

## 03.03 Epidemiología clínica: metodología e aplicações



A avaliação da validade e da confiabilidade dos testes diagnósticos e da eficácia das intervenções terapêuticas mediante o desenho de estudos adequados constitui a principal tarefa da epidemiologia clínica, cujo fim último é a melhoria da qualidade da prática clínica.

**Autor: Miguel Ángel Royo Bordonada**

*Jefe de Área de Estudios de la Escuela Nacional de Sanidad*

*(Texto original de Miguel Ángel Royo Bordonada, revisado e adaptado por Suzana Alves da Silva - Pesquisadora do LIGRESS/HCor – Laboratório de Inovação em Planejamento, Gestão, Avaliação e Regulação de Políticas, Sistemas, Redes e Serviços de Saúde / Hospital do Coração)*

Se recomienda imprimir 2 páginas por hoja

### Citación recomendada:

Royo Bordonada MA. Epidemiología clínica: metodología e aplicações[Internet]. Madrid: Escuela Nacional de Sanidad; Traducción 2015 [consultado día mes año]. Disponible en: direccion url del pdf.



TEXTOS DE ADMINISTRACION SANITARIA Y GESTIÓN CLINICA  
by UNED Y ESCUELA NACIONAL DE SANIDAD  
is licensed under a Creative Commons  
Reconocimiento- No comercial-Sin obra Derivada  
3.0 Unported License.



## Introdução

A epidemiologia clínica surge diante da necessidade, por parte dos profissionais da saúde, de dispor de evidência científica contrastada com o que se enfrenta no contexto de incerteza, no qual se desenvolve com frequência a prática clínica. A aplicação do método científico epidemiológico no ambiente clínico responde a esta demanda, facilitando ao profissional o uso mais apropriado do arsenal diagnóstico e terapêutico disponível em cada momento.

## *Introdução*

### *2 Viés (Erro Sistemático) versus Acaso (Erro Aleatório)*

### *3 Reprodutibilidade*

#### *3.1.1 Índice de Concordância Global (IC)*

#### *3.1.2 Índice Kappa (IK)*

#### *3.1.3 Coeficiente de correlação intraclasse (CCI)*

### *4 Avaliação da Acurácia de Testes Diagnósticos*

#### *4.1 Validade de testes diagnósticos*

#### *4.2 Análise da validade (risco de viés) de estudos seccionais para avaliação da acurácia de testes diagnósticos*

#### *4.3 Apresentação dos Resultados*

##### *4.3.1 Sensibilidade (S)*

##### *4.3.2 Especificidade (E)*

##### *4.3.3 Razão de Verossimilhança (RV)*

##### *4.3.4 Curvas ROC*

#### *4.4 Utilização dos testes diagnósticos*

##### *4.4.1 Limiar de decisão*

##### *4.4.2 Probabilidade pré-teste*

##### *4.4.3 Probabilidade pós-teste*

### *5 Ensaio Clínico Randomizado*

#### *5.1 Análise da validade (risco de viés) de ensaios clínicos randomizados*

##### *5.1.1 Randomização*

##### *5.1.2 Cegamento*

##### *5.1.3 Perda de seguimento*

##### *5.1.4 Intenção de tratar*

##### *5.1.5 Interrupção precoce por benefício*

#### *5.2 Apresentação dos Resultados*

#### *5.3 Utilização de ensaios clínicos randomizados*

##### *5.3.1 Aplicabilidade dos resultados*

##### *Rastreamento*

### *6 Referências*

#### *Referências bibliográficas*

*A epidemiologia clínica pretende dar resposta às perguntas que surgem na prática clínica, mediante o uso de um método científico comparado que garante a validade e a confiabilidade das evidências obtidas<sup>5</sup>.*

Os principais incentivadores da epidemiologia clínica têm proposto as seguintes definições para esta disciplina:

- Estudo dos determinantes e dos efeitos nas decisões clínicas<sup>1</sup>.
- É a aplicação dos métodos biométricos epidemiológicos ao estudo do diagnóstico e tratamento por um clínico que proporciona atendimento médico direto a pacientes<sup>2</sup>.
- Utilização da ferramenta epidemiológica em subpopulações hospitalares, com o objetivo fundamental de avaliar estratégias de diagnóstico e tratamento<sup>3</sup>.
- Prática da epidemiologia na solução de problemas pelos clínicos<sup>4</sup>.

Independente da definição que escolhermos, todos os autores concordam que a epidemiologia clínica busca dar resposta às perguntas que surgem na prática clínica relacionadas com o diagnóstico, o tratamento, o prognóstico e a prevenção dos pacientes, mediante o uso de um método científico comparado que garante a validade e confiabilidades das evidências obtidas<sup>5</sup>. Desta forma, a epidemiologia clínica estabeleceu os fundamentos sobre os quais se desenvolveu durante as últimas décadas o fenômeno conhecido como Medicina Baseada em Evidências<sup>6</sup>. O desafio para o profissional da saúde, na prática clínica diária, consiste em vincular da melhor forma possível a evidência científica obtida a partir da abordagem de grupo (populacional) da epidemiologia clínica com cada paciente individual, e aí é onde reside o que Sackett tão acertadamente denominou a ciência da arte da medicina clínica<sup>7</sup>.

## **2. Viés (Erro Sistemático) versus Acaso (Erro Aleatório)**

Viés é “qualquer tendência na coleta, análise, interpretação, publicação ou revisão de dados que possa levar a conclusões que sejam sistematicamente diferentes da verdade”.<sup>8</sup> Acaso é uma variação aleatória entre a observação de um fenômeno em diferentes amostras e o valor verdadeiro na população. Ao contrário do viés que tende a distorcer o resultado de forma sistemática para um sentido ou outro, o erro aleatório resulta

em observações em torno do valor verdadeiro, mas sem uma direção específica. Nos estudos clínicos o erro aleatório é avaliado através do intervalo de confiança dos resultados, que demonstra os extremos da variação de um resultado onde provavelmente reside o valor verdadeiro, com uma precisão estimada geralmente de 95%.

A presença ou não de erros sistemáticos em uma pesquisa é que determina a validade interna ou acurácia daquela pesquisa, ou seja, o grau com que os resultados da pesquisa estão corretos para a amostra de pacientes analisada. O erro aleatório diz respeito à confiabilidade ou precisão dos resultados daquela pesquisa realizada, naquela amostra selecionada de pacientes, para representar o valor verdadeiro na população geral.

Na Figura 1 observam-se quatro alvos sobre os quais foi realizada uma série de disparos consecutivos com a mesma arma. Assumindo que o atirador não tenha cometido nenhum tipo de erro ao apontar e disparar, cada alvo representa uma situação diferente: o superior esquerdo representa uma medida válida (os disparos acertaram o alvo) e confiável (os disparos estão agrupados neste alvo); a inferior esquerda é bastante válida (os disparos estão ao redor do centro do alvo), mas pouco confiável (os disparos estão dispersos em torno do alvo); a superior direita é muito confiável, mas pouco válida (os disparos estão agrupados, mas afastados do centro do alvo) e a inferior direita nem é válida nem confiável.

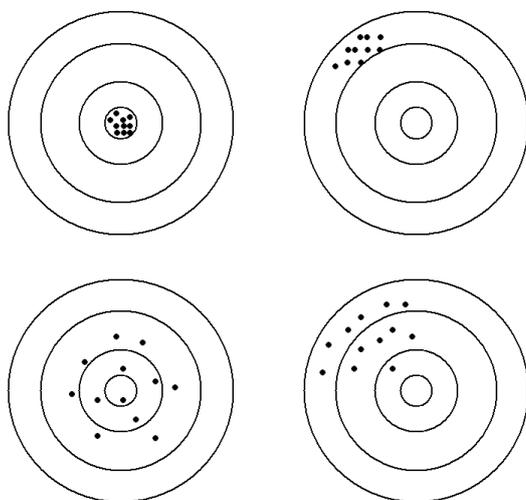


Figura 1. Avaliação da validade e confiabilidade do método de aferição.

Na prática o viés (erro sistemático) e o acaso (erro aleatório) são tratados de forma diferente. O viés pode ser evitado pela condução de investigações clínicas e interpretação dos dados coletados com um método apropriado. Ele pode ser detectado através da leitura criteriosa da seção de métodos do estudo publicado. O acaso, por outro lado, não tem como ser eliminado ou evitado, mas a precisão dos resultados pode ser aumentada com um aumento no tamanho da amostra, por exemplo, até que o intervalo de confiança do resultado não mais gere incertezas em relação à conclusão final da pesquisa, se positiva ou negativa.

---

### **3. Reprodutibilidade**

---

Outra forma de se avaliar a confiabilidade de um método de aferição é através da análise da sua reprodutibilidade. A reprodutibilidade do método representa sua capacidade de proporcionar resultados similares quando sua aplicação é repetida em condições idênticas; ou seja, o grau de precisão ou estabilidade alcançado em sucessivas aplicações do mesmo método de aferição do fenômeno. Por exemplo, a aferição dos níveis de colesterol em um indivíduo com um valor real de colesterol plasmático de 200 mg/dl poderia, com as técnicas laboratoriais habituais, apresentar em medições consecutivas variação dos valores de dosagem de 170 mg/dl a 230 mg/dl, com implicações opostas na decisão terapêutica. Isto ocorre porque qualquer variável (biológica ou não) está submetida às seguintes fontes de variação:

- a variabilidade que depende do aparelho de medