# ENSAIO CLÍNICO RANDOMIZADO: VERDADEIRO OU FALSO?

## *RANDOMIZED CLINICAL TRIAL: TRUE OR FALSE?*

"A fé independe da ciência, mas a ciência depende da fé."

Reinaldo José Gianini*

Recente artigo publicado no *Journal of the American College of Cardiology*,[1] nos remete às limitações dos Ensaios Clínicos Randomizados (Randomized Clinical Trials - RCT).

Os autores focam sua discussão em três aspectos críticos: relevância clínica versus significância estatística dos resultados, composição de desfechos, e análise de subgrupos.

No mesmo periódico, Stone e Pocock[2] comentam, aprofundam e ampliam esta discussão.

O objetivo do presente trabalho é fornecer subsídios para melhor compreensão, por parte dos médicos, dos argumentos publicados nesses dois artigos.

### CARACTERÍSTICAS DO RCT

O RCT é considerado o modelo de estudo com maior força de evidência clinicoepidemiológica.

A evidência clinicoepidemiológica se caracteriza por: valorizar desfechos clínicos de significância ao paciente e à sociedade, permitir a definição de graus de evidência científica para as condutas clínicas, apresentar dados para análise objetiva do potencial impacto das condutas clínicas.

RCT com desfechos bioquímicos, fisiológicos ou celulares não são classificados como forte evidência, mas como evidências de força intermediária.

Esta classificação define a recomendação favorável ou contrária à intervenção clínica (Quadro 1).

| CATEGORIA | DEFINIÇÃO | JUSTIFICATIVAS PARA INTERVENÇÃO |
|---|---|---|
| A | Evidência forte e benefício clínico significativo apóiam a recomendação. | Benefício medido supera risco e custo calculados. |
| B | Evidência intermediária, mas benefício potencial alto, com risco e custo aceitáveis ou evidência forte com benefício limitado apóiam a recomendação. | Dificuldade em alcançar evidência mais conclusiva (por razões éticas, logísticas ou financeiras). |
| C | Evidência fraca apóia a recomendação a favor ou prognóstico sabidamente ruim e ausência de alternativa contra a intervenção. | Risco (custo) da intervenção seguramente baixo e benefício potencial alto. |

Quadro 1. Categorias de recomendação de uma intervenção clínica. Adaptado de Duncan e Schmidt[3]

**Rev. Fac. Ciênc. Méd. Sorocaba, v. 12, n. 4, p. 27 - 31, 2010**
* Professor do Depto. de Medicina - FCMS/PUC-SP
*Recebido em 9/4/2010. Aceito para publicação em 16/7/2010.*
Contato: rgianini@pucsp.br

Os Ensaios clínicos compreendem as seguintes fases: Fase pré-clínica (não humana - ensaios com modelos em laboratório); Fase clínica I - segurança em indivíduos sadios; II - segurança e eficácia em pacientes; III - eficácia comparada do medicamento; IV - efetividade e segurança ou teste no mundo real.[4]

Os RCT que aqui discutimos referem-se à Fase III.

O RCT padrão-ouro é duplo-cego, no qual nem pesquisador nem paciente sabe se pertence ao grupo de tratamento ou ao grupo controle, pois esse procedimento reduz o viés de classificação (ou de informação).

O viés de classificação pode ocorrer em relação ao tratamento ou ao desfecho. Recomenda-se a análise por intenção de tratamento (e não por tratamento completado) porque tem a vantagem de agregar não só a informação sobre o efeito específico do tratamento como também de outros efeitos (adversos, colaterais) e fatores (acesso, adesão), traduzindo benefício mais próximo do real. Mesmo assim, deve ser realizada a análise das perdas de seguimento ou inversões não-intencionais de tratamento, e a análise da adesão ao tratamento, de preferência com marcador biológico. A escolha do desfecho a ser analisado também é de fundamental importância para a atenuação do viés de classificação. O desfecho deve ser objetivo e claro, passível de observação e mensuração inequívocas.

Outra fonte de erro dos RCT é o viés de seleção. O ponto forte dos RCT é validade interna. A randomização tende a equilibrar a distribuição dos fatores associados (conhecidos ou não) ao efeito de tratamento entre grupo controle e grupo de tratamento. Entretando, o confundimento residual pode ocorrer mesmo com a randomização, principalmente em pequenas amostras. E isso tende a se agravar na análise de subgrupos.

O ponto fraco dos RCT é a validade externa (generalização), muito comprometida, pois são estudos restritos a amostras muito especiais, com características diferentes da população. RCT multicêntricos podem melhorar a validade externa. De qualquer modo, é isto que justifica a necessidade da Fase IV.

