REVISÃO

Conceitos básicos de epidemiologia e estatística para a leitura de ensaios clínicos controlados

Basic concepts in epidemiology and statistics for reading controlled clinical trials

Evandro Silva Freire Coutinho,[1] Geraldo Marcelo da Cunha[1]

Versão original aceita em Português

Resumo

Os autores apresentam conceitos básicos de epidemiologia e de estatística necessários para a compreensão adequada do desenho e dos resultados de ensaios clínicos controlados. No texto apresentam-se, através de exemplos, os conceitos de medidas de associação e de efeito, teste de significância estatística, p-valor, intervalo de confiança e poder do estudo, e discutem-se os erros mais comuns em suas interpretações.

Descritores*: Ensaios clínicos controlados; Eficácia; Estatística; Testes de hipótese*

Abstract

The authors present the basic concepts in epidemiology and statistics needed for understanding properly the design and results of controlled clinical trials. Through a set of examples, the concepts of measures of association and effect, statistical significance tests, p-value, confidence interval and statistical power are presented and common their misinterpretations are discussed.

Keywords: *Controlled clinical trial; Efficacy; Statistics; Hypothesis-testing*

[1] Departamento de Epidemiologia e Métodos Quantitativos em Saúde; Escola Nacional de Saúde Pública e Fundação Oswaldo Cruz

Financiamento: CNPq nr. 471145/03-3
Conflito de interesses: Inexistente
Recebido: 12.08.2004
Aceito: 19.08.2004

Correspondência
Evandro Silva Freire Coutinho
Rua Leopoldo Bulhões, 1480, 8º andar – Manguinhos
21041-210 Rio de Janeiro, RJ, Brasil
E-mail: evandro@ensp.fiocruz.br

Introdução

Os ensaios clínicos constituem-se numa poderosa ferramenta para a avaliação de intervenções para a saúde, sejam elas medicamentosas ou não. O primeiro ensaio clínico, nos moldes que hoje conhecemos, foi publicado no final da década de 40,[1] quando o estatístico Sir Austin Bradford Hill alocou aleatoriamente pacientes com tuberculose pulmonar em dois grupos: os que receberiam estreptomicina e os que não receberiam o medicamento. Desta forma, ele pode avaliar, de maneira não-viesada, a eficácia deste medicamento.

Em que pese a publicação crescente de ensaios clínicos controlados, alguns aspectos do desenho e da análise ainda são mal compreendidos e interpretados de forma equivocada. O objetivo deste artigo é apresentar alguns conceitos básicos de epidemiologia e de estatística presentes em grande parte desses estudos, assim como chamar a atenção para as peculiaridades e equívocos em sua interpretação. Para isso, abordaremos os seguintes tópicos:

1) Randomização;
2) Medidas de efeito: razão e diferença;
3) Testes de significância estatística;
4) Intervalo de confiança;
5) Poder do estudo.

Randomização

No ensaio clínico ideal para se avaliar a eficácia de um tratamento, um grupo de pacientes deveria receber o placebo e ser acompanhado por um período de tempo para se medir a ocorrência de certo evento (ex: óbito, cura). Em seguida, o pesquisador faria o tempo recuar a um momento imediatamente anterior à administração do placebo e administraria, a esse mesmo grupo de pacientes, o tratamento que se quer avaliar. O desfecho nessa segunda situação seria contabilizado e comparado com aquele observado na primeira situação. Por se tratarem dos mesmos pacientes, num mesmo momento de suas vidas, qualquer diferença quanto à ocorrência do desfecho (ex: óbito, cura) nas duas situações poderia ser atribuída, sem qualquer dúvida, à intervenção.

Como este desenho imaginário não é viável, os pesquisadores realizam uma randomização com intuito de gerar grupos comparáveis. Este procedimento consiste em alocar os indivíduos aleatoriamente (ao acaso) nos grupos a serem comparados. Com isso, busca-se constituir grupos com características muito semelhantes (comparáveis), com exceção das intervenções que se quer avaliar. Com a distribuição eqüitativa de fatores de risco ou de prognóstico, pode-se atribuir as diferenças observadas entre os grupos às intervenções que estão sendo comparadas. Embora a randomização não assegure a distribuição homogênea dos fatores nos grupos comparados em todas as ocasiões em que é implementada, a probabilidade de que isso ocorra aumenta conforme cresce o número de participantes no estudo.

