ARCH SOC ESP OFTALMOL. 2010;**85(1)**:2-10

ARCHIVOS DE LA SOCIEDAD ESPAÑOLA DE OFTALMOLOGÍA

www.elsevier.es/oftalmologia

Review

Evidence based ophthalmology: critical appraisal skills on clinical trials on treatments

J.C. Mesa-Gutiérrez[a,*], *A. Rouras-López*[b], *I. Cabiró-Badimón*[b], *V. Amías-Lamana*[b], *and J. Porta-Monnet*[a]

[a]*Ph.D. in Medicine, Master in Evidence-based Medicine, FEBO (European Board of Ophthalmologists), Ophthalmology Service, Esperit Sant Hospital, Santa Coloma de Gramenet, Barcelona, Spain.*
[b]*Graduate in Medicine.*

ARTICLE INFORMATION

Article history:
Received on Oct. 23, 2008
Accepted on Jan. 25, 2010

Keywords:
Randomization
Internal validity
Sampling
Allocation concealment
Worst case analysis
Intention to treat
Relative risk
Number needed to treat

A B S T R A C T

Purpose: To analyse and provide the tools to answer three questions that must be asked about a clinical trial on treatments: evaluation of validity (internal validity, so that results are not biased); clinical importance of these results and their application to individual patients. We provide simple statistics tools, clinical trial design, and clinical epidemiology for the evaluation and analysis of clinical trials on treatments.
Methods: Review of the medical literature.
Results: For the evaluation of papers on treatment and for using a treatment efficiently, we need to know if the clinical trials that support it are trustworthy, if the treatment has clinical importance and if it can be applied to my patients.
Conclusions: It is of paramount importance to check that the sample of the clinical assay has been correctly randomized, because randomization is necessary to comply with mathematical laws of application of statistical tests. Follow-up of patients must be complete in order to measure precision of results and evaluate the efficacy of treatment.

*Author for correspondence.
E-mail: jcarlosmesa@mixmail.com (J.C. Mesa-Gutiérrez).

Oftalmología basada en evidencias: evaluación crítica de los ensayos clínicos sobre tratamiento

RESUMEN

Objetivo: Analizar y proporcionar las herramientas para contestar a las tres preguntas que debe responder un ensayo clínico aleatorizado (ECA) sobre tratamiento: la evaluación de la validez de los resultados (o validez interna, que garantiza que los resultados del estudio no están sesgados); la importancia clínica de dichos resultados y si se pueden aplicar a pacientes individuales. Suministraremos una serie de conocimientos sencillos de estadística, diseño de investigaciones y epidemiología clínica que nos permitirán evaluar y analizar los ECA de tratamiento.

Método: Revisión de la literatura.

Resultados: Para evaluar artículos sobre tratamiento y para utilizar eficientemente un tratamiento necesitamos saber si los resultados de los ECA que los recomiendan son fiables, si tienen importancia clínica y si se pueden aplicar dichos resultados a mis pacientes.

Conclusiones: Es absolutamente fundamental comprobar que en el ensayo clínico exista una distribución aleatoria, ya que el muestreo aleatorio es imprescindible para que se cumplan los requisitos matemáticos de aplicación de las pruebas estadísticas. El seguimiento de los pacientes debe ser completo y debe medir la magnitud de los resultados y su precisión para así poder medir la eficacia del tratamiento.

Palabras clave:
Aleatorización
Validez interna
Muestreo
Ocultación secuencia aleatorización
Análisis de sensibilidad
Análisis por intención de tratar
Riesgo relativo
Número necesario de tratamiento

Introduction

Medical literature must be assessed to appraise its level of evidence. This is done for us by secondary journals such as *ACP Journal Club*, the Cochrane collaboration or, with increasing frequency, ourselves with the help of the so-called critical reading guides. Although this "evaluation" may seem complicated at first, it allows us to quickly determine if an article has sufficient interest to give it full attention and also to limit the number of articles we need to remain up to date. Basically, critical reading guides are verification lists which must be filled in with the data we obtain from the study, which will help us to decide about its adequacy to respond to clinical questions.

An article about treatment must address three aspects: validity of results, importance thereof and applicability to individual patients. In this revision we provide the tools to address the following points: assessment of the internal validity (which ensures that the study results are not biased), the treatment effect and thirdly its usefulness for our daily work.

One example: in the latest Congress of the Spanish Ophthalmology Society, you attend a presentation about the convenience of treating ocular hypertension by means of antihypertensive drugs. After the presentation a debate arises. When you return to your practice, you decide to search for more conclusive evidence. You will know that the highest evidence on the efficacy of a treatment is obtained by means of a randomized clinical trial and you search references through MEDLINE utilizing "ocular hypertension" and "medical treatment" as keywords, limiting the search to "randomized controlled trials" as article type. You get 18 references including trials with various drugs, comparisons between them and also a multicentre trial published in *Archives of Ophthalmology*[1], a journal available in your hospital library.

Table 1 shows the critical reading guide we shall utilize for articles on treatments. Similar guides can be obtained in

Table 1 – Critical reading guide for articles on treatments

Are the results of the study valid?
Primary criteria
Is there a clearly defined clinical question?
Was the allocation of patients random?
Was the randomization sequence maintained hidden?
Was the follow up of patients complete?
Were the patients analyzed in the group to which they were assigned?
Secondary criteria
Was a masked design maintained for patients, clinicians and research personnel?
Were the groups similar at the beginning of the trial?
Apart from the intervention, were both groups treated in the same way?
Which results were obtained?
Which was the magnitude of the treatment effect?
Which was the precision with which the effect was estimated?
Will the results be useful for me?
Can the results be applied to my patients?
Have all the clinically important results benn taken into account?
Do the treatment benefits compensate possible adverse effects and costs?

the websites of the *Critical Appraisal Skills Programme* Spain (CASPe)[2] network or the Evidence-based Medicine Centre of McMaster University[3].

To analyze the results of a randomized clinical trial (RCT), the reader must understand its design, development, analysis and interpretation. This can only be achieved if the authors have applied complete transparency. Several researchers and editors have developed the CONSORT statement (*Consolidated Standards of Reporting Trials*) to assist authors in their research and publication process by means of verification lists and flowcharts in various sections (title, abstract, introduction, material and methods, results and conclusions). The verification list comprises 22 points with essential information to determine the reliability or relevance of findings. In turn, the flow chart comprises information about the four main stages of a clinical trial (recruitment, intervention, follow-up and analysis)[4].

Subjects, material and method

We carried out a bibliographical revision of available information discussing the above topic. As Evidence-based Medicine (EBM) is a relatively recent discipline, it's working system is disseminated mainly through the Internet and therefore information is found mostly and freely in the Net. Accordingly, in this paper almost all references, spreadsheets and tools were obtained from the numerous sites and webs found in the net about this new "medical procedure style" which is progressively becoming established in the medical community.

After analyzing the information we selected a clinical trial on treatment published in an impact journal (*Archives of Ophthalmology*) and proceeded to a critical analysis by way of example[1].

Results: what can we expect of a clinical trial on treatment?

This question, necessary for the practice of EBM, comprises at least three elements which are resumed in the PICO acronym (Patient, Intervention, Comparison, Outcome), and cannot be responded by consulting a textbook but searching articles or systematic revisions. This type of question is like many that arise in our daily practice in relation to topics we are familiar with and utilize daily. The "C" for "comparison" in the acronym is not always necessary. A well formulated clinical question will greatly facilitate the search for evidence because it allows us to translate easily our terms into keywords (descriptors).

Let us begin the critical reading of the article. What shall we require of an article on treatment? The three generic questions we must answer are:

1. Are the results of the study valid?
2. What is the clinical importance of the results?
3. Are the results applicable to my patients?

Table 2 – Response to the clinical OHTS question

Participants	1,636 patients between 40-80 years, with IOP 24-32 mmHg in one eye and 21-32 mmHg in the contralateral eye; with open angle, 2 previous normal visual fields and with normal papillary stereoscopic photograph
Interventions	Anti-hypertensive topical treatment to diminish IOP 20% and < 24 mmHg
Results	IOP Reduction Aggregate open angle primary glaucoma prevalence Follow-up 78 months

IOP: intraocular pressure.

Validity analysis

Identification of the clinical question

By identifying the elements of the clinical question we will be able to determine its applicability in addition to evaluating its internal validity (lack of bias), i.e., the generalization of its results to our patients in the case that the sample utilized is representative. Table 2 shows the inclusion criteria of the chosen study, the interventions that must be compared and the recorded results, the main one being the development of a visual field defect or the involvement of the papilla.

The importance of randomization

The second question, and possibly the most important, is to determine if the trial was randomized. There are several reasons why it is necessary for a clinical trial to apply random distribution. All the statistical inferences and probabilities calculations are based on the study of randomly selected samples. Randomization is the only point of experimental design in which the randomness laws are explicitly introduced: these laws rule the distribution of theoretical frequencies that will be compared with observed frequencies. Accordingly, only randomization gives true sense to the use of statistical tests (the search of the "p" or 95% Confidence Interval [CI]) utilized to measure the power of randomness in the emerging results.

The second question is that randomization of large samples is the only known method to avoid the confusion bias. If we start an experiment with two comparable samples and we carry out an intervention on one of them but not on the other, if differences emerge between both samples at the end we can say that the cause of the differences is the intervention. On the contrary, if we start an experiment with samples that are not comparable, after carrying out an intervention in only one sample and differences emerge at the end, we will not be able to attribute the cause of these differences to said intervention. Thus, the comparability of samples is an essential requirement to attribute the cause, in order to conclude that the cause of differences in the final results is the only different variable between both groups, i.e., the intervention that is under study.

On the basis of the theorem known as the "Law of Large Numbers" (LLN)[5], the randomization of large samples tends

to produce uniform groups in all variables (including the unknown variables), prior to applying the intervention under study. As a consequence of this law, the control of confusion variables will be proportional to the sample size. The theorem is valid if the sample is random when n (the sample size) is 8. As in our clinical trials we will never have an infinite number of samples, it is essential to prove to the readers of our study that the effect of the LLN to standardize samples and control the effect of the main confusion factors has been fulfilled for the sample size we have chosen. The so-called "Table 1" of papers is utilized to achieve this. By consensus[4] it has been established that the first table of a clinical trial must show the frequency of appearance of the main demographic and/or confusion variables in both samples prior to the intervention. In said table we will not find any statistical evidence for comparing both groups, nor any "p" indicating the absence of statistically significant differences. "p" does not measure homogeneity apart from the fact that a small difference between two groups would reach statistical significance if the sample size was sufficiently large. In view of the results, the reader will be able to determine whether the groups are comparable by applying judgment. Table 3 shows the "Table 1" of the selected trial. Note the power of randomness to generate two very similar but not identical samples, as in the case of patients with a family history of glaucoma or association with myopia or high blood pressure, which are more numerous in the group treated with topical anti-hypertensive drugs. You, the reader, would have to determine on the basis of your knowledge and experience whether that difference is relevant or able to influence the result.

How can we verify that the randomization was correct?

The term "*random sample*" has a strict mathematical interpretation which is different from the usual meaning. Theoretically, for a sample to be random it is necessary that each member of the sample population must have the same probability of being selected for inclusion in the sample: The probability of allocation to different groups is *fixed and equal* for each and every one of the participating individuals. This requirement is not fulfilled if the sample is chosen fortuitously (that is, without a specific plan), systematically (odd or even numbers in clinical history or alternate days for visiting the practice) or conveniently (all those who returned a filled-in questionnaire).

In order to obtain a genuinely random sample of the population of patients, special maneuvers must be executed such as flipping a non-biased coin, utilizing a random numbers table or a computerized random numbers generator.

From now on we shall discuss simple randomization. As we have seen, randomization is the most sensitive part of the design. Therefore, to endow the trial results with validity, we must learn a few "tricks" to identify if the randomization has been done properly. The first, already mentioned above, is to check that randomness has generated similar or comparative samples by going through the data of the so-called "Table 1". But perhaps the most striking fact is knowing that a simple random sampling tends to produce a homogeneous sample *but with different sample sizes*. Most people think that, for example, when we randomly distribute a total sample of 40 patients in two groups (probability of belonging to each group=0.5), each of the two resulting samples must comprise 20 patients. Nothing further from the truth, because the probabilities theory allows us to calculate with binomial distribution (fig. 1) that the probability of obtaining that result is of 12. 54%.

