RCT, however, you will already have some evidence that the treatment works, perhaps from laboratory or small-scale testing. This allows you to estimate the effect size.

Image courtesy of iStockphoto / wwing

Normal blood pressure is between 90 and 120 mmHg

In a healthy patient, blood pressure should be between 90 and 120 mmHg. But patients with high blood pressure will consistently have measurements of more than 140 mmHg, putting them at increased risk of heart attack and stroke. You might estimate that the new treatment will reduce a patient's maximum blood pressure by 5 mmHg: after treatment, you would expect that the average blood pressure of the study group would be at least 5 mmHg lower than the average blood pressure of the control group.

There are statistical formulae to determine the sample size you need to have a good chance of detecting that estimated effect[w2]. For your blood-pressure treatment, these formulae tell you that you would need around 64 patients in each group to detect a treatment difference of 5 mmHg[w3].

How different is different?

The trial has run its course, the participants have been monitored, and you have recorded a difference in blood pressure between the patients in the study and control groups. Thanks to randomising, you know that the two groups were comparable before the trial. So either your new treatment has had an effect, or a very surprising event has occurred: the treatment really has no effect at all and the difference you recorded in your RCT was due to chance alone.

Imagine that the average blood pressure of the study group was 5.2 mmHg lower than the average blood pressure of the control group. How do you decide if that difference is due to chance or to a real effect of the treatment? After all, blood pressure can vary for many reasons, not all of which can be controlled in your RCT.

What statisticians do is to allow for some variation; rather than rely on one average for each group, they calculate a range of values for each group that they are pretty confident will include the true value. This range of values is called a *confidence interval*. If the confidence intervals in your blood-pressure study are 141.2-148.9 mmHg in the control group and 133.7-139.3 mmHg in the study group, you can see that the two

Image courtesy of iStockphoto / hipokrat

Before running an RCT, the treatment is tested in the laboratory and on small groups of volunteers

Image courtesy of iStockphoto / scubaluis

For a medical trial to work, the people taking part must be representative of the real-world population of people to be treated

BACKGROUND

Evidence changes medical practice

Before 1994, doctors recommended that patients with lower back pain rest in bed. However, after reviewing all the available evidence, the Clinical Standards Advisory Group realised that bed rest was not beneficial and was perhaps even harmful. This led to a radical change in treatment, with patients being advised to remain active[w1].

confidence intervals do not overlap. Statisticians, therefore, say that the observed difference between the two groups is statistically significant – and you can assume that it really was caused by the treatment.

But how confident is confident? Statisticians usually say that 95% confident is good enough; this means that they are prepared to live with the fact that 5% of the time (or 1 in 20 times) they will be wrong due to chance alone. To be even surer that you have the right value, you have to measure more patients and even then, the only way to be 100% sure is to measure the whole population!

If the result turns out not to be statistically significant, one of the key questions to ask is whether you included enough patients in the trial. Perhaps the effect of the treatment is smaller than you estimated – with a larger sample size, you might have detected a difference between the two groups of patients.

Applying RCTs to the real world

A well-designed and properly analysed RCT is a very powerful tool for medical researchers – providing doctors with the information they need to make the right decisions when treating their patients. Nonetheless, RCTs do have limitations.

Firstly, it is not enough to know that the new treatment makes a statistically significant difference. Is the difference also clinically significant – for example, does a decrease in maximum blood pressure of 5 mmHg make a real difference to a patient's health and well-being? After all, in our example, the treatment still did not reduce the blood pressure to the

BACKGROUND

Evidence can change views

A systematic review of the evidence for minocycline, an antibiotic that was heavily promoted as the best cure for acne, was recently conducted to investigate its efficacy and its safety.

One side effect of minocycline is potentially fatal autoimmune liver problems. These problems are rare and can have a number of causes. Most doctors do not come across them, and even if they do, the connection might not necessarily be made with the drug.

It was only when all the information was reviewed together that the link was made. A systematic review showed that there was no evidence that minocycline was any better at curing acne than any other known treatment. Given the risks, the authors of the review concluded it should not be used in preference to other treatments (Garner et al., 2003).

normal range of 90-120 mmHg. To judge if this is clinically relevant, doctors may have to turn to other types of research.

A further limitation of RCTs is that patients in the trial may not represent the real-world population of people to be treated. Because trials aim to control as many factors as possible, they usually have strict inclusion and exclusion criteria. For example, pregnant women are not included due to potential risks to the unborn child; this meant that no one realised that thalidomide caused birth defects until it was introduced into general practice in the late 1950s[w4].