O ocultamento do processo de randomização é importante para evitar manipulações da alocação que podem comprometer a comparabilidade dos grupos. Num ensaio clínico bem conduzido, a decisão de incluir ou não um paciente no estudo deve anteceder a sua randomização.

Ainda que a randomização constitua um aspecto central dos ensaios clínicos, não é raro encontrarmos estudos nos quais esse procedimento é implementado de forma inadequada. Estratégias de alocação por ordem de chegada, numeração corrida e dias da semana não devem ser usadas, pois facilitam a identificação da intervenção a que será submetido um paciente selecionado para o estudo. Com isso, o responsável pela alocação pode manipular o processo (ainda que inconscientemente), comprometendo a comparabilidade dos grupos.

Para tornar esse problema mais claro vamos imaginar que o pesquisador acredite que o novo tratamento é superior ao tratamento convencional. Se ele sabe que um paciente mais grave será alocado no grupo de tratamento convencional, ele pode não incluir esse indivíduo no estudo, aguardando a chegada de um paciente menos grave. Com isso, os grupos tendem a perder a desejada comparabilidade, ocorrendo um predomínio de pacientes mais graves no grupo que receberá o novo tratamento. O uso de uma seqüência aleatória de números, obtida através de tabelas de números aleatórios ou de algoritmos computacionais, facilita o encobrimento da seqüência de alocações e da conseqüente manipulação da alocação.*

A Tabela 1 ajuda a entender esse fenômeno. No caso das estratégias apresentadas nas colunas 1 e 2, basta que o responsável pela randomização descubra o dia ao qual cada tratamento está ligado para que ele desvende toda a seqüência da randomização e saiba em qual grupo será alocado o próximo paciente. No caso da coluna 3, mesmo que o pesquisador saiba que números ímpares correspondem ao tratamento A e números pares correspondem ao tratamento B, ele não tem como "deduzir" a seqüência das alocações. Estudos que adotam os procedimentos descritos nas colunas 1 ou 2 costumam ser denominados "quasi-experimentais", sendo mais vulneráveis a manipulações.

Tabela 1 – Randomização de 10 pacientes segundo três estratégias

Estratégia de randomização		
Numeração corrida	**Dias alternados**	**Números aleatórios**
Números ímpares: tratamento A Números pares: tratamento B	2ª, 4ª, 6ª feiras: Tratamento A 3ª, 5ª, sábado: Tratamento B	Números ímpares: Tratamento A Números pares: Tratamento B 3, 4, 7, 1, 8, 4, 0, 9, 6, 5
Seqüência: ABABABABAB	ABABABABAB	ABAABBBABA

Medidas de efeito

Considerando-se que o processo de randomização se deu de modo adequado, que os pacientes receberam as intervenções de modo apropriado e que as variáveis de interesse foram aferidas corretamente, a próxima etapa será a análise dos dados. Existem diferentes maneiras de se mensurar o desfecho de interesse em um ensaio clínico. Quando os participantes são classificados em dois grupos, segundo a presença

***NOTA:** Existem diversas metodologias para implementar a randomização de modo mais eficiente, como o procedimento por blocos, estratificado, não fixo. Esses métodos fogem ao escopo deste artigo e podem ser encontrados nos artigos de Pocock e Meinert et al. [2-3]

ou não de certo acontecimento, diz-se que esta variável é dicotômica. Por exemplo, os participantes podem ser classificados como vivos ou mortos, curados e não curados, com ou sem efeito adverso, e assim por diante.

Quando fazemos uso de uma variável dicotômica para classificar o desfecho dos participantes do estudo, podemos usar diferentes medidas para comparar o resultado observado entre os grupos de intervenção e de controle. Essas medidas são construídas através de razões ou de diferenças e trazem informações distintas.

1. Risco relativo ou redução relativa do risco (RR) - Eficácia

O risco é a probabilidade de ocorrência de certo desfecho. Varia entre 0 e 1 e pode ser transformado em percentual ao se multiplicar por 100.

Os dados apresentados na Tabela 2 foram extraídos de um ensaio clínico controlado[4] em que se alocou aleatoriamente 838 pacientes esquizofrênicos hospitalizados, de ambos os sexos, em dois grupos: clorpromazina e placebo. Os pacientes foram acompanhados por 24 semanas. Os dados referentes na tabela permitem estimar os riscos de agravamento dos sintomas psicóticos nos dois grupos de pacientes:

No grupo tratado: R(t) = 37/416 = 0,089 ou 8,9%.

No grupo controle: R(c) = 70/212 = 0,33 ou 33%.

Tabela 2 – Evolução de pacientes esquizofrênicos em uso de clorpromazina e de placebo

	Piora dos sintomas		Total
	Sim	Não	
Clorpromazina	37	379	416
Placebo	70	142	212
Total	107	521	628

Dados extraídos de Prein[2]

Após obtermos os riscos em cada grupo, uma maneira de compararmos as duas intervenções é através do cálculo de uma razão desses riscos, conhecida como risco relativo (RR). Desse modo, quando o risco nos dois grupos for o mesmo, o RR será igual a 1. Se o risco no grupo de intervenção for menor do que o risco no grupo controle, então o RR será menor que 1; caso contrário, ele será maior do que 1. No exemplo da Tabela 2, o RR é:

RR = R(t) / R(c) = 0,089 / 0,33 = 0,27.

Portanto, o grupo de pacientes com esquizofrenia que fez uso de clorpromazina apresentou um risco cuja magnitude equivale a 27% do risco encontrado para os pacientes que fizeram uso de placebo; isto é, a magnitude do risco no grupo que recebeu clorpromazina foi de aproximadamente ¼ da magnitude do risco no grupo placebo.

Pode-se ainda calcular a redução de risco relativo, também conhecida como eficácia, através da seguinte fórmula: RRR ou Eficácia = (1-0,27) x 100 = 73%.

A eficácia representa a redução relativa do risco obtida com a intervenção. No exemplo da Tabela 2, conclui-se que o uso da clorpromazina reduziu em 73% o risco de piora de pacientes.

No caso do tratamento provocar um aumento do risco de algum evento, teremos o excesso relativo de risco (ERR) calculado como (RR-1) x 100. No mesmo estudo apresentado na Tabela 2, o risco de distonia foi de 5% nos usuários de clorpromazina contra 2% no grupo com placebo, levando a um risco relativo de 2,5. Desse modo, o excesso relativo de risco foi de 150%.

ERR = (2,5-1) x 100 = 150%.

Em outras palavras, a clorpromazina elevou em 150% o risco de distonia em comparação com o grupo que recebeu placebo.

É preciso cautela ao se interpretar o RR, pois nem sempre um valor maior do que 1 indica algo ruim, indesejado. Tudo depende do modo como as variáveis estão sendo mensuradas. Num estudo com pacientes agitados/agressivos, realizado em três emergências psiquiátricas do Rio de Janeiro,[5] comparou-se o uso de midazolam IM (intramuscular) em relação à combinação haloperidol + prometazina (H+P), também por via IM. Observou-se que 89% dos pacientes do primeiro grupo foram tranqüilizados em até 20 minutos após o uso da medicação, contra 67% do segundo grupo. Nesse caso, o RR foi 1,33, o que significa um aumento de 33% da probabilidade de estar tranqüilo 20 minutos após o uso intramuscular do midazolam, em comparação com a combinação H+P.

2. Redução absoluta de risco (RAR)

A RAR representa a redução, em termos absolutos, do risco no grupo que sofreu a intervenção de interesse, em relação ao grupo controle.

RAR = [R(c) – R(t)] x 100

No caso do estudo da Tabela 2, a RAR foi de 24,1%.

RAR = (0,33-0,089) x 100 = 24,1%

Para entendermos melhor o conceito de RAR e a sua diferença em relação à redução do risco relativo (RRR), observemos os dados fictícios apresentados na Tabela 3.