Table 3 – Table I of the OHTS clinical trial

Characteristic	Medication (n=817)	Observation (n=819)	Total (n=1,636)
Sex, n (%)			
Male	359 (43.9)	346 (42.2)	705 (43.1)
Female	458 (56.1)	473 (57.8)	931 (56.9)
Age, n (%)			
40-≤50	291 (35.6)	287 (35)	578 (35.3)
>50-≤60	270 (33.0)	259 (31.6)	529 (32.3)
60-70	202 (24.7)	210 (25.6)	412 (25.6)
70-80	54 (6.6)	63 (7.7)	117 (7.2)
Race, n (%)			
Native American	1 (0.1)	3 (0.4)	4 (0.2)
Asian	4 (0.5)	10 (1.2)	14 (0.9)
Afro-American	203 (25)	205 (25)	408 (25)
Hispanic	24 (2.9)	35 (4.3)	59 (3.6)
Caucasian	577 (70.6)	560 (68.4)	1,137 (69.5)
Others	8 (1)	6 (0.7)	14 (0.9)
IOP; mean (SD), mmHg	24.9 (2.6)	24.9 (2.7)	24.9 (2.7)
Cup/horizontal disc ratio, DM	0.39 (0.19)	0.36 (0.18)	0.36 (0.18)
Cup/vertical disc ratio, DM	0.39 (0.20)	0.39 (0.19)	0.39 (0.19)
CV, DM, mean (SD) (dB)	+0.27 (1.07)	+0.21 (1.03)	+0.24 (1.05)
CV, mean deviation pattern (dB) (SD), dB	1.92 (0.21)	1.90 (0.21)	1.91 (0.21)
CV, mean corrected deviation pattern (SD), dB	1.12 (0.34)	1.12 (0.36)	1.12 (0.35)
Mean central corneal thickness (SD), microns	570.5 (38.9)	274.5 (37.7)	572.5 (38.4)
Prior use of ocular hypotensive medication, %	35	39.3	37.2
First degree familial glaucoma history	34	35.6	34.8
Myopia SE>1D, %	34.4	33.7	34.1
Oral beta blockers, %	5.4	4.6	5
Oral calcium channel antagonists	12.8	14	13.4
Migraine history, %	10.4	11.7	11.1
Diabetes history, %	11.5	12.1	11.8
Hypertension history, %	37.5	38.1	37.8
Hypotension history, %	4.8	4	4.4
Cardiovascular disease history, %	5.8	6.5	6.1
Infarct history, %	0.9	1.6	1.2

dB: decibels; SD: standard deviation; SE: spherical equivalent; IOP: intraocular pressure.

$$P(k)=[n!(n-k)!]\ p^{k}\ (1-p)^{n-k}$$

Figure 1 – Binomial probability law. n: total sample size; P(k): probability of obtaining k subjects.

It is surprising to find in the literature a large amount of small clinical trials (under 100 patients) in which the size of compared groups is equal when the sample is an even number (30/30, probability of presentation=10.3%; 40/40, probability=8.9%) or there is an imbalance of 1 when the total sample is an odd number (17/16, probability=13.6%; 25/26, probability 11%) when the authors stated in the methods that the distribution was random. The fact that they contradict the probability theory renders these studies suspect of the lack of true randomization even though its authors state that they have done it.

It is well known among mathematicians that random sampling tends to produce unequal sample size groups, to the extent that statistical experts working with small samples generally utilize a special sampling method such as the balanced permuted block sampling. Accordingly, a good way to determine whether randomization has been effectively carried out is that, if a simple randomized sample has been done, it must have a different size. If the groups are equal in number, the methods must specify the specific randomization system applied (for instance, balanced blocks).

In any case we already have empirical proof[6] that, regardless of the randomization methods, the factor which introduces the largest bias in results is not respecting the *randomization allocation concealment (RAC)*. The RAC is different from masking and consists in that, after the first patient has entered the study, it must be impossible for the staff administering the intervention to determine the group in which the next patient entering the study will be placed. For randomization to be correct, the RAC must be correct and to this end researchers utilize a number of methods. One of the most widely utilized methods is centralized randomization (through the telephone or the Internet) in a coordination centre that is different from the rest of centers that provide patients for the trial. Other valid RAC methods involve the utilization of opaque sealed envelopes containing the intervention (or the label thereof, the inclusion of masking, etc.) or the balanced block method for randomization, by changing the size of the blocks.

A secondary issue is the mask design vis-à-vis the treatment. In an ideal study, the patient, the clinician and the data analyst should not know the group of the patient, although this is not always possible. The case of surgical therapies vis-à-vis medical therapies is a good example, like medical treatment vis-à-vis observation as in the case of the article we have selected. In these cases, at least the authors will have made an effort to keep the analyst from knowing the group to which the patients belong.

