3
Avaliação crítica e limitações dos ensaios clínicos

Bernardo Rangel Tura [1,2,3], Nelson Albuquerque de Souza e Silva [3,4] e Basílio de Bragança Pereira [5,6]

Universidade Federal do Rio de Janeiro

Palavras-chave:
avaliação critica, ensaios clínicos, medicina baseada em evidências

Introdução

Na última década, os cardiologistas têm sido "bombardeados" por um número crescente e avassalador de "ensaios clínicos randomizados" ("*clinical trials*") [1], identificados por siglas, por vezes criativas, outras vezes nem tanto. As análises dos resultados desses estudos procuram direcionar a prática clínica. Induz-se a idéia de que, se o clínico não se comportar, nos cuidados com os seus pacientes, de acordo com o que foi "demonstrado" em tal ou qual ensaio clínico, ou nas análises dos conjuntos destes, as "meta-análises", está defasado no conhecimento e praticando "medicina sem evidências científicas". O clínico então busca orientar a sua prática, utilizando os princípios da "medicina baseada em evidências". Surge um número crescente de "diretrizes" ("guidelines"). Estas, em geral, consideram os resultados dos ensaios clínicos dentro de um modelo hierarquizado, atribuindo graus de evidências de acordo com a metodologia empregada nos diferentes estudos. Justamente por serem tão valorizados, tornou-se de suma importância que os médicos, ao lerem um ensaio clínico, sejam capazes de avaliar a sua qualidade e compreender suas limitações. Os idealizadores da medicina baseada em evidências, apropriadamente, deram grande ênfase ao "Julgamento Clínico", como a base para interpretar qualquer resultado científico e aplicá-lo à prática clínica. Apesar disto, procura-se desvalorizar o julgamento clínico como estando "fora de moda", "não-científico", e que todo e qualquer resultado que tenha por base uma metodologia de pesquisa clínica, hoje aceita como adequada, deva ser posta em prática. A ciência evolui, os métodos mudam ou se aperfeiçoam e só o clínico experiente, conhecedor profundo da metodologia de pesquisa clínica, do ambiente de sua prática e do paciente e seus familiares, poderá julgar, dentro do conhecimento científico atual, o que é mais adequado para aquele paciente sob seus cuidados.

O primeiro passo deste processo é definir e classificar os ensaios clínicos. Chama-se ensaio clínico um estudo que avalia o impacto de uma intervenção numa determinada condição clínica. Estes estudos podem ser classificados de várias formas.

O ensaio é chamado controlado [2] ("*controled clinical trials*") se o grupo da intervenção estudada foi comparado com um outro grupo dito controle. Se o grupo controle for um grupo tratado com placebo, este estudo será chamado ensaio clínico controlado por placebo ("*placebo-controlled trial*") [3]. Se o grupo de controle utilizar a terapia-padrão, o estudo será chamado de ensaio clínico ativamente controlado ("*active controlled trial*") [4,5].

O ensaio clínico pode ou não ser randomizado ("*randomized clinical trial*"), sendo classificado como aberto ou cego ("*open-label*" ou "*blinded*") [2,6]. A randomização implica em sortear, dentro de uma determinada metodologia, a intervenção que o paciente receberá, por isso, em certos casos, principalmente por questões éticas, esta não pode ou não deve ser realizada [7]. O estudo é chamado simples cego, duplo cego ou triplo cego, se somente o paciente, o médico e o paciente ou ainda o paciente, o médico e o responsável pela análise de dados, respectivamente, desconhecem qual paciente fez parte do grupo controle ou do grupo de estudo.

Avaliação crítica de um ensaio clínico

O pensamento científico hegemônico desde o século XVII tem se desenvolvido por um processo dedutivo-indutivo [8] e nem sempre acumulativo. Isto significa que evolui através de experimentos ou estudos, que são baseados na observação dos fenômenos naturais ou não, e considera que o conhecimento é construído a

1. Coordenação de Ensino e Pesquisa do INCL/Ministério da Saúde.
2. Médico-pesquisador da DIM/FCM-UERJ.
3. Programa de Pós-graduação em Clínica Médica, área de concentração Pesquisa Clínica, do Departamento de Clínica Médica da Faculdade de Medicina da UFRJ.
4. Professor Titular de Cardiologia do Departamento de Clínica Médica da Faculdade de Medicina/UFRJ.
5. Professor Titular do Departamento de Medicina Preventiva da Faculdade de Medicina/UFRJ.
6. NESC- Núcleo de Estudos em Saúde Coletiva

partir dos resultados e sua interpretação, não de um, mas de vários experimentos relacionados a um determinado objeto de pesquisa. O pensamento indutivo pode ser assim expresso: desde que certas condições sejam satisfeitas, é legítimo generalizar a partir de uma lista finita de proposições de observação singulares para uma lei universal. As condições que devem ser satisfeitas para uma generalização ser considerada legítima, pelo pensamento indutivo são:

1. o número de proposições de observação que forma a base de uma generalização deve ser grande;
2. as observações devem ser repetidas sob ampla variedade de condições;
3. nenhuma proposição de observação deve conflitar com a lei universal derivada.

Baseado neste pensamento é fácil concluir que um único ensaio clínico não irá resolver totalmente uma questão e, ainda, que é essencial ter bons ensaios e avaliar criticamente os seus resultados como alicerce da prática médica.

Mas como é possível saber se um determinado ensaio clínico é ou não de boa qualidade? Esta não é uma pergunta nova ou ainda não respondida. De fato, vários autores já se dedicaram a criar critérios para avaliar a qualidade de artigos publicados, inclusive de ensaios clínicos. Cabe um especial destaque a série de artigos "User's *guide to the medical literature*" que é publicada no JAMA desde 1993 [9] e que vem a ser um dos materiais mais citados sobre o assunto.

O primeiro tipo de estudo que foi enfocado por esta série foram os estudos sobre prevenção e tratamento (no caso ensaios clínicos). Trata-se de dois artigos: o primeiro enfoca a avaliação da validade do estudo[10], enquanto o segundo se os resultados do estudo podem ser usados nos pacientes do médico que lê o artigo [11]. A questão levantada pelo segundo artigo será vista adiante.

Neste artigo, o autor utiliza um interessante esquema hierárquico de oito perguntas a serem respondidas em três níveis: dois sobre metodologia e um sobre resultados. A idéia é que estas perguntas encaminhariam um esquema semelhante a um conjunto de triagens. A primeira triagem seria composta de três perguntas essenciais, aquelas que deveriam ser verificadas mesmo numa leitura rápida. A segunda triagem, também com três perguntas, deveria ser utilizada numa leitura mais cuidadosa, caso o artigo tivesse passado pela primeira. Por fim, mais duas perguntas deveriam ser feitas sobre os resultados, se o artigo tivesse sido aprovado nas triagens anteriores (Tabela 1).

Nos últimos dez anos, os ensaios clínicos aumentaram em número e em complexidade, o que justifica, na nossa avaliação, a inclusão de mais dois conjuntos de perguntas: um referente à metodologia e outro referente aos resultados. A função dessas novas triagens seria garantir uma maior qualidade metodológica do estudo e facilitar o entendimento da evidência obtida. É importante observar que a estrutura hierárquica seja mantida. Neste sentido só deveríamos utilizar as novas triagens se o estudo já tivesse sido aprovado nas anteriores.

Cada uma destas perguntas tem um motivo para ser feita, portanto é interessante uma breve explicação sobre cada uma delas para permitir a melhor compreensão de sua importância. Uma discussão mais detalhada fugiria do objetivo deste artigo.