Tabela 3 – Reinternações e recaídas em pacientes com psicose, em dois grupos de tratamento (dados fictícios)

	Reinternação			Recaída		
	Sim	Não	Total	Sim	Não	Total
Tratamento	5	95	100	30	70	100
Placebo	10	90	100	40	60	100
Total	15	185	200	70	130	200
	RRR = [1-(0,05 / 0,10)] x 100 = 50% RAR = (0,10-0,05) x 100 = 5%			RRR = [1-(0,30 / 0,40)] x 100 = 25% RAR = (0,40-0,30) x 100 = 10%		

No caso do desfecho reinternação, temos uma eficácia (RR) do tratamento de 50% e uma redução absoluta de risco (RAR) de 5%; isto é, o tratamento reduziu o número de reiternações à metade (de 10% para 5%), o que representou a eliminação de um total de 5% desses eventos. Quanto ao desfecho recaída, a eficácia foi de 25% e a redução absoluta de risco foi de 10%; em outras palavras, o tratamento reduziu o número de recaídas em apenas ¼ (de 40% para 30%), mas isso representou a eliminação de um total de 10% desses eventos. Portanto, ainda que a eficácia da intervenção seja maior para o desfecho reinternação, o maior benefício se dá para o desfecho recaída, onde houve uma redução de 10% do total de casos contra 5% do total das reinternações. Isso ocorreu porque a freqüência de recaídas é maior do que a freqüência de

reinternações. A Figura 1 ajuda a entender essa conclusão. Sendo a área clara das barras aquela correspondente à diminuição das reinternações e recaídas, observa-se uma redução de maior volume no evento recaída. As reinternações foram reduzidas à metade, mas isso representou um volume menor do que aquele alcançado para as recaídas.

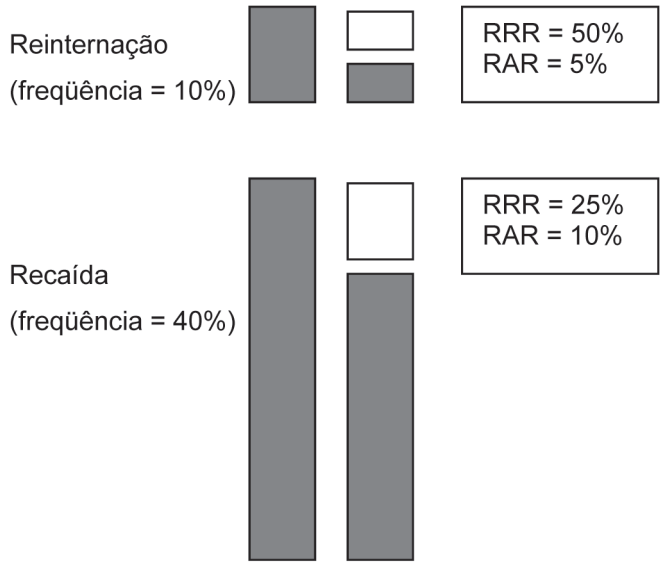

Figura 1 - Comparação da eficácia e da redução absoluta de risco para dois desfechos (baseada nos dados da Tabela 3)

3. Número necessário para tratar (NNT)

Um modo adicional de se medir o impacto de uma intervenção que vem se tornando popular nos últimos anos é o número necessário para tratar (NNT). Essa medida representa o número de pacientes que se precisa tratar para se prevenir um evento indesejado (ex: morte, recaída). O NNT é calculado como o inverso da RAR. No caso da Tabela 2, onde a RAR foi de 0,241 (ou 24,1%), o NNT será 1/0,241 = 4. Portanto, previne-se um caso de piora dos sintomas psicóticos em cada quatro pacientes com esquizofrenia que fazem uso de clorpromazina.

Vimos no item anterior que a RAR é influenciada pela freqüência do evento que se está avaliando. Pelo fato do NNT ser o inverso da RAR, ele também será influenciado pela freqüência do evento. No caso dos dados da Tabela 3, temos um NNT de 10 para o evento recaída contra um NNT de 20 para o evento reinternação. Em outras palavras, para evitar uma reinternação seria necessário tratar o dobro de pacientes do que seria necessário para impedir uma recaída.