Complete follow-up and analysis per intention to treat

Secondly, the follow-up of patients must be complete, which means that all patients recruited for the study must be taken into account up to completion, although in all studies there will be incomplete follow-ups. If the latter are excessive, the results of the study may be questionable because quite frequently the prognosis of patients without follow-up is different to that of the rest. How can we know if the follow up losses of a study invalidate their results? By means of the *worst case analysis*. This consists in assuming that all patients lost in the control group have evolved positively and all those lost in the treated group had evolved negatively, and recalculating the results. If the results of the study are not modified, the losses can be assumed. Otherwise, the strength of the result weakens in proportion to the probability that the treated patients may have evolved positively and in the control group patients may have evolved negatively.

Applying this criterion to our article, both branches of the study exhibited 84 and 89 follow-up losses up to 78 months (fig. 2). For the worst-case analysis, we would assume that all patients lost in the treatment group have developed glaucoma and those not treated in the control group have not. It seems that the results change: The risk of developing glaucoma in the treated group is modified (0.05 to 0.15%) whereas the observation group remains at 0.11%.

In addition to the follow-up losses, it is possible that patients allocated to one group or the other do not comply with the treatment, that it is not possible to apply it or even some patients changed from one group to the other. Losses can also be due to secondary effects or the lack of efficacy of the treatment. In any case, the patient shall be analyzed within the group to which he/she was allocated regardless of the treatment received. This is what is known as *analysis per treatment intention*. Although it may seem that if a patient has not received the allocated treatment, he/she must be excluded of the analysis of that branch, this is not the case. In many studies, the patients who did not take the medication evolved poorly even if they took placebo. In the case of patients with surgical pathology, maybe some are not submitted to surgery perhaps due to severity. If these patients are included in the control arm, even a futile operation will appear as effective because all the patients with the worst prognosis were assigned to the control group[7].

This form of analysis preserves the randomization effect and brings the study results closer to actual clinical practice. In fact, in our example, 702 of the 817 patients randomly allocated to the topical treatment group did receive the treatment. For the remaining 115 patients, the application was not possible but they were analyzed in that group anyway.

Equal treatment apart from the intervention

Both groups of the study should receive the same attention. If a stricter follow-up of one branch is decided, some facts which did not "appear" in the other branch might be identified and this could affect results. Interventions which are different from the treatment under study, also called "co-interventions", could also be a problem (for example, utilization of systemic beta blockers due to ischemic cardiopathy), above all it the physicians know about the treatment they are applying (not

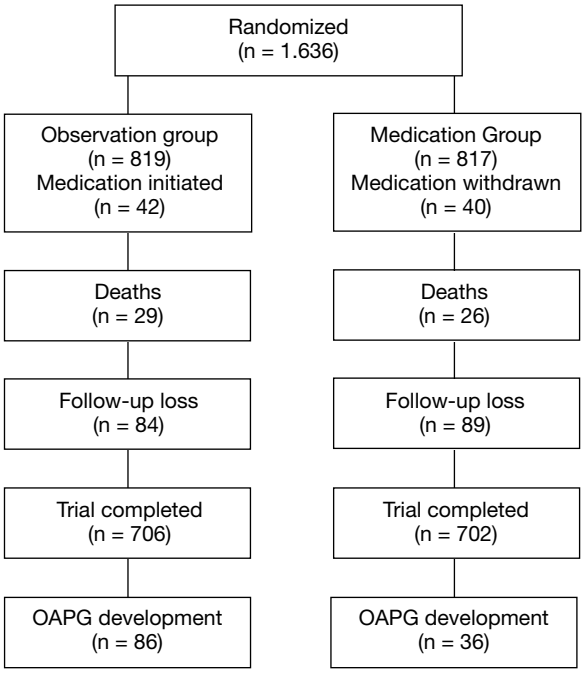

Figure 2 – Evolution of recruited patients in the clinical trial. OAPG: open angle primary glaucoma.

a double blind study) or the physician authorized the use of efficient therapies different to those under study[8].

Discussion

Is the article we selected valid?