A alocação dos pacientes foi randomizada?

A randomização é importante nos ensaios clínicos por três motivos: reduz a possibilidade do viés de seleção de pacientes; aumenta a possibilidade dos grupos de pacientes serem comparáveis - especialmente se o tamanho da amostra for suficientemente grande - e valida o uso de testes estatísticos habitualmente utilizados [12].

De fato a randomização torna os grupos estudados comparáveis em relação aos determinantes, conhecidos ou não, do desfecho estudado. Além disso a ausência da randomização tende a superestimar o efeito da intervenção estudada [13]. Ressalte-se que não basta haver menção de que o estudo foi randomizado, é necessário descrever o processo de randomização, e este processo deve também ser "cego".

Houve perda de acompanhamento ("*follow-up*")?

Todos os pacientes que entram no ensaio clínico devem ser acompanhados até o final do mesmo. Quando isto não acontece, a conclusão do estudo pode ficar comprometida, pois a perda de informação pode alterar o efeito da intervenção estudada. Isto depende do tamanho da perda de acompanhamento ("*lost*

to follow-up"). Quanto maior a perda, maior a chance da alteração ocorrer. Pode-se considerar o valor de 20% dos pacientes seguidos, como um limite de referência para esta situação.

No estudo HPS [14] a perda de acompanhamento foi de 0,33%, enquanto noutro estudo, o CADILAC [15], foi de 27,1%. É claro que no primeiro esta perda de pacientes não é preocupante, porém no segundo é possível se acreditar que esta pode ter alterado o resultado do estudo. Mais adiante será explicado como é possível avaliar o impacto da perda de acompanhamento.

A análise foi realizada seguindo a intenção de tratar ("*intention-to-treat*")?

Durante um estudo, os pacientes podem não continuar no grupo para o qual foram inicialmente alocados. Isto pode ocorrer por vários motivos: não-aderência, intercorrência clínica, efeito adverso da intervenção etc.. Se, ao analisar o estudo, for considerado que o paciente permanece no grupo alocado, mesmo que tenha migrado para outro, dizemos que a análise foi realizada seguindo a intenção de tratar. O motivo pelo qual se prefere este tipo de análise, é que esta garante a manutenção dos benefícios obtidos com a randomização, caso contrário, o estudo deixa de ser randomizado e portanto, seus resultados comprometidos.

A randomização foi cega?

O próprio processo de randomização tem que ser "cego". A randomização utilizando envelopes fechados, por exemplo, pode possibilitar a escolha, indevida, do caso depois de aberto o envelope, o que a compromete. A vantagem de ter um ensaio cego ou mascarado, desde a randomização, é evitar o chamado viés de expectativa,

Tabela 1

Esquema de avaliação crítica da qualidade de um ensaio clínico

1. Primeira Triagem (o essencial)
- 1.1. A alocação dos pacientes foi randomizada?
- 1.2. Houve perda de acompanhamento (*follow-up*)?
- 1.3. A análise foi realizada seguindo a intenção de tratar (*intention-to-treat*)?

2. Segunda Triagem (o indicado)
- 2.1. A randomização foi cega?
- 2.2. A randomização foi eficiente?
- 2.3. Só a intervenção estudada diferenciou o tratamento dos pacientes?

3. Terceira Triagem (evidência e metodologia)
- 3.1. O objetivo do estudo foi claramente formulado?
- 3.2. Se o estudo tem mais de um objetivo, eles podem ser analisados no mesmo estudo?
- 3.3. Os critérios de inclusão e exclusão são consistentes com o objetivo do estudo?
- 3.4. Os critérios de inclusão e exclusão favorecem alguma intervenção?
- 3.5. A intervenção estudada e a co-intervenção foram padronizadas?
- 3.6. As padronizações da intervenção estudadas e da co-intervenção foram adequadas em relação ao objetivo do estudo?
- 3.7. O acompanhamento foi por tempo suficiente?
- 3.8. Houve troca de grupo (*crossover*)?

4. Resultados – Primeira Triagem (o essencial)
- 4.1. Foi avaliado o efeito do tratamento?
- 4.2. Foi avaliada a precisão deste efeito?

5. Resultados – Segunda Triagem (evidência e metodologia)
- 5.1. O desfecho estudado é composto por um único evento?
- 5.2. Se for composto por mais de um evento, todos têm a mesma importância clínica?
- 5.3. Se for composto por mais de um evento, o efeito da intervenção é estatisticamente significativo para todos os eventos?
- 5.4. Se houve perda de acompanhamento, ela foi igualmente distribuída?
- 5.5. Se houve perda de acompanhamento, foi calculado o impacto da perda?
- 5.6. Se houve troca de grupos, foi calculado o impacto da troca?

ou seja, uma tendência a interpretar os resultados da intervenção de forma melhor ou pior de acordo com a sua visão sobre a mesma. Além disso, os profissionais de saúde podem ter diferentes interpretações de pequenas modificações ou de abordagem do paciente em estudos abertos [16].

A randomização foi eficiente?

Os ensaios clínicos avaliam a capacidade de uma intervenção influir em um determinado desfecho. Por isso é importante que os grupos do estudo tenham o mesmo risco para o desfecho analisado. Se a randomização for eficiente, as características clínicas dos grupos serão semelhantes e os grupos terão um risco semelhante. Por exemplo, no estudo PROGRESS [17], o grupo de pacientes que recebia diurético e betabloqueador ou recebia dois placebos era mais novo, com mais homens, mais doença isquêmica, com a pressão arterial mais alta e tinha mais hipertensos que o grupo que recebeu betabloqueador ou apenas um placebo.

Só a intervenção estudada diferenciou o tratamento dos pacientes?

Justamente para termos certeza de que o efeito sobre o desfecho foi determinado única e exclusivamente pela intervenção estudada, é necessário que os pacientes de ambos os grupos recebam o mesmo tratamento, ou seja, tenham o mesmo tipo de acompanhamento, o mesmo número de consultas, o mesmo acompanhamento ambulatorial, a mesma co-intervenção etc.

O objetivo do estudo foi claramente formulado?

Esta pergunta é essencial para permitir o entendimento da evidência que o artigo pode nos fornecer. Neste caso, o importante é que seja dito não só o que se pretende estudar, como também qual a definição de cada manifestação clínica estudada.

Se o estudo tem mais de um objetivo, eles podem ser analisados no mesmo estudo?

Para cada tipo de estudo e, às vezes, até para cada tipo de objetivo são necessários tipos de amostras diferentes, por isso, nem sempre, a amostra que serve para um determinado estudo ou objetivo serve para outro. Por exemplo, nos estudos de prognóstico, é essencial que os pacientes que compõem a coorte inicial do estudo, estejam no mesmo estado evolutivo da doença. Caso contrário, poderemos inferir uma associação ou causalidade entre um determinado fator e um determinado evento, quando na realidade o estágio inicial de evolução do paciente é que determinou o melhor ou o pior prognóstico e não o fator em estudo. Para os ensaios clínicos, isto também pode ser verdade. Se selecionarmos pacientes com amplo espectro clínico da doença, pode ocorrer que a intervenção em estudo seja benéfica apenas para pacientes com melhor ou com pior prognóstico e isto não será observado no grupo todo. Portanto, estudar o efeito das intervenções em formas clínicas semelhantes da doença e, portanto, prognóstico semelhante, pode ser mais informativo para o clínico do que tentar estudar efeitos generalizáveis para toda e qualquer situação, embora isto por vezes possa ser o desejável. Um determinado tratamento pode ter resultados diversos em pacientes com prognósticos diversos. Já nos estudos que avaliam testes diagnósticos é essencial que haja o maior espectro possível de apresentação da doença [18]. Portanto, não é possível fazer uma avaliação de um teste diagnóstico em uma amostra de um ensaio clínico ou de um estudo de prognóstico.