Existe, ainda, o erro aleatório. Sempre que se tratar de amostra, há a possibilidade de o acaso incidir sobre os resultados do estudo, de modo que o efeito do tratamento encontrado seja diferente do real (efeito na população). Quanto maior a amostra menor o erro aleatório ou, em outras palavras, maior a precisão da medida de efeito. O cálculo do Intervalo de Confiança traduz exatamente esta dimensão.

## SIGNIFICÂNCIA ESTATÍSTICA E RELEVÂNCIA CLÍNICA

A significância (p) é função do tamanho da amostra e da diferença entre tratamentos. Então: $n = (2.s/d^2).[Z(p/2)+Z\ beta].^2$ Onde: s = variância; d = diferença entre grupo de tratamento e grupo controle; p = erro tipo I = significância; beta = erro tipo II = poder estatístico; Z = escala na distribuição de Gauss.

Na figura 1 podemos observar que os estudos A e E têm o mesmo nível de significância (p = 0,05), mas por razões diferentes: o estudo A pela elevada diferença entre tratamentos, apesar do pequeno n; o estudo E pelo grande tamanho da amostra, apesar da pequena diferença entre tratamentos. O estudo A apresenta problema de validade externa - generalização, mas elevada relevância clínica. O estudo E apresenta melhor validade externa, mas pouca relevância clínica.

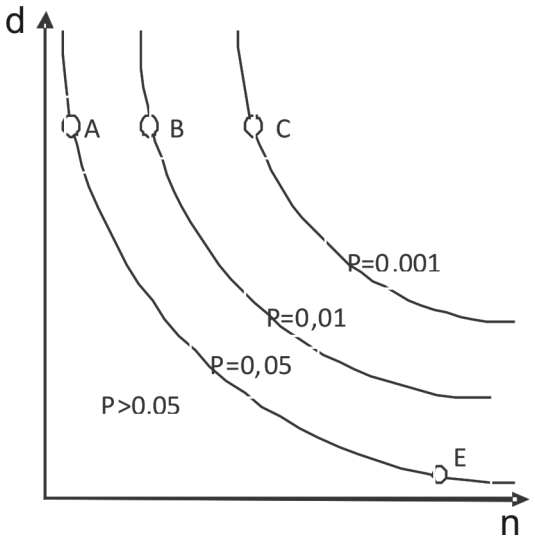


Figura 1. Relação entre Níveis de significância (P), Diferença entre tratamentos (d) e Tamanho da amostra (n). Exemplos de estudos: Validade externa de A < B < C; E – menor relevância clínica; Razão de Verossimilhança de C > B > A.

Finalmente, quando se compara A, B e C, estudos com a mesma diferença entre tratamentos, mas com tamanhos de amostras diferentes, observa-se que a significância de C > B > A de acordo com o n. Se forem calculadas as RV desses exemplos também teríamos C > B > A. Pois para uma mesma d, quanto maior o n menor o p, e maior a RV. Ou, para um mesmo n, quanto maior a d menor o p, e maior a RV.[5]

Na figura 2, procura-se demonstrar como a Razão de Verossimilhança (RV) calculada para H1 aumenta em função da diferença de tratamento. Apresentam-se 11 ensaios, todos com n =100, com diferenças de tratamento (ou benefício) que variam de 0% a 10%. H0 foi fixada como 0,5 (igual probabilidade de benefício para tratamento e controle) e H1 foi fixada como 0,55 (5% de benefício para tratamento). A RV é calculada dividindo-se a probabilidade dos resultados observados em determinado estudo dado que ocorreu H1 pela probabilidade dos resultados observados em determinado estudo dado que ocorreu H0, ou RV = P(resultados|H1)/P(resultados|H0), segundo a distribuição Binomial de probabilidades. Para um benefício de 10% a RV de H1 é igual a 6, em outras palavras 6 chances de H1 ser verdadeira para 1 chance de H0 ser verdadeira (6/7 ou 86%). Para um benefício de 5% a RV seria igual a 3.