É importante fazer uma distinção entre os efeitos benéficos e os efeitos indesejados da intervenção. No caso desses últimos, o NNT é denominado número necessário para produzir um dano (NNH – *number needed to harm*).

4. Diferença de médias

Existem ensaios clínicos nos quais o desfecho é medido através de escores de escalas como, por exemplo, a *Brief Psychiatric Rating Scale* (BPRS) ou a *Abnormal Involuntary Movement Scale* (AIMS). Essas escalas produzem escores para cada paciente, ao invés de resultados dicotômicos do tipo "sim/não". Esse tipo de variável é denominada contínua, sendo comum o cálculo de sua média nos dois grupos que se deseja comparar. Para avaliar o tratamento de melhor resultado, costuma-se comparar as médias dos dois grupos ao longo do estudo. Em outras ocasiões calculam-se esses escores no início e no final do tratamento, e compara-se a mudança desses escores em cada grupo.

O estudo de Borison et al compararam, entre outros desfechos, os escores médios do BPRS ao final de 8 semanas nos pacientes alocados para o grupo da Clorpromazina com os escores observados entre pacientes do grupo placebo. A média do grupo que recebeu clorpromazina foi 46,4, contra 50,5 no grupo placebo.[6]

Um dos problemas desse tipo de desfecho é que, embora seja possível afirmar que os pacientes que fizeram uso de clorpromazina tivessem uma pontuação mais baixa para os sintomas psiquiátricos, é difícil extrair um significado clínico dessa diferença. É mais fácil entender uma redução de 25% das recaídas do que uma diferença de 4,1 pontos numa escala de sintomas psicóticos.

Nível de significância - valor de p

Até o momento apresentamos diferentes medidas para estimar o tamanho da diferença de um determinado evento (ex: reinternação, agravamento dos sintomas) em grupos expostos a diferentes intervenções (ex: clorpromazina *vs* placebo). No entanto, ao lermos os ensaios clínicos é freqüente nos deparamos com expressões do tipo "a diferença entre os grupos foi estatisticamente significativa" ou "$p < 0,05$".

O que se deseja com essas expressões é discutir o papel do acaso nos resultados obtidos em um ensaio clínico. Em outras palavras, ainda que um estudo estime uma eficácia de 30%, esta diferença entre os grupos pode ser casual.

Em estatística, uma das maneiras de abordar essa questão é avaliando a evidência contra o que se denomina hipótese nula, segundo a qual não existe diferença entre os efeitos das intervenções que se está comparando. A força evidência contra a hipótese nula é avaliada através do valor de p, que representa a probabilidade de se observar uma diferença entre os grupos como a que foi encontrada no estudo, quando, na verdade, esta diferença não existe. O valor de p também é chamado de nível de significância e, quanto menor ele for, maior a evidência contra a hipótese nula. Por se tratar de uma probabilidade, o valor de P varia entre 0 e 1.

Os dados de três ensaios clínicos controlados comparando clorpromazina com placebo (Tabela 4) ajudam a entender essa questão. Todos os estudos foram conduzidos com pacientes hospitalizados, de ambos os sexos. Clark et al randomizaram 55 pacientes, os quais foram acompanhados por 12 semanas,[7] enquanto o estudo de Hall et al incluiu 175 pacientes, acompanhados por 66 dias.[8] O estudo de Ban et alocou aleatoriamente 30 pacientes, sendo o período de seguimento de 12 semanas.[9] A hipótese nula no caso desses estudos é que a clorpormazina e o placebo não diferem quanto aos seus efeitos sobre a sintomatologia psicótica. Os testes estatísticos apresentados na última linha da tabela mostram que, no caso do ensaio de Hall et al, o valor de p é 0,01 (ou 1%).[8] Portanto,