Os critérios de inclusão e exclusão são consistentes com o objetivo do estudo?

Aqui é importante determinar se os critérios de inclusão e exclusão são coerentes com o objetivo do estudo, ou seja, se realmente irão selecionar pacientes com as características clínicas necessárias para a realização do ensaio clínico e retirar aqueles que podem falsear o resultado ou que, por razões clínicas ou éticas, não devem receber a intervenção estudada. A população estudada deve ser representativa de pacientes que encontramos em nossa prática diária. Os critérios diagnósticos para caracterizar a doença selecionada, ou para caracterizar os desfechos clinicamente relevantes escolhidos, devem ser muito bem definidos para que não ocorram erros de classificação das doenças.

Os critérios de inclusão e exclusão favorecem alguma intervenção?

Neste caso é importante observar se, ao selecionar os pacientes, um grupo foi beneficiado ou prejudicado, ou seja, se para uma determinada intervenção tornou-se mais ou menos eficaz, porque naqueles pacientes ela apresenta melhores ou piores resultados. Por exemplo, no estudo LIFE [19] os fatores de exclusão retiraram os

pacientes com história de infarto do miocárdio e angina. Estes pacientes teriam clara indicação de usar o betabloqueador,e ao excluí-los, o estudo tende a favorecer a outra intervenção (lisinopril).

A intervenção estudada e a co-intervenção foram padronizadas?

Como já foi dito, é essencial que os pacientes tenham recebido a mesma intervenção, portanto ela tem que estar padronizada. Por exemplo, no caso de uma droga, a sua dosagem, a via de administração, o intervalo entre as doses e o tempo de tratamento têm que ser descritos. Se houve a necessidade de ajustar a dose do remédio, a forma como isto foi feito e como se avaliou o efeito desta mudança, também devem ser registrados.

As co-intervenções, quaisquer outras intervenções realizadas no paciente, em paralelo com a intervenção estudada, devem ser padronizadas, principalmente se tiverem um impacto importante no desfecho. Por exemplo, no estudo PROGRESS [20], os pacientes que eram hipertensos podiam participar do estudo, desde que se encontrassem estáveis há pelo menos duas semanas e não usassem inibidores da ECA. Entretanto, o estudo não faz nenhuma referência ao uso de diuréticos, drogas reconhecidas por prevenirem Acidente Vascular Encefálico (AVE) na hipertensão [21,22].

As padronizações da intervenção estudadas e da co-intervenção foram adequadas em relação ao objetivo do estudo?

A padronização utilizada no estudo foi adequada para se observar o efeito da intervenção ou ela pode ter fatores que alterem o efeito da mesma no desfecho estudado? Por exemplo, no estudo LIFE [19] os pacientes receberam, inicialmente, losartan ou atenolol; se após dois meses persistissem hipertensos, acrescentava-se hidroclorotiazida; se após mais dois meses ainda persistissem hipertensos, aumentava-se a dose da primeira droga (losartan ou atenolol). Teria sido melhor que os pacientes recebessem a dose máxima da droga do grupo, inicialmente, para somente após este passo, ser adicionado o diurético. Isto facilitaria a avaliação do efeito da intervenção isoladamente, sem a adição da segunda droga.

O acompanhamento foi por tempo suficiente?

Aqui é importante garantir que o tempo de acompanhamento foi planejado para um período suficientemente longo, de modo a garantir que ocorram, e sejam detectados, os desfechos clinicamente relevantes escolhidos para o estudo, bem como os efeitos adversos esperados da intervenção. Esta pergunta é crítica quando o desfecho estudado é de evolução longa.

Houve troca de grupo (“*crossover*”)?

Como foi dito anteriormente, os estudos devem ser analisados segundo a intenção de tratar. Portanto, se um paciente não receber o tratamento que originalmente deveria ter recebido, ele não deve ser trocado de grupo. Se muitos pacientes saírem do seu grupo original, a avaliação do efeito da intervenção sobre o desfecho pode ser prejudicada. A randomização, que garante a igualdade entre os grupos de estudo, será “quebrada” caso o paciente não seja mantido no grupo original.

No estudo GUSTO [23], não houve diferença significativa dos desfechos entre os grupos que usavam heparina subcutânea ou intravenosa, porém 36% dos pacientes que foram originalmente alocados no grupo da heparina subcutânea, a receberam por via intravenosa[24]. Como esta troca de grupo foi muito grande, é possível que isto tenha alterado o resultado do estudo.

Foi avaliado o efeito do tratamento?

Uma vez que o objetivo do ensaio clínico é saber o efeito da intervenção sobre um determinado desfecho, é importante que este seja quantificado. É desejável que este efeito não seja descrito apenas na forma de risco relativo ou de redução relativa do risco. Outras medidas como a redução absoluta do risco (RAR) ou o Número Necessário a ser Tratado (NNT) devem ser fornecidas para melhorar a compreensão do real efeito da intervenção[25].

Foi avaliada a precisão deste efeito?

Quando se realiza um ensaio clínico, o efeito da intervenção sobre o desfecho, é na verdade, uma estimativa

do efeito na população geral. O real efeito é um número desconhecido que deve ser próximo do valor obtido no estudo. O ***intervalo de confiança*** permite uma melhor estimativa do valor deste efeito na população geral e não na amostra utilizada no estudo e, por isso, é importante que o estudo forneça o intervalo de confiança dos efeitos descritos no estudo [26].

O desfecho estudado é composto por um único evento?

É preferível, mas não essencial, que o estudo analise um desfecho descrito por um único evento. Desta forma é fácil avaliar o efeito da intervenção. Além disso, quando o evento é único não há outros problemas de interpretação que podem surgir e que serão abordados nas perguntas seguintes

Se o desfecho for composto por mais de um evento, todos têm a mesma importância clínica?

Normalmente, por questões operacionais, os estudos utilizam, para fins de análise, um desfecho composto por mais de um evento. Isto se torna muito importante quando os riscos dos eventos isolados são baixos, o que levaria à necessidade de um tamanho amostral muito grande. Nesse caso, é preciso avaliar se os eventos, utilizados para compor o desfecho têm a mesma relevância clínica ou se pelo menos não é muito diferente. Por exemplo, no estudo MIRACL [27], o desfecho principal é composto por quatro eventos: óbito, ressuscitação de parada cardiorrespiratória, infarto agudo do miocárdio não-fatal e isquemia miocárdica sintomática. É razoável admitir que a isquemia miocárdica é uma condição com menos importância clínica que as outras.

Se for composto por mais de um evento, o efeito da intervenção é estatisticamente significativo para todos os eventos?

Se o desfecho analisado for composto por mais de um evento, pode ocorrer que o efeito da droga esteja relacionado com alguns eventos e não esteja relacionado a outros eventos que compõem o desfecho. Então é importante observar, através do valor de p ou do intervalo de confiança, se o efeito é ou não estatisticamente significante em todos os eventos .