Then there is the question of how RCTs are reported. No one wants to publish bad news, particularly people who have spent time and effort to develop a new treatment. Historically, therefore, researchers did not publish trials that showed no difference or even that an older treatment was better. Some unscrupulous researchers have also reported selective or incomplete results, which made a new treatment look better than it really was. The research community has taken steps to stop both these biases by making companies and researchers register the start of a trial, so that it is more difficult to hide unfavourable outcomes, although there is still no requirement to report all outcomes. Journals are also standardising the information they require researchers to submit with their manuscripts, which makes it more difficult for bad results to be hidden.

Above all, RCTs are expensive and time consuming. As a result, many trials are not conducted at all, or their sample size or duration is limited. This may mean that the trial is not powerful enough to detect whether a treatment is effective, when in fact it is. Smaller trials may also miss important adverse effects (which may be rare), and shorter trials are unable to capture long-term outcomes.

Clinical researchers, therefore, often review the outcomes of a number of trials together in a meticulous analysis known as a systematic review – this effectively increases the sample size. Organisations such as the Cochrane Collaboration[w5] and the UK's National Institute for Health and Clinical Excellence[w6] base their recommendations to the medical community on systematic reviews.

Since the 1940s, the use of RCTs has significantly changed medical practice. Doctors are no longer reliant on their own observations but can rely on rigorous evaluation to ensure that the benefit of a new treatment outweighs the risks.

Acknowledgement

If you enjoyed this article but would like to learn more about the mathematics involved, read the original, longer version of this article[w3], which appeared in *Plus* magazine[w7], a free online magazine which opens a door to the world of mathematics with all its beauty and applications.

Reference

Garner SE (2003) Minocycline for acne vulgaris: efficacy and safety. *Cochrane Database of Systematic Reviews* **1**: CD002086. doi: 10.1002/14651858.CD002086

Web references

w1 – For more information about recommendations for bed rest, see the 'management' section of the article 'low back pain and sciatica' on the Patient UK website (www.patient.co.uk) or use the direct link: http://tinyurl.com/y9gghww

w2 – You can read a good explanation of how treatment effects and sample size can affect statistical power in Jerry Dallal's *Little Handbook of Statistical Practice*: www.jerrydallal.com/LHSP/sizenotes.htm

w3 – For the original version of this article, including more details of the

Image courtesy of stevecoleccs / iStockphoto

statistics, see:
Garner S, Thomas R (2010) Evaluating a medical treatment – how do you know it works? *Plus Magazine*. http://plus.maths.org/latestnews/jan-apr10/rct

w4 – To learn more about the thalidomide disaster and also about recent research into thalidomide and limb formation, see:

Zimmer C (2010) Answers begin to emerge on how thalidomide caused defects. *New York Times* **16 Mar**: D3. www.nytimes.com

w5 – The Cochrane Collaboration is an international network of people helping healthcare providers, policy makers, patients, their advocates and carers make well-informed decisions about human healthcare. See: www.cochrane.org

w6 – The National Institute for Health and Clinical Excellence (NICE) is an independent organisation responsible for providing UK national guidance on promoting good health and preventing and treating ill health. See: www.nice.org.uk

w7 – To learn more about *Plus*, the free online mathematics magazine, visit: http://plus.maths.org

Resources

For a brief description of the four phases of a clinical trial, see the information box in:

Wynne K, Bloom S (2007) Oxyntomodulin: a new therapy for obesity? *Science in School* **6**: 25-29. www.scienceinschool.org/2007/issue6/oxyntomodulin

Ledford H (2010) Companies pledge to make more trial data public. *Nature News* **15 Jun**. doi: 10.1038/news.2010.299.

To listen to the podcast that accompanied the original version of this article (*Plus* podcast 22, February 2010: Evaluating a medical treatment), visit: http://plus.maths.org/podcast

The charity Sense About Science has produced a useful guide about interpreting medical claims in the press (I've got nothing to lose by trying it). It can be downloaded free from the Sense About Science website (www.senseaboutscience.org.uk) or via the direct link: http://tinyurl.com/63zv4l

Freiberger M (2010) Medical research plagued by bad reporting. *Plus Magazine.* http://plus.maths.org/latestnews/jan-apr10/reporting

Plus Magazine offers a range of articles, podcasts and classroom activities addressing the mathematics behind heath and medicine: 'Do you know what's good for you?' See: http://plus.maths.org/wellcome

If you enjoyed this article, you might like to browse the other medicine-related articles in *Science in School*. See: www.scienceinschool.org/medicine

---

Dr Sarah Garner is the associate director for research and development at the National Institute for Health and Clinical Excellence (NICE)[w6], which bases its recommendations to the medical community on systematic reviews.

Rachel Thomas is co-editor of *Plus*[w7] magazine.