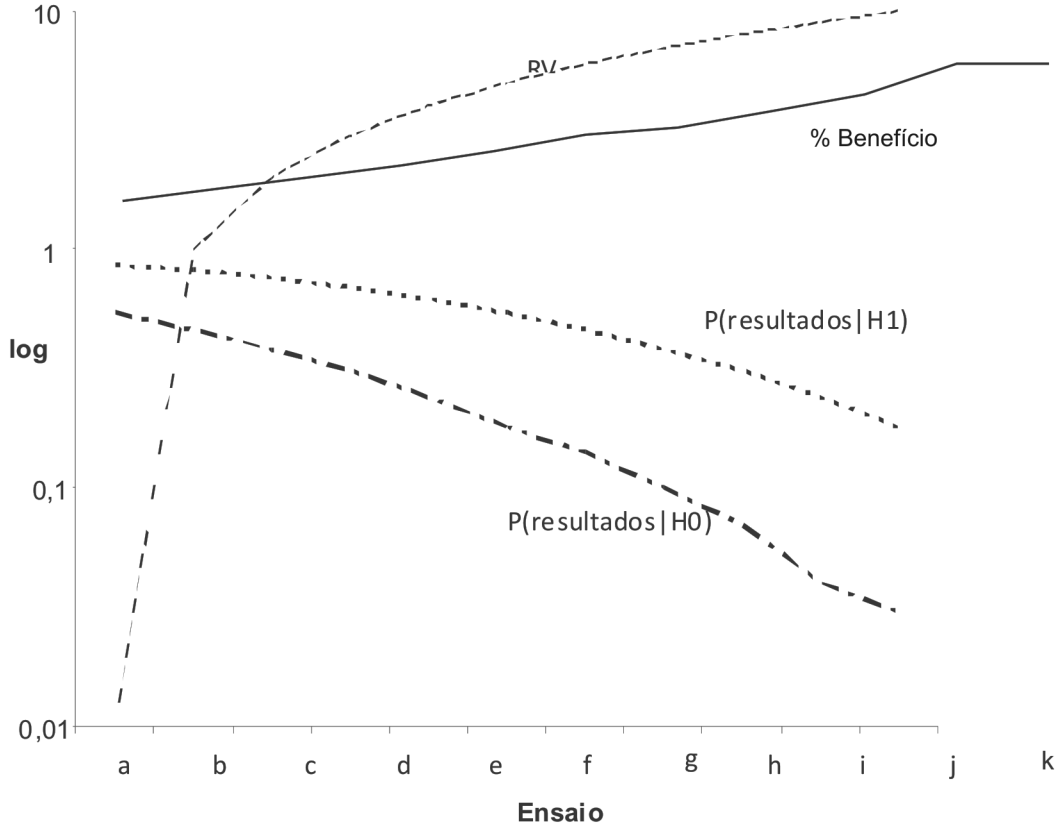


Figura 2. Análise Bayesiana dos Resultados de Ensaios

## COMPOSIÇÃO DE DESFECHOS E ANÁLISE DE SUBGRUPOS

Outro modo de analisar a relevância clínica da intervenção é o Número Necessário para Tratar (NNT), que seria o número mínimo de pacientes que precisa ser tratado para se ter um caso beneficiado pelo tratamento. O NNT é função da diferença entre tratamentos e da frequência do evento na população (incidência ou prevalência): NNT = 1/risco dos não-tratados – risco dos tratados.

Na figura 3, observamos que os estudos A e E têm o mesmo NNT apesar de suas características diferentes.

O estudo A apresenta elevada diferença entre tratamentos, mas trata de evento pouco frequente na população, tendo, por isso, um NNT elevado.

O estudo E trata de evento frequente na população, mas apresenta baixa diferença entre tratamentos, também resultando em NNT elevado (NNT menores que 50 são considerados clinicamente relevantes). Ainda, comparando-se A, B e C, observa-se que a relevância clínica de C > B > A apesar de apresentarem a mesma diferença entre tratamentos (o que ocorre por causa da frequência dos eventos de que eles tratam C > B > A). O mesmo ocorreria para um mesmo n se tivéssemos d diferentes: quanto maior a d menor o NNT, maior a relevância clínica.

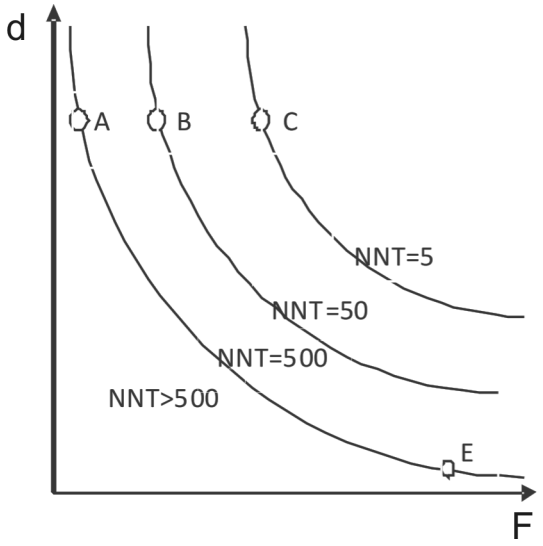


Figura 3. Relação entre Número Necessário para Tratar (NNT), Diferença entre tratamentos (d) e Frequência do evento na população (F) . F = casos/população. Exemplos de estudos: A – menor impacto populacional; E – menor valor clínico-individual; Relevância C>B>A.

Por este motivo a composição de desfechos é uma estratégia amplamente utilizada para aumentar o poder estatístico do estudo graças à maior frequência de eventos.