As discussed above, the article specifically defines patients for the study by means of well-defined inclusion criteria, establishing the therapeutic alternative to be compared and the result to be measured. In fact, Table 3 illustrates all the collected variables, although the text defines the primary event as a development of alterations in the visual field or in papillary stereoscopic photographs, leaving the remaining events as secondary.

The study is defined as randomized, utilizing an analysis based on intention to treat and a minimization sequence to ensure a balance in the variables and a revision process to distinguish between visual field changes caused by glaucoma or other causes.

The size of both groups is credible and it can be seen that the follow-up was almost complete, with 84 and 89 losses per branch. The analysis was based on intention to treat, maintaining in the topical treatment branch the cases in which it could not be applied. Due to the characteristics of the study, it wasn't a double blind study: overall, 1,636 patients with ocular hypertension and without glaucomatous damages were randomized to receive treatment or for observation. However, the masked observer method was applied to determine papillary or visual field alterations which emerged during follow-up. We are certain that the analyst did not know the treatment branch of each case. A randomization central site was arranged.

In conclusion, this seems to be a valid article and therefore we would continue reading it to respond the following questions: effect of treatment and usefulness for our daily work.

¿Do results have clinical relevance?

The importance of results is determined by means of the *magnitude* of the effect as well as by their *precision*. Accordingly, statistically significance will not be utilized. In fact, the well-known "p" is indicating the probability of making a type 1 error, in other words stating that differences between treatments exist when they don't because we have found them only randomly. If the sample is large, small differences can produce statistically significant differences. It must be emphasized that statistical significance need not match clinically relevant differences. Thus, we might find that a 3-point difference in a pain scale from 0 to 100 is statistically significant but it would have no importance in daily clinical practice.

As regards the *magnitude* of the effect, in the case of continuous variables such as survival time or pain scale scores, the result would be expressed as a difference from mean or average values (bearing in mind that clinical differences and statistical differences need not match). However, usually a study utilizes binary variables(visual field involvement yes/no, papillary involvement yes/no, demise yes/no, tumor relapse yes/no, etc.). In this case the magnitude of the effect is expressed through the relative risk (RR), the relative risk reduction (RRR), absolute risk reduction (ARR) and number of patients to be treated (NNT). If the article does not provide these results, it should at least provide the data required for calculating them.

In the article being used as example, the authors did not provide said values but we were able to calculate them from table 3. The recorded event is the development of open angle primary glaucoma after 78 months of follow-up. Therefore, an RR below 1 indicates that the treatment under study has a protective effect.

The RR of the treatment is 0.40 or, in other words, the risk of developing glaucoma in treated patients is 0.40 times the risk of non-treated ones. This parameter is better understood if we utilize RRR: the risk of treated patients is reduced 60% compared to controls. If the RR were above 1 it would have a damaging effect because treated patients would exhibit a higher probability of developing glaucoma. Thus, a RR of 1.55 would indicate a 55% higher risk vis-à-vis the control group (table 4).

Neither RR nor RRR take into account the baseline risk of the population, but ARR does include it and this will allow us to calculate the absolute risk. ARR has the characteristic that it is small when the risks in groups are low, while RRR remains constant. By way of example, Table 5 illustrates the results obtained with two drugs compared against placebo to cure a disease. It can be seen how both drugs have the same RR and RRR, but ARR (and therefore NNT) is the parameter that indicates that the effect of drug A is higher because it

is applied on a population having a higher baseline risk. This peculiarity is frequently utilized by the pharmaceutical industry to promote its products, either emphasizing the RR or RRR and omitting ARR if it is very small, or providing data obtained in high risk populations but providing the product to low risk groups too, with frequently meaningless treatment benefits. In this example the importance of declaring conflicts of interest becomes apparent.

In any case, the best parameter to express the *clinical* efficacy of a therapeutic measure is NNT or number of patients to be treated with the experimental treatment —compared to what would have happened had they received the control treatment (placebo)— to avoid a negative event (for instance, developing glaucoma) or producing a positive event (for instance, remission). It is also the best parameter because that is exactly what the clinician needs to know, i.e., how many patients must be treated with the new treatment to cure one. In the case that adverse effects are also studied, the so-called Number Needed to Harm (NNH) is utilized, that is: how many patients must be treated to produce an undesirable effect. As shown in table 4, the NNT is obtained on the basis of ARR.

In the OHTS[3] study, for each 16.66 patients treated with antihypertensive drugs, one open angle primary glaucoma would be avoided, meaning that the treatment is extremely inefficient in this population group.