Tabela 4 – Comparação da clorpromazina *versus* placebo em três ensaios clínicos

Estudo	Hall, 1955[8]		Ban, 1975[9]		Clark, 1972[7]	
	Ausência de melhora	Total	Ausência de melhora	Total	Ausência de melhora	Total
Clorpromazina	56	87	6	10	10	19
Placebo	72	88	8	10	14	18
Risco relativo	0,79		0,75		0,68	
Eficácia	21%		25%		32%	
p-valor	0,01		0,63		0,11	

a probabilidade de se observar uma eficácia de 21% em favor da clorpromazina, quando esta não difere do placebo, é de apenas 1%. No caso do estudo de Ban et al, a probabilidade de se encontrar uma eficácia de 25% na ausência de uma superioridade da clorpromazina em relação ao placebo é de 63%.[9] Desse modo, o estudo de Hall et al apresenta uma forte evidência contra a hipótese nula (valor baixo de p),[8] enquanto no caso do estudo de Ban et al a evidência contra a hipótese nula é fraca (valor alto de p).[9] As razões para essas discrepâncias serão discutidas adiante.

Para efeito de tomada de decisão, muitos ensaios clínicos consideram a probabilidade menor do que 5% ($p < 0,05$) como o valor limite para considerar que um efeito observado no estudo é real, não sendo decorrente do acaso. Isto é, a hipótese nula será rejeitada caso o valor de p seja inferior a 0,05. Em outras palavras, quando a probabilidade de concluirmos equivocadamente que uma intervenção é superior à outra for menor que 5%. Esse erro é denominado erro tipo I ou α.

Embora esse limite de 5% para aceitar ou rejeitar a hipótese nula seja habitual em ensaios clínicos, não existe nenhuma obrigação de que o valor de p seja fixado nesse nível. Dependendo dos riscos em se assumir uma conclusão falso-positiva, esse valor pode ser reduzido.**

Os resultados dos testes de significância estatística, através de seus valores de p, costumam ser interpretados equivocadamente como medidas da magnitude do efeito de uma intervenção. Quem mede a magnitude do efeito de uma intervenção é a eficácia, a redução absoluta de risco, o número necessário para tratar. Os valores de p apenas informam a probabilidade de que uma associação, identificada no estudo, seja um achado falso-positivo decorrente do acaso. Em outras palavras, um valor de p igual a 0,10 ou 10% significa que existe uma probabilidade de 10% de se encontrar uma eficácia como a observada no ensaio clínico na ausência de superioridade de uma das intervenções.

Considerando-se um nível de significância de 5% e observando as estimativas de eficácia, vemos que o estudo de Clark et al foi aquele com maior eficácia (32%), embora não seja estatisticamente significativo ($p = 0,11$).[7] Por outro lado, o estudo de Hall et al foi o único a apresentar significância estatística ($p = 0,01$), embora seja o de menor eficácia (21%).[8] Portanto, um nível de significância ou valor de p baixo (ex: p = 5% ou 1%) não quer dizer que exista uma forte associação (ex: grande eficácia), mas apenas que existe uma forte evidência de que o efeito observado não seja decorrente do acaso.

Mas se o teste de significância estatística não avalia a magnitude da associação (eficácia, neste caso), por que ele só foi significativo no estudo de Hall et al, exatamente aquele com menor eficácia? Porque esses testes dependem não só da magnitude da eficácia, mas também do tamanho da amostra.[8] Em outras palavras, o nosso grau de certeza de que um efeito observado não decorre do acaso aumenta quando temos um número maior de indivíduos no estudo. No caso da Tabela 4, o estudo de Hall et al tem uma amostra cerca de 5 vezes maior que a do estudo de Clark et al e cerca de 8 vezes maior que no estudo de Ban et al.[7-9]

É importante ainda ressaltar que o fato de um resultado não ser estatisticamente significativo não deve ser interpretado como evidência de ausência de efeito da intervenção, mas sim de que as evidências contra a hipótese nula são fracas. Na Tabela 5 apresentamos dados fictícios sobre o risco de efeitos adversos de dois medicamentos (A e B). Observa-se que a única distinção entre os estudos 1 e 2 é o tamanho do grupo investigado, já que a redução de efeitos adversos observada para o tratamento A é a mesma em ambos os casos: 34%.