Por exemplo, no estudo CURE [28], o desfecho principal é composto de três eventos: morte por causa cardiovascular, infarto agudo do miocárdio não-fatal e acidente vascular encefálico (AVE). Este desfecho composto é significativamente menor nos pacientes que usaram o clopidogrel do que no tratamento comparativo (segundo o estudo: $p<0,001$). Entretanto, analisando o intervalo de confiança do risco relativo é possível observar que, se considerarmos isoladamente morte por causa cardiovascular (segundo o estudo de 0,79 a 1,08) e por AVE (segundo o estudo de 0,63 a 1,18), estes eventos não tiveram diferenças significativas relacionadas à intervenção.

Se houve perda de acompanhamento, ela foi igualmente distribuída?

Se as perdas de acompanhamento forem mais importantes num grupo que no outro, é possível que o efeito detectado não esteja relacionado à intervenção estudada. Pode ser que o grupo que teve maior perda de acompanhamento tenha um risco maior ou menor para o desfecho do estudo e a retirada em grande número destes indivíduos, altere o resultado avaliado. Este efeito é conhecido como viés de perda seletiva.

Se houve perda de acompanhamento, foi calculado o impacto da perda?

A perda de acompanhamento é uma ocorrência comum em estudos e pode comprometer seus resultados, porém é possível diminuir o impacto negativo desta no artigo avaliado. A forma como isto é feito é bem simples: basta atribuir aos pacientes que foram perdidos o pior desfecho do estudo (normalmente o óbito) e a seguir, refazer a análise e observar se o efeito da intervenção ainda persiste.

Por exemplo, no estudo PROWESS [29] houve a perda de acompanhamento de apenas um paciente, mesmo assim os autores consideraram que este faleceu e toda a análise foi realizada desta forma.

Se houve troca de grupos, foi calculado o impacto da troca?

A troca de grupo, da mesma forma que a perda de acompanhamento, pode ser um fator determinante no resultado de um estudo, podendo inclusive torná-lo inconclusivo. Novamente é possível calcular o impacto de uma troca excessiva e reverter o problema, bastando para isto fazer um procedimento simples como o citado

acima. Os pacientes que trocaram de grupo receberiam o pior desfecho do estudo e a análise seria refeita com esta adaptação, avaliando-se a diferença do resultado.

Limitações de um ensaio clínico

Nesta segunda parte do artigo, iremos enfocar as limitações dos ensaios clínicos. Isto significa que iremos abordar os motivos pelos quais, mesmo um estudo bem delineado e executado não pode e não deve ter seus resultados transpostos integral e instantaneamente a todo e qualquer paciente [11,30]. Esta também não é uma questão nova e há muito material escrito sobre o tema. O que iremos enfocar serão algumas questões consideradas importantes e que devem ser consideradas pelos médicos ao lerem um artigo.

Por exemplo, baseados no estudo SOLVD [31], podemos afirmar que o uso do inibidor da ECA na insuficiência cardíaca previne morte e internações? A resposta é não, por vários motivos, como veremos a seguir.

Generalização dos resultados

Os ensaios clínicos, mesmo que de excelente qualidade metodológica, têm uma limitação importante e constantemente ignorada, que é a baixa capacidade de generalização dos seus resultados. Isto ocorre porque uma amostra fornece uma ou mais estimativas válidas em relação à população que lhe deu origem. Como a população de um ensaio clínico é muito uniforme, o efeito passa a ser válido só em pacientes desta população.

Para avaliar se os resultados do estudo podem ser utilizados em determinado grupo de pacientes, devemos considerar os aspectos biológicos, epidemiológicos e socioeconômicos do mesmo [30].

O estudo SOLVD só avaliou pacientes com idade entre 21 e 80 anos, fração de ejeção ventricular esquerda menor ou igual a 35% e que não apresentassem doença coronária isquêmica instável; portanto seu resultado pode ser transposto a pacientes com características clínicas que sejam compatíveis com estas. É biologicamente aceitável que um paciente um pouco mais novo, por exemplo, com 19 anos, seja beneficiado pelo tratamento proposto, porém um paciente com uma fração de ejeção mais elevada pode não ter nenhum benefício.

Nos aspectos epidemiológicos é importante considerar vários componentes. Por exemplo, qual o perfil de risco dos pacientes? O risco do desfecho no seu paciente é semelhante ao do estudo? No estudo SOLVD, a mortalidade no grupo placebo foi de 39,7% e apresentou uma redução relativa de risco de 16%. Se os pacientes em que você deseja utilizar o tratamento apresentam um risco de óbito de 15%, o grupo tratado teria um risco de 12,6%, uma diferença não-significativa (p=0,075).

Por outro lado, estes pacientes podem apresentar co-morbidades que aumentam o risco de desfecho que você pretende reduzir ou que impeçam o uso da intervenção. Além disso, devemos considerar a população de origem dos pacientes. No exemplo que estamos utilizando, cada vez surgem mais estudos relacionando mudança na distribuição alélica do gene da ECA com a população estudada [32,33] e sobre a alteração da resposta do paciente com insuficiência cardíaca ao inibidor da ECA relacionada com o genótipo [34,35].

Por fim, mas não menos importante, devemos avaliar os aspectos socioeconômicos de nossos pacientes. Eles podem arcar com o ônus do tratamento? Se o Estado lhes fornecer o remédio, eles têm condições de tomá-lo corretamente? É garantido que ele receba seu tratamento sem interrupções?

Se o grupo de pacientes no qual você pretende utilizar a droga tiver características biológicas, epidemiológicas e socioeconômicas semelhantes às apresentadas no estudo, ainda assim a redução de risco obtida na sua população pode não ser a mesma que a do estudo. Esta foi uma estimativa pontual baseada numa amostra. A provável resposta do seu paciente deve estar dentro do intervalo de confiança indicado no estudo, e é este um dos motivos pelos quais é importante que este seja apresentado no artigo.

Efeito da classe

Por definição, uma classe de drogas é um conjunto de medicamentos que compartilham características comuns (a utilização, o efeito ou o mecanismo de ação). Por exemplo, os anti-hipertensivos são drogas utilizadas para tratar a hipertensão, os diuréticos aumentam a diurese e os inibidores da ECA inibem a enzima de conversão da angiotensina. Quando um determinado efeito é comum a todos os medicamentos de uma mesma classe, este é chamado de efeito da classe.

Apesar de existir, o efeito da classe não pode ser provado com um simples ensaio clínico, ou seja, o fato do enalapril ter tido êxito no estudo SOLVD não garante que todos os inibidores da ECA são benéficos na insuficiência cardíaca. Um exemplo disto é que existem vários estudos que mostram como os betabloqueadores são benéficos na ICC [36-39], posterior ou não ao infarto, porém há pelo menos um estudo que mostra que o sotalol pode aumentar a mortalidade dos pacientes com infarto agudo do miocárdio e insuficiência cardíaca [40].

É importante ter em mente que o efeito da classe é uma característica que só deve ser atribuída a uma classe de medicamentos, tendo em vista não apenas alguns ensaios clínicos mas vários estudos incluindo meta-análises.

Efeito do tratamento

Apesar de muitos dizerem o contrário, os efeitos terapêuticos das intervenções em cardiologia são modestos [41,42]. É raro encontrarmos um ensaio clínico que apresente uma redução relativa de risco superior a 25%. Há que se destacar que o placebo tem um efeito terapêutico que corresponde a uma redução relativa de risco entre 5 a 9% [43].

Um conceito muito utilizado para avaliar o impacto de uma terapia em relação a um desfecho é o NNT ("*number needed to treat"* ou número necessário para tratar) que expressa o número de pacientes que necessitamos tratar para prevenir um desfecho.