It is important to assess not only RRR, but also ARR. With an ARR=25%, NNT is 4, while an ARR=0.25%, NNT is 400, i.e., we would have to treat 400 patients to avoid one evolving from ocular hypertension ocular to glaucoma. NNT goes down while IOP increases.

In addition, we must also take into account that the majority of treatments have adverse effects which will also occur with a certain degree of frequency. For treatments with high NNTs it will be necessary to ponder possible adverse effects and cost. As a general rule, *with large NNT use only if the treatment is cheap, easy to apply and innocuous!*

Unfortunately we cannot determine with certainty the true risk reduction in treated patients. The results obtained in table 4 are merely an estimate of the true effect on the selected sample. If we had the entire population and not only this sample, would the effect be the same? To determine the true estimation we must calculate the *precision* of our result. If it is not very precise, the actual value may be far from the estimated effect. This precision is quantified by means of the CI calculation: we can estimate an interval where the true value will be found in 95% of cases. This 95% is accepted by consensus. It is possible to work with a CI of 90% or 99% but the higher the percentage the larger the population will have to be in order to estimate a tight confidence interval.

After obtaining the CI, we must determine whether the results are statistically significant. We must remember that for the RR the CI should not include the unit and for NNT it should not include a zero. Otherwise, no useful conclusion could be drawn from the study because any differences between both treatment will not have been demonstrated. In the case that the NNT does not include a zero (i.e., there are differences between treatments) the interval limits must be observed to decide, on the basis of clinical experience, if they seem acceptable. By way of example, the CAPRIE[9] trial for Aspirin® against clopidogrel for prevention of cardiovascular and cerebral ischemic events in a risk population — a trial financed by the pharmaceutical industry — even though the RR reduction was statistically significant (RRR 8.7%, p=0.043), the NNT was of 197, with a CI of 95% between 84 and 1,001. This means that it is possible that 1,000 patients would have to be treated with clopidogrel to cure one or more of those which would be cured by taking Aspirin®. If we compare the prices of both drugs (clopidogrel is much more expensive) it doesn't seem rational to treat an entire population at risk of suffering an ischemic event with the new drug. Clopidogrel can be efficient but its price renders it extremely inefficient.

If the article does not include the CI, it can be obtained by means of a calculator found in the CASPe[10] website (a simple spreadsheet), or by means of the formula supplied by Oxford University[11]:

Table 4 – Magnitude and decision on the effect in the OHTS study

	Name	Formula	OHTS	CI 95%
Re	Risk in patients exposed to study treatment	No. of events/ total patients in the branch	0.04 (4%)	3.1-6%
Rc	Risk in control group	No. of events/ total patients in the branch	0.10 (10%)	8.7-13.1%
RR	Relative risk	Re-Rc	0.4 (40%)	0.29-0.61
RRR	Relative risk reduction	1-RR or Rc-Re/ Rc	–0.58 (–0.58%)	–71.3 to –39%
ARR	Absolute risk reduction	Rc-Re	–0.06 (–6%)	–8.9 to –3.7%
NNT	Number of patients to be treated	1/ARR (if ARR expressed as proportion). 100/ARR (if ARR expressed as percentage)	–16.6	–28 to –12

Table 5 – Example with two imaginary drugs

	Death		RR	RRR	ARR	NNT
	Control	Experimental				
Drug A	0.2	0.12	0.6	0.4	0.08	12
Drug B	0.015	0.009	0.6	0.4	0.006	167

NNT: number of patients to be treated; ARR: absolute risk reduction; RR: relative risk; RRR: relative risk reduction.

IC95%ARR=ARR ± 1.96vRc(1-Rc)/N control patients + Re(1-Re)/N exposed patients

The upper and lower ARR limits are utilized to obtain the upper and lower NNT limits.

Will the results be useful for me?

Can these results be applied to my patients?

To answer this question we should consider whether we would have included our patients in the study had they been recruited. However, it may be easier to approach the issue the other way round and ask ourselves if there is any reason rendering the results of the study inapplicable to our patients.