Tabela 5 – Proporção de efeitos adversos observados em quatro ensaios clínicos (dados fictícios)

	Efeitos adversos			
	Sim	Não	Total	
Estudo 1				
Droga A	4	146	150	RRR = 34%
Droga B	6	144	150	P-valor = 0,52
Estudo 2				
Droga A	40	1460	1500	RRR = 34%
Droga B	60	1440	1500	P-valor = 0,04
Estudo 3				
Droga C	30	270	300	RRR = 50%
Droga D	60	240	300	P-valor < 0,001
				IC 95%: 25%-67%
Estudo 4				
Droga C	300	2700	3000	RRR = 50%
Droga D	600	2400	3000	P-valor < 0,001
				IC 95%: 43%-56%

Se arbitrarmos um nível de significância de 5%, valores de p acima desse nível levarão à aceitação da hipótese nula e valores de P abaixo desse nível levarão à sua rejeição. Nesse caso, o estudo 1 não permitiria concluir que a droga A apresenta um risco de efeitos adversos superior ao da droga B, pois o valor de P é alto (0,52); isto é, a probabilidade de que se trate de um achado ao acaso está acima do limite de 5% que arbitramos a priori. No entanto, no estudo 2, nossa conclusão seria de que o tratamento A está mais sujeito à ocorrência de efeitos adversos dado que o valor de P é baixo (0,04); isto é, a probabilidade de que este achado decorra de uma casualidade é menor do que o limite de 5% que estabelecemos a priori. Do mesmo modo que na Tabela 4, essa mudança na conclusão deve-se a um aumento do tamanho amostral.

****NOTA:** Parece que esta opção pelo valor de 5% vem dos escritos do estatístico Sir R. A. Fisher, que expressou sua preferência por este ponto de corte.

Intervalo de confiança

A cada dois anos somos expostos aos resultados das pesquisas eleitorais sobre as preferências dos eleitores. O percentual de votos de cada candidato é apresentado sempre seguido da seguinte informação: "a margem de erro da pesquisa é de 2% ou de 3%". Isso significa que sempre que fazemos uma pesquisa, seja ela eleitoral ou um ensaio clínico, utilizando uma fração da população, existe certo grau de incerteza sobre o real valor da estimativa que fazemos.

O intervalo de confiança define os limites inferior e superior de um conjunto de valores que tem certa probabilidade de conter no seu interior o valor verdadeiro do efeito da intervenção em estudo. Desse modo, o processo pelo qual um intervalo de confiança de 95% é calculado é tal que ele tem 95% de probabilidade de incluir o valor real da eficácia da intervenção em estudo.

Na Tabela 5 estão os dados fictícios de dois estudos (estudo 3 e estudo 4) comparando a proporção de pacientes com efeitos adversos observados ao longo do tratamento com dois neurolépticos (C e D). Ambos os estudos tiveram redução no risco de efeitos adversos de 50% com o uso do medicamento C e um nível de significância estatística menor que 0,001 ou 0,1%. Entretanto, o intervalo de confiança do estudo 4 é mais estreito do que o intervalo do estudo 3. Por essa razão, dizemos que o estudo 4 é mais preciso do que o estudo 3, pois a região de incerteza quanto ao verdadeiro valor da RR é menor. No caso do estudo 3, há uma probabilidade de 95% do intervalo entre 25% e 67% conter o valor verdadeiro da RRR, enquanto no caso do estudo 4 este intervalo varia apenas de 43% a 56%. Não há obrigatoriedade de que o intervalo de confiança seja de 95%, podendo ser de 90%, 99% ou ainda outro valor diferente.

O uso do intervalo de confiança permite não só conhecermos a precisão com que o estudo estima certo efeito, como também possibilita dizermos se o achado é estatisticamente significativo para um dado nível de significância. Quando o intervalo de confiança contiver o valor nulo de efeito, o estudo será inconclusivo (sem significância estatística). Como vimos anteriormente, entende-se por valor nulo de efeito o valor que expressa riscos iguais em ambos os grupos. No caso do RR, da eficácia e da RAR os valores nulos são um, zero e zero, respectivamente.

Voltando aos estudos da Tabela 4, o ensaio clínico de Hall et al,[8] cujo valor de p de 0,01 foi significativo, tem um intervalo de confiança de 95% para a RR que exclui o valor nulo um (0,65-0,95). Já o ensaio clínico de Ban et[9] al apresentou um valor de p não significativo de 0,63, o que se expressa num intervalo de confiança de 95% que inclui o valor nulo (0,41-1,36).