Consideremos que um ensaio clínico, com mais de 10.000 pacientes, mostrou que no grupo tratado observou-se uma mortalidade de 7% após cinco anos de tratamento e que o grupo placebo apresentou uma ocorrência de morte de 10%. Temos, portanto, uma redução do Risco Absoluto de 3% ou uma redução relativa de risco de 30% ou ainda um NNT de 33. Estes resultados seriam considerados "aceitáveis", pois são "estatisticamente significantes". No entanto, se olharmos estes mesmos resultados, de outro modo, vemos que não são brilhantes. Olhemos pelo lado dos que não se beneficiam com o tratamento. Se precisarmos tratar 33 pacientes durante cinco anos para reduzir um evento (NNT=33), 32 pacientes não tiveram benefício do tratamento. Portanto, ao transportamos estes resultados para os nossos pacientes, fica claro que eles precisam ser informados de que embora o tratamento seja benéfico, e comprovado pelos ensaios clínicos, a maioria deles não terá benefícios do tratamento.

Os grandes ensaios clínicos, com grande número de pacientes, são montados para mostrar pequenas diferenças que se tornam "estatisticamente relevantes" e servem para "vender" a droga ou o procedimento. No entanto, quando estamos lidando com pacientes de baixo risco, estas pequenas diferenças podem não ter nenhum significado clínico, pois a grande maioria dos pacientes não terá benefícios e pode ter malefícios importantes, além do custo, por vezes insuportável, destas novas drogas ou procedimentos terapêuticos.

Por exemplo, a partir do estudo HPS [14] é possível calcular que são necessários tratar com sinvastatina 83 pacientes por 5 anos para prevenir 1 óbito por doença coronária. Mas, e os outros 82 pacientes, o que ocorreu com eles? A resposta a este questionamento pode ser encontrada no NRR [44] (*number remaining at risk* ou número que permanecem em risco) que expressa o número de desfechos que ocorrerá na população que não foi beneficiada pela terapia. Por exemplo, no mesmo estudo a cada 83 pacientes tratados, iremos evitar 1 óbito, em relação ao outro tratamento, porém ocorreram 5 nos outros 82 pacientes do grupo submetido ao novo tratamento. Em um outro exemplo, baseado no estudo MADIT II [45], podemos concluir que é possível prevenir um óbito para cada 18 pacientes infartados com fração de ejeção menor que 30%, se utilizarmos um desfibriliador implantável profilático; porém ocorreram três óbitos nos outros pacientes.

É importante também considerar que os efeitos terapêuticos de duas, três ou mais drogas não podem ser adicionados aritmeticamente Por exemplo, considerando que a aspirina [46] e os betabloqueadores [47] apresentam uma redução relativa de risco de 25% na prevenção secundária de óbito, infarto ou AVE, é errado dizer que as duas drogas juntas reduziriam o risco em 50%.

Uma forma de tentar estimar este efeito conjunto seria calcular a redução de risco da seguinte forma: 1. [(1–efeito da aspirina) x (1–efeito do betabloqueador)] = 1–(0,75 x 0,75) = 0,4375 ou 43,75%. Mesmo este cálculo pode não representar o efeito real das drogas combinadas. Podemos imaginar também que duas drogas tenham efeito benéfico quando utilizadas isoladamente; quando combinadas podem ter efeitos adversos ou ainda uma reduzir o efeito da outra. Conforme aumentamos o número de drogas associadas, mais imprevisível

torna-se a avaliação dos seus efeitos conjuntos, pois as interações farmacológicas aumentam exponencialmente e não representam apenas aquelas ações que desejamos como benéficas.

A prevenção primária das doenças é uma abordagem, em geral, muito mais eficaz que as intervenções medicamentosas. Sabe-se que existe uma relação entre o peso ao nascer e o risco de infarto agudo do miocárdio. Uma criança que tem menos de 2.268 g ao nascer tem um risco relativo de 1,49 em relação àquelas que nascem com o peso normal [48]. Isto pode estar relacionado a um polimorfismo do gene do fator 1 de crescimento semelhante à insulina[49] (IGF-1). Uma intervenção que evitasse o nascimento de crianças de baixo peso, levaria a uma redução relativa de risco de 32,9% e também diminuiria o risco do surgimento de diabetes melittus tipo II. Do mesmo jeito, a interrupção do fumo após um infarto reduz o risco de um novo infarto, óbito ou AVE em 50% [42]. Estas duas intervenções não-medicamentosas são mais eficazes que qualquer droga utilizada na prevenção secundária pós-infarto.

Tamanho do estudo

É muito comum referências ao tamanho do estudo como um indicador de qualidade, mas isto não é em si uma verdade. O tamanho da amostra do estudo é determinado por alguns fatores: nível de significância do estudo; poder do estudo; risco do desfecho esperado no grupo controle; redução relativa do risco ou risco do desfecho esperado no grupo tratamento.

Por exemplo, o estudo INSIGHT [50] relata um nível de significância de 95%, um poder de 90%, uma taxa de risco do evento em três anos de 8% e uma redução de 25% deste risco com a intervenção proposta. Isto corresponderia a um tamanho amostral de 6830 pacientes.

A fórmula para o cálculo do tamanho da amostra de um ensaio clínico envolvendo 2 grupos [51] (controle e tratamento) é:

$$N = \frac{2[p_c(1-p_c)+p_i(1-p_i)]\ (Z_{1-a}/2+Z_{1-b})^2}{(p_c - p_i)^2}$$

O nível de significância do estudo [52] (Z_{1-a}) representa a probabilidade de se considerar que existe uma diferença, se esta de fato existir. Há uma relação estreita entre o nível de significância e o valor de p de um estudo. Para um nível de significância de 95% , o valor de p limite é 0,05; se a significância do estudo for de 99% o mesmo seria 0,01. Quanto maior o nível de significância de um estudo, maior será o tamanho amostral do mesmo.

O poder do estudo (Z_{1-b}) representa a probabilidade de se considerar uma diferença como estatisticamente não-significativa, se ela de fato não é estatisticamente significativa. O poder do estudo não influencia diretamente o valor de p, mas nem por isso deixa de ser importante. Sua importância surge quando a diferença é não-significativa [53], pois quanto maior for o poder do estudo menor será a probabilidade disto estar errado. Quanto maior o poder de um estudo maior será o tamanho amostral do mesmo.

Se o tamanho da amostra for fixo, a significância e o poder do estudo têm uma relação inversa, ou seja, se aumentarmos um deles o outro se reduzirá. Só com o aumento do tamanho amostral é possível aumentar a ambos.

O risco do desfecho esperado no grupo controle (p_c) é um dos principais determinantes do tamanho amostral de um estudo. Quanto menor for este risco esperado, maior será o tamanho da amostra. Por exemplo, no estudo INSIGHT [50] foi calculado um tamanho amostral de 6830 pacientes para um risco esperado de 8% em três anos; entretanto, para um risco de 5% este mesmo estudo precisaria de 11.243 pacientes. A influência deste risco no tamanho amostral é um dos motivos pelos quais os ensaios clínicos tendem a estudar populações de alto risco para um determinado evento e, por este mesmo motivo, muitos estudos utilizam um desfecho composto por mais de um evento, visando aumentar o risco esperado e, conseqüentemente, diminuir o tamanho amostral.

O último elemento do cálculo pode ser considerado como o efeito do tratamento e está representado pelo risco do desfecho esperado no grupo da intervenção (p_i). Quanto maior for o efeito da intervenção, ou seja, a redução relativa de risco, menor será o risco esperado no grupo do tratamento. Em relação ao tamanho da amostra, quanto maior for o efeito do tratamento, menor será o número de pacientes no estudo.