A different question is if the patient matches a sub-group of patients included in the study. What to do in this case? The first is to see if the authors have made any subgroup analysis, which is relatively frequent when global results do not show a clear superiority of the treatment under study and the aim is to achieve better results in a subgroup. On many occasions these analyses were not planned at the initial study stage and are resorted to after obtaining the overall result. These types of analyses can be accepted if any of the following criteria is fulfilled[12]:

— High treatment difference between subgroups
— Low probability of differences being caused by randomness
— The subgroup analysis had been foreseen as a hypothesis in the study design phage
— The difference between subgroups is reproduced in other studies.

However, even when all the above criteria are met, in theory it would still be necessary to obtain confirmation through an *ad hoc* RCT.

Were all the clinically important results taken into account?

It is important that the variable resulting from the study *(end point)* is clinically relevant, demonstrating an improvement which justifies the use of the treatment. In our example, the improvement is preventing the development of glaucoma. However, some studies utilize substituted end points such as: 2-3 mmHg IOP reduction without specifying the lower limit (IOP reduction <24 mmHg), "hardly detectable" or "clinically significant" differences in the retinal ring. *The fact that the treatment improves these parameters does not make it clinically beneficial for the patient.*

The study must also consider possible deleterious effects of the treatment to enable the physician to assess the risk of applying it. In the study of our example, secondary results of the treatment included the development of systemic or ocular symptoms secondary to the treatment applied.

Do the treatment benefits counterbalance the possible adverse effects and costs?

Replying this question involves a global assessment of the benefits/drawbacks of the treatment in our patient together with other evaluations which may initially appear to be complex such as treatment cost. Here, personal experience is highly valuable to evaluate issues such as the difficulty in applying the treatment for technical reasons or due to low patient compliance, or even the logistical difficulties it may entail. If a new anti-hypertensive eye drop is being assessed, the funding for the treatment would have to be considered among other variables. Taken together, these questions can only be answered on the basis of experience and knowledge about the peculiarities of each case.

Conclusions: ocular hypertension and treatment

Even though it may seem complicated in the beginning, we have seen how, on the basis of data provided by a valid article and utilizing simple tools, it is possible to quantify the effect of giving treatment and derive useful conclusions for daily clinical practice. In our case the conclusions would be, in the first place, that not all individuals with high IOP have to be treated. The decision of initiating treatment must consider a number of factors such as:

— the low prevalence of open angle primary glaucoma between individuals with high IOP in the various population-based studies
— the convenience of a long-term treatment (adverse effects, cost)
— the individual risk of developing open angle primary glaucoma
— the individual probability of deriving benefit from the treatment
— the general condition of the individual's health and life expectancy

All the above justifies what Cochrane revisions usually indicate, i.e., that larger randomized studies are required and long-term revisions to establish guidelines and determine the type of benefits that can be obtained for these patients before establishing definitive indications for an expensive treatment which could involve complications.

Conflict of interests

The authors state that they have no conflict of interests.

REFERENCES

1. Kass M, Heuer D, Higginbotham E, Johnson C, Keltner J, Millar P, et al. The Ocular Hypertension Treatment Study: a randomized trial determines that topical ocular hypotensive medication delays or prevents the onset of primary openangle glaucoma. Arch Ophthalmol. 2002;120:701-8.
2. www.redcaspe.org/herramientas/lectura/11ensayo.pdf
3. www.cche.net/usersguides/therapy.asp

4. Moher D, Schulz KF, Altman D. The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials. JAMA. 2001;285:1987-91.
5. www.stat.berkeley.edu/~stark/Java/Html/lln.htm
6. Davidoff F, Haynes B, Sackett D, Smith R. Evidence based medicine: a new journal to help doctors identify the information they need. BMJ. 1995;310:1085-6.
7. Egger M, Ebrahim S, Smith GD. Where now for meta-analysis? Int J Epidemiol. 2002;31:1-5.
8. Guyatt GH, Sackett DL, Cook DJ. Users' guides to medical literature II. How to use an article about therapy or prevention. A. Are the results of the study valid? JAMA. 1993;270:2598-601.
9. CAPRIE steering committee. A randomised, blinded, trial of clopidogrel versus aspirin in patients at risk of ischaemic events. Lancet. 1996;348:1329-39.
10. www.redcaspe.org/herramientas/descargas/tratamientos.xls
11. www.cebm.net/?o=1023
12. Guyatt GH, Sackett DL, Cook DJ. Users' guides to the medical literature II. How to use an article about therapy or prevention. B. What were the results and will they help me in caring for my patients? JAMA. 1994;271:59-63.