Poder do estudo

O poder de um ensaio clínico pode ser definido como a probabilidade do estudo identificar uma diferença entre os tratamentos (efeito), quando esta diferença é real. O poder é influenciado por quatro fatores: a natureza do teste estatístico, o nível de significância, o tamanho da amostra e a diferença esperada no efeito dos dois tratamentos.

Na Tabela 4, o poder do estudo de Hall et al foi de 77%, enquanto o poder no ensaio de Ban et al[9] foi de 16%. Como os estudos observaram eficácias bastante próximas, o que está levando a poderes tão distintos é a diferença no tamanho amostral.[8]

Levando em conta o conceito de poder, fica mais claro porque um estudo com resultado sem significância estatística não pode ser interpretado como evidência de ausência de efeito. Pode ser apenas um caso de falta de poder estatístico para evidenciar este efeito.

Por esta razão, é de extrema importância para os ensaios clínicos que:

1) O tamanho amostral propicie um poder elevado. Estudos para detectar efeitos pequenos necessitam amostras maiores.

2) O estudo informe o poder, sobretudo quando seus resultados não alcançam significância estatística. Se o poder for baixo, nada se pode concluir. Se o poder for alto, pode-se considerar, com um pouco mais de segurança, que os tratamentos tenham efeitos semelhantes.

Conclusões

1) A magnitude de uma associação ou efeito de uma intervenção é dada pelo risco relativo, redução de risco relativo (eficácia), diferença de riscos ou diferença de médias e não pelo valor de p.

2) O fato de uma intervenção num ensaio clínico apresentar maior eficácia (redução relativa do risco) não significa que ela é responsável pela maior redução de risco em termos absolutos.

3) O valor de p não indica se o efeito de uma intervenção é forte ou fraco. Ele apenas indica a probabilidade de se observar determinado um efeito quando este se deve ao acaso.

4) O valor de p é influenciado, entre outros fatores, pelo tamanho da amostra.

5) Estudos com amostras maiores tendem a obter estimativas de efeito mais precisas (menor intervalo de confiança) e costumam apresentar maior poder (probabilidade de detectar um efeito quando este existe).

Referências

1. Medical Research Council. Streptomycin treatment of pulmonary tuberculosis. A Medical Research Council Investigation. BMJ. 1948;2:769-82.
2. Pocock SJ. Clinical trials: a practical approach. Chichester: John Wiley & Sons; 1983.
3. Meinert CL. Clinical trials: design, conduct and analysis. New York: Oxford University Press; 1986. (Monographs in epidemiology and Biostatistics, 8).
4. Prien RF, Cole JO. High dose chlorpromazine therapy in chronic schizophrenia. Report of National Institute of Mental Health-psychopharmacology research branch collaborative study group. Arch Gen Psychiatry. 1968;18(4):482-95.
5. TREC Collaborative Group. Rapid tranquillisation for agitated patients in emergency psychiatric rooms: a randomised trial of midazolam versus haloperidol plus promethazine. BMJ. 2003;327(7417):708-13. Comment in: Evid Based Ment Health. 2004;7(2):42.
6. Borison RL, Diamond BI, Dren AT. Does sigma receptor antagonism predict clinical antipsychotic efficacy? Psychopharmacol Bull. 1991;27(2):103-6.
7. Clark ML, Ramsey HR, Rahhal DK, Serafetinides EA, Wood FD, Costiloe JP. Chlorpromazine in chronic schizophrenia. The effect of age and hospitalization on behavioral dose-response relationships. Arch Gen Psychiatry. 1972;27(4):479-83.
8. Hall RA, Dunlap DJ. A study of chlorpromazine: methodology and results with chronic semi-disturbed schizophrenics. J Nerv Ment Dis. 1955;122(4):301-14.
9. Ban TA, Lehmann HE, Sterlin C, Climan M. Comprehesiven clinical studies with thiothixene. Dis Nerv Syst. 1975;36(9):473-7.