Um exemplo seria um estudo que incluísse na análise do desfecho a Presença de Onda de lesão no ECG (Onda Q do Infarto do Miocárdio), além da Mortalidade por Infarto Agudo do Miocárdio. No quadro 2 descrevem-se os critérios elencados por Kaul e Diamond para a composição de desfechos:

| |
|---|
| Justificar a validade dos componentes individuais – medem o mesmo desfecho clinicoepidemiológico. |
| Evitar componentes individuais mal definidos ou que não sejam clinicamente importantes para o desfecho estudado. |
| Evitar componentes individuais do desfecho que não sejam claramente passíveis de modificação pelo tratamento. |
| Descrever resultados do desfecho primário composto e dos componentes individuais do desfecho separadamente. |
| Examinar modificação de efeito de tratamento segundo componentes individuais do desfecho e testar heterogeneidade. |
| Atribuir pesos aos componentes individuais do desfecho, prospectivamente, de acordo com sua importância clínica. |
| Realizar e descrever análise de sensibilidade relativa aos pesos atribuídos aos componentes individuais do desfecho. |

Quadro 2. Critérios para a utilização correta de desfechos compostos

A análise de subgrupos tem por função verificar se o efeito do tratamento é mais benéfico quando os pacientes apresentam alguma característica específica. Orienta desse modo a indicação do tratamento para o clínico, e a política de saúde para o gestor. Mas seu resultado, em termos estatísticos, é inverso à composição de desfechos: reduz o poder estatístico porque divide o n. Assim, aumentam as chances de se aceitar uma diferença falsa ou de se rejeitar uma diferença verdadeira devido ao acaso. O quadro 3 apresenta critérios para a análise de subgrupos.[3]

| |
|---|
| Definir hipóteses prospectivamente. |
| Limitar análise ao biologicamente plausível, com base em evidências pregressas. |
| Limitar a análise aos estudos que apresentaram significância estatística para o efeito do tratamento quando da análise global pré-definida. |
| Identificar modificações de efeito de tratamento estatisticamente significantes associadas aos subgrupos. |
| Ajustar significância (p) para análise de comparações múltiplas* |
| Descrever resultados de subgrupos como exploratórios, que demandam futura pesquisa. |
| Evitar interpretações que superestimem as diferenças encontradas nos subgrupos. |

Quadro 3. Critérios para a análise de subgrupos
*p = 1- (1-0,05) x, onde x = número de comparações de subgrupos e 0,05 corresponde ao erro tipo I

## CONCLUSÃO

Apesar de o RCT, com razão, ser considerado o melhor modelo de estudo para fornecer evidências clínico-epidemiológicas, os estudos variam muito em qualidade, são passíveis de vieses e apresentam importantes limitações. Portanto, não dispensa a boa crítica, oriunda da vivência e do conhecimento acumulados por nós, médicos.

A estatística é a ciência do provável. Isoladamente, não tem o poder de decidir sobre o que é verdade ou o que é real. Se o médico está convicto de que determinada intervenção clínica é segura, apresenta custo aceitável e beneficia o paciente, a falta de significância estatística ($p > 0,05$) não deve impedir a indicação do tratamento. E o contrário também é válido: se o médico não está convicto de que determinada intervenção é segura, apresenta custo aceitável e beneficia o paciente, a significância estatística ($p < 0,05$) não é suficiente para a indicação daquele tratamento.

## Nota sobre o autor

Reinaldo José Gianini é médico (FCM Sorocaba), especialista em Saúde Pública (FSPUSP), mestre em Medicina Preventiva (FMUSP), doutor em Medicina Preventiva (FMUSP), com pós-doutorado em Políticas de Saúde (London School of Hygiene & Tropical Medicine). É professor titular do Depto. de Medicina da FCMS Sorocaba - Área de Medicina Preventiva, e Pesquisador do LIM-39 HC-FMUSP.

## REFERÊNCIAS

1. Kaul S, Diamond GA. Trial and error: how to avoid commonly encountered limitations of published clinical trials. J Am Coll Cardiol. 2010; 55(5):415-27.
2. Stone GW, Pocock SJ. Randomized trials, statistics, and clinical inference. J Am Coll Cardiol. 2010; 55(5):428-31.
3. Duncan BB, Schmidt MI, Giugliani, ERJ. Medicina ambulatorial: condutas de atenção primária baseadas em evidências. 3ª ed. São Paulo: Artmed; 2004.
4. Fletcher RH, Fletcher SW. Epidemiologia clínica: elementos essenciais. 4ª ed. São Paulo: Artmed; 2006.
5. Gianini RJ. Bayesianismo. Rev Fac Ciênc Méd Sorocaba. 2009; 11(1):27-9.