Para melhor avaliarmos o impacto da mudança de cada um destes componentes no tamanho da amostra de um estudo, é possível fazer a seguinte comparação. Se aumentarmos o poder de um estudo de 80% para 90%, isto acarretaria um aumento de aproximadamente 33% no tamanho do estudo; se este aumento fosse para 95%, atingiria algo em torno de 65% e, no caso de 99%, seriam necessários 134% de pacientes a mais do que antes. Em relação à significância, um aumento de 95% para 99% levaria a um aumento de cerca de 49% no tamanho da amostra; um aumento menor que o correspondente, quando se alterou o poder do estudo.

No caso do risco de desfecho e no efeito do tratamento, este impacto é mais variável e, portanto, não nos permite generalizar como foi feito em relação ao poder e significância de um estudo. Porém observe o seguinte exemplo: um estudo com significância de 95%, poder de 80%, risco no controle de 30% e com uma redução relativa de risco de 20%, precisaria de 1712 pacientes. Se o risco esperado fosse de 20%, seriam necessários 2889 pacientes (68,75% de aumento). Se por outro lado, quiséssemos detectar como significante uma diferença de 10%, seriam necessários 7101 pacientes (um aumento de mais de 300%).

Um estudo de boa qualidade informa quais foram os valores atribuídos a cada um dos componentes do cálculo do tamanho amostral. Além disso este cálculo deve ser realizado antes do estudo ocorrer. A prática de calcular o poder após o término da coleta dos dados é questionável [54].

É importante frisar que, uma vez que haja um número suficiente de pacientes, é possível que qualquer diferença, por menor que ocorra, seja considerada estatisticamente significativa.

Significância estatística e significância clínica

Para melhor compreender a evidência produzida por um estudo, é de suma importância entendermos o conceito de significância estatística. A generalização indevida dos resultados e o mal entendimento do conceito da significância estatística são os principais responsáveis por uma série de conclusões conflitantes e/ou errôneas que levaram alguns autores a sugerir o fechamento de todos os departamentos de epidemiologia do mundo [55].

Para melhor entendermos a questão da significância estatística é preciso entender a idéia do teste de hipótese. Este é um procedimento que nos permite escolher entre duas hipóteses: a nula e a alternativa. A hipótese nula é assim chamada por significar que não há, de fato, uma diferença entre os tratamentos, enquanto a outra, chamada alternativa, diz que há uma diferença entre os tratamentos. O teste de hipótese pode ter dois tipos de erros. O primeiro (tipo I) consiste em rejeitar a hipótese nula quando ela é verdadeira e o segundo (tipo II) seria aceitar a hipótese nula sendo ela falsa. É importante lembrar que existe uma relação entre o nível de significância e o poder de um ensaio clínico e a possibilidade dos erros tipo I e II do mesmo.

O tamanho amostral de um estudo é calculado para permitir um teste de hipótese, normalmente para verificar se um determinado desfecho ocorre menos no grupo tratado que no grupo controle. Por exemplo, o estudo ALLHAT [22] foi realizado para avaliar se a clortalidona poderia reduzir em 16% na ocorrência do desfecho (doença coronária isquêmica fatal ou infarto do miocárdio não-fatal) quando comparada com amilodipina ou lisinopril. Este estudo tinha um poder de 83% e significância de 98,2%, o que resultou num estudo com 33.357 pacientes.

No resumo deste mesmo artigo, foi descrita, entre outras, uma redução estatisticamente significativa da pressão arterial diastólica. Novamente para fins de exemplo, observemos a diferença entre a amilodipina e a clortalidona: o desenvolvimento de insuficiência cardíaca em 6 anos foi de 10,2% e de 7,7%, respectivamente. Da mesma forma a pressão arterial diastólica foi, em média, 0,8 mmHg menor no grupo do diurético.

Apesar de serem estatisticamente significativas, existe uma diferença importante entre estes efeitos. A redução da ocorrência de insuficiência cardíaca foi clinicamente importante, a da pressão arterial diastólica não. A insuficiência cardíaca é uma doença grave que leva à incapacidade e à morte e que normalmente não apresenta possibilidade de reversão. No estudo em questão, a clortalidona apresentou uma redução de risco relativo de 24,5% em relação à amilodipina. Esta redução é realmente digna de nota porque é feita em relação a outro tratamento eficaz e não a um placebo. Se utilizarmos o conceito do NNT [56], esta intervenção irá prevenir uma insuficiência cardíaca a cada 40 pessoas tratadas o que, considerando o custo, é uma boa opção.

Por outro lado, uma redução da pressão arterial diastólica de 0,2 mmHg não tem tanta importância, principalmente se considerarmos que os pacientes estudados são hipertensos do estágio I ou II. Esta diferença é tão pequena que não é sequer aferível na maior parte dos esfignomanômetros habitualmente utilizados, ou

seja, mesmo sendo uma diferença significativa no aspecto estatístico, isto não significa que tenha importância clínica. Embora tenhamos feito o raciocínio acima, devemos chamar atenção para o fato de que a diferença encontrada foi entre as médias dos grupos e que diferenças de médias não têm o mesmo significado do que uma variação do mesmo nível na pressão arterial de um paciente.

Esta dissociação entre os dois tipos de significância (clínica e estatística) ocorre por um motivo já enfocado previamente neste artigo: o tamanho da amostra. O tamanho da amostra de um estudo é calculado a partir de uma diferença que é considerada clinicamente importante para um determinado desfecho. Neste caso, um teste será ou não clínica e estatisticamente significativo. Por outro lado, ao avaliar qualquer outra diferença dentro deste estudo, o leitor deve usar de seu discernimento, pois como a amostra não foi calculada para aquele teste, este pode apresentar falsos resultados. Portanto, temos duas possibilidades: a primeira é que o estudo tenha menos pacientes do que o necessário para se realizar o teste. Terá, então, um poder menor que o descrito no texto e poderá considerar uma diferença clinicamente importante como não sendo estatisticamente significativa. A outra possibilidade é que o estudo tenha mais pacientes que o necessário; agora, uma diferença clinicamente não-importante poderia ser estatisticamente significativa.

Portanto, sempre que detectarmos num artigo uma dissociação entre as significâncias, devemos buscar outros estudos para formarmos nossa opinião.

A interpretação do valor de p

Quando se faz a leitura de um artigo devemos ter especial atenção para interpretarmos corretamente o valor de p e, por conseqüência, termos melhor entendimento da evidência e da conclusão.

O valor de p surgiu há cerca de 60 anos[57]. Nesta época havia duas formas de se avaliar se um tratamento era superior ao outro: o teste de significância de Fisher o e o teste de hipóteses de Neyman e Pearson.

Na visão de Fisher, o valor de p deveria ser interpretado como a capacidade da hipótese nula explicar os fatos observados naquele estudo. A visão de Fisher é baseada em minimizar a ocorrência do erro do tipo I. Foi ele que propôs o famoso limite de $p<0,05$ para a significância do seu teste, porém ele mesmo recomenda, enfaticamente, que haja uma posição de interpretação em relação a este valor. Fisher chega a afirmar que um valor próximo do limite proposto não poderia nem confirmar nem afastar a hipótese nula, e por isso outro experimento deveria ser realizado[58].

Posteriormente, uma nova abordagem suplantou a proposta de Fisher. Foi o chamado teste de hipótese. Nesta nova abordagem, além do erro tipo I, o pesquisador se preocupa em evitar o erro tipo II. É interessante o fato de que o limite do valor de p, posposto por Fisher para seu teste de significância, foi eternizado para o teste de hipótese [58,59].

Tradicionalmente o valor de p tem sido interpretado de forma equivocada. É errado interpretá-lo como a probabilidade da conclusão do teste estar errada [58]. De fato o valor de p corresponde à probabilidade de encontrarmos um resultado igual ou mais extremo que o do estudo e a hipótese nula ser verdade [59]. Apesar de parecerem iguais, estas duas abordagens são diferentes. A conclusão de um teste de hipótese pode ser errada mesmo que o valor de p diga o contrário. O valor de p foi proposto por Fisher como uma medida da discrepância entre a hipótese nula e os dados do teste e assim deve ser interpretado. O que significa que, quanto menor for o valor de p, mais distante os dados do estudo estão da possibilidade representada pela hipótese nula e, portanto, devemos escolher a hipótese alternativa. Por outro lado, num estudo com 95% de significância e poder de 80%, uma conclusão estatisticamente significativa (aquela com $p < 0,05$) tem, segundo alguns autores, 36% de chance de estar errada [58,60].

Pelo mesmo motivo, é errado interpretá-lo como uma medida de eficiência. Ou seja, se num estudo o valor de p de um determinado tratamento é menor que o valor de p de outro tratamento, não significa que o primeiro seja melhor que o segundo, e sim que, naquele estudo, os dados mostram maior discrepância entre não existir efeito para o primeiro tratamento que para o segundo.

Outro engano comum é usar o valor de p como uma medida de evidência a favor do tratamento ser diferente do placebo. O valor de p serve para orientar a escolha entre duas hipóteses; a evidência é quanto uma hipótese é mais verossímil que a outra a partir dos dados do estudo. Neste aspecto o que deve ser utilizado para medir a evidência é a razão de verossimilhança ("*likelihood ratio*") e não o valor de p [61].

Conclusão

Os ensaios clínicos são uma ferramenta útil para auxiliar o cardiologista na sua prática diária, mas não podem ser usados de forma simples e massificados. Devemos buscá-los como fonte de evidência para orientar nossas decisões, mas não podemos fazê-lo como se este fosse um dogma inquestionável.

Alguns pontos devem ser frisados para maximizar o que aqui foi colocado. Primeiro: os médicos não devem transpor um ensaio clínico para a prática médica sem antes avaliá-lo criticamente. Isto é essencial, pois evita que estudos de qualidade ruim afetem o tratamento dos pacientes. Segundo: mesmo que o artigo tenha qualidade, não significa que iremos ter a mesma resposta que foi descrita no estudo; mas significa que não devemos operacionalizar grandes mudanças baseadas em um único artigo.

O conhecimento em medicina pode ser desenvolvido de forma cumulativa, mas também pode ocorrer quebrando os conceitos anteriores, portanto um único ensaio clínico, por melhor que seja, dificilmente resolve completamente a questão do tratamento de determinada doença.

O simples fato de um estudo ter muitos pacientes, não significa que seja bom, bem como o fato de uma diferença ser estatisticamente significativa não quer dizer que isto afetará o paciente. São necessários discernimento e bom senso por parte do cardiologista para avaliar o real significado da informação.

Por último, mas não menos importante, por mais atrativo que seja, devemos evitar sucumbir à tentação de simplificar o tratamento dos pacientes com um único e milagroso medicamento. Os efeitos dos remédios em cardiologia, ou na medicina em geral, são modestos e o “efeito da classe” nem sempre está presente.

Podemos dizer ainda que, mesmo para as drogas que tiveram ensaios clínicos evidenciando respostas significantes, os tratamentos não beneficiam a maioria dos pacientes. Mais ainda, em geral, os maiores efeitos são para os pacientes com risco mais alto de apresentar um desfecho relevante. Para os pacientes de baixo risco, em geral, os efeitos das drogas são inexistentes ou muito pequenos.

Devemos ter sempre presente em nosso pensamento e em nossa conduta que precisamos abordar o paciente com um todo, adotar políticas de prevenção (por ser o tratamento mais efetivo) para evitar a sua doença e, só se esta falhar, usarmos os medicamentos com a melhor relação custo-efetividade possível.

Concluindo, para utilizarmos os resultados dos ensaios clínicos em nossos pacientes são fundamentais: o “julgamento clínico” consciencioso e baseado nas evidências científicas existentes; a avaliação das expectativas dos pacientes e de seus familiares; o conhecimento do ambiente da prática clínica e suas limitações; e ainda a compreensão do processo de tomada de decisões com base em incertezas.

Referências bibliográficas

1. Woolf SH. The need for perspective in evidence-based medicine. JAMA 1999; 282:2358-65.
2. Gomes MM, Kale PL. Qualidade das evidências: desenhos de pesquisa. In Gomes MM, ed. Medicina baseada em evidências: princípios e práticas. Rio de Janeiro: Reichmann & Affonso editores; 2001. p 17-35.
3. Antman EM. Clinical trials in cardiovascular medicine. Circulation 2001; 103:E101-E104.
4. Ellenberg SS, Temple R. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 2: practical issues and specific cases. Ann Intern Med 2000; 133:464-70.
5. Temple R, Ellenberg SS. Placebo-controlled trials and active-control trials in the evaluation of new treatments. Part 1: ethical and scientific issues. Ann Intern Med 2000; 133:455-63.
6. Food and Drug Administration. International Conference on Harmonisation; choice of control group and related issues in clinical trials; availability. Notice Fed Regist 2001; 66:24390-1.
7. Spodick DH. The randomized controlled clinical trial. Scientific and ethical bases. Am J Med 1982; 73:420-5.
8. Pereira BB. Estatística: a tecnologia da ciência I. Boletim da ABE 1997; 27-35.
9. Guyatt GH,.Rennie D. Users’ guides to the medical literature. JAMA 1993; 270:2096-7.
10. Guyatt GH, Sackett DL, Cook DJ. Users’ guides to the medical literature. II. How to use an article about therapy or prevention. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1993; 270:2598-601.
11. 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? Evidence-Based Medicine Working Group. JAMA 1994; 271:59-63.
12. Friedman L, Furberg C, DeMets D. Fundamentals of Clinical Trials. New York: Springer Verlag, 1998.

13. Kunz R,.Oxman AD. The unpredictability paradox: review of empirical comparisons of randomised and non-randomised clinical trials. BMJ 1998; 317:1185-90.
14. Heart Protection Study Collaborative Group. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. Lancet 2002; 360:7-22.
15. Stone GW, Grines CL, Cox DA, Garcia E, Tcheng JE, Griffin JJ et al. Comparison of angioplasty with stenting, with or without abciximab, in acute myocardial infarction. N Engl J Med 2002; 346:957-66.
16. Guyatt GH, Pugsley SO, Sullivan MJ, Thompson PJ, Berman L, Jones NL et al. Effect of encouragement on walking test performance. Thorax 1984; 39:818-22.
17. PROGRESS Collaborative Group. Randomised trial of a perindopril-based blood-pressure-lowering regimen among 6,105 individuals with previous stroke or transient ischaemic attack. Lancet 2001; 358:1033-41.
18. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1994; 271:389-91.
19. Dahlof B, Devereux RB, Kjeldsen SE, Julius S, Beevers G, Faire U et al. Cardiovascular morbidity and mortality in the Losartan Intervention For Endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. Lancet 2002; 359:995-1003.
20. Randomised trial of a perindopril-based blood-pressure-lowering regimen among 6,105 individuals with previous stroke or transient ischaemic attack. Lancet 2001; 358:1033-41.
21. Dahlof B, Lindholm LH, Hansson L, Schersten B, Ekbom T, Wester PO. Morbidity and mortality in the Swedish Trial in Old Patients with Hypertension (STOP-Hypertension). Lancet 1991; 338:1281-5.
22. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT). JAMA 2002; 288:2981-97.
23. The GUSTO investigators. An international randomized trial comparing four thrombolytic strategies for acute myocardial infarction. N Engl J Med 1993; 329:673-82.
24. White HD. Is heparin of value in the management of acute myocardial infarction? Cardiovasc Drugs Ther 1997; 11:111-9.
25. Myerburg RJ, Mitrani R, Interian A, Jr., Castellanos A. Interpretation of outcomes of antiarrhythmic clinical trials: design features and population impact. Circulation 1998; 97:1514-21.
26. Lopez-Jimenez F. Clinical interpretation of statistical significance. Rev Invest Clin 1996; 48:231-8.
27. Schwartz GG, Olsson AG, Ezekowitz MD, Ganz P, Oliver MF, Waters D et al. Effects of atorvastatin on early recurrent ischemic events in acute coronary syndromes: the MIRACL study: a randomized controlled trial. JAMA 2001; 285:1711-8.
28. Yusuf S, Zhao F, Mehta SR, Chrolavicius S, Tognoni G, Fox KK. Effects of clopidogrel in addition to aspirin in patients with acute coronary syndromes without ST-segment elevation. N Engl J Med 2001; 345:494-502.
29. Bernard GR, Vincent JL, Laterre PF, LaRosa SP, Dhainaut JF, Lopez-Rodriguez A et al. Efficacy and safety of recombinant human activated protein C for severe sepsis. N Engl J Med 2001; 344:699-709.
30. Dans AL, Dans LF, Guyatt GH, Richardson S. Users' guides to the medical literature: XIV. How to decide on the applicability of clinical trial results to your patient. Evidence-Based Medicine Working Group. JAMA 1998; 279:545-9.
31. The SOLVD Investigators. Effect of enalapril on survival in patients with reduced left ventricular ejection fractions and congestive heart failure. N Engl J Med 1991; 325:293-302.
32. Mettimano M, Lanni A, Migneco A, Specchia ML, Romano-Spica V, Savi L. Angiotensin-related genes involved in essential hypertension: allelic distribution in an Italian population sample. Ital Heart J 2001; 2:589-93.
33. Sagnella GA, Rothwell MJ, Onipinla AK, Wicks PD, Cook DG, Cappuccio FP. A population study of ethnic variations in the angiotensin-converting enzyme I/D polymorphism: relationships with gender, hypertension and impaired glucose metabolism. J Hypertens 1999; 17:657-64.
34. O'Toole L, Stewart M, Padfield P, Channer K. Effect of the insertion/deletion polymorphism of the angiotensin-converting enzyme gene on response to angiotensin-converting enzyme inhibitors in patients with heart failure. J Cardiovasc Pharmacol 1998; 32:988-94.
35. Sica DA,.Ghosh S. Pharmacotherapy in congestive heart failure: Effect of the insertion/deletion polymorphism of the ACE gene on response to ACE inhibitors in patients with heart failure. Congest Heart Fail 1999; 5:125-8.
36. Dargie HJ. Effect of carvedilol on outcome after myocardial infarction in patients with left-ventricular dysfunction: the CAPRICORN randomised trial. Lancet 2001; 357:1385-90.
37. CIBIS Investigators and Committees. A randomized trial of beta-blockade in heart failure. The Cardiac Insufficiency Bisoprolol Study (CIBIS). Circulation 1994; 90:1765-73.
38. Packer M, Coats AJ, Fowler MB, Katus HA, Krum H, Mohacsi P et al. Effect of carvedilol on survival in severe chronic heart failure. N Engl J Med 2001; 344:1651-8.
39. The Cardiac Insufficiency Bisoprolol Study II (CIBIS-II): a randomised trial. Lancet 1999; 353:9-13.

40. Waldo AL, Camm AJ, deRuyter H, Friedman PL, MacNeil DJ, Pauls JF et al. Effect of d-sotalol on mortality in patients with left ventricular dysfunction after recent and remote myocardial infarction. The SWORD Investigators. Survival With Oral d-Sotalol. Lancet 1996; 348:7-12.
41. Califf RM, DeMets DL. Principles from clinical trials relevant to clinical practice: Part I. Circulation 2002; 106:1015-21.
42. Yusuf S. Two decades of progress in preventing vascular disease. Lancet 2002; 360:2-3.
43. Hrobjartsson A, Gotzsche PC. Is the placebo powerless? An analysis of clinical trials comparing placebo with no treatment. N Engl J Med 2001; 344:1594-602.
44. Massel D, Cruickshank MK. The number remaining at risk: an adjunct to the number needed to treat. Can J Cardiol 2002; 18:254-8.
45. Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS et al. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. N Engl J Med 2002; 346:877-83.
46. Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. BMJ 2002; 324:71-86.
47. Yusuf S, Peto R, Lewis J, Collins R, Sleight P. Beta blockade during and after myocardial infarction: an overview of the randomized trials. Prog Cardiovasc Dis 1985; 27:335-71.
48. Rich-Edwards JW, Stampfer MJ, Manson JE, Rosner B, Hankinson SE, Colditz GA et al. Birth weight and risk of cardiovascular disease in a cohort of women followed up since 1976. BMJ 1997; 315:396-400.
49. Vaessen N, Janssen JA, Heutink P, Hofman A, Lamberts SW, Oostra BA et al. Association between genetic variation in the gene for insulin-like growth factor-I and low birthweight. Lancet 2002; 359:1036-7.
50. Brown MJ, Palmer CR, Castaigne A, de Leeuw PW, Mancia G, Rosenthal T et al. Morbidity and mortality in patients randomised to double-blind treatment with a long-acting calcium-channel blocker or diuretic in the International Nifedipine GITS study: Intervention as a Goal in Hypertension Treatment (INSIGHT). Lancet 2000; 356:366-72.
51. Moyé LA. Alpha calculus in clinical trials: considerations and commentary for new millennium. Statist Med 2000; 19:767-79.
52. Whitley E, Ball J. Statistics review 4: sample size calculations. Crit Care 2002; 6:335-41.
53. Schriger DL. How do we draw inference from "negative" studies? Ann Emerg Med 2003; 41:69-71.
54. Goodman SN, Berlin JA. The use of predicted confidence intervals when planning experiments and the misuse of power when interpreting results. Ann Intern Med 1994; 121:200-6.
55. Le Fanu J. The Rise and Fall of Modern Medicine. New York: Avalon Publishing Group; 2002.
56. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. BMJ 1995; 310:452-4.
57. Goodman SN. P values, hypothesis tests, and likelihood: implications for epidemiology of a neglected historical debate. Am J Epidemiol 1993; 137:485-96.
58. Sterne JAC, Smith GD, Cox DR. Sifting the evidence - what's wrong with significance tests? Another comment on the role of statistical methods. BMJ 2001; 322:226-31.
59. Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. Ann Intern Med 1999; 130:995-1004.
60. Browner WS, Newman TB. Are all significant P values created equal? The analogy between diagnostic tests and clinical research. JAMA 1987; 257:2459-63.
61. Goodman SN, Royall R. Evidence and scientific research. Am J Public Health 1988; 78:1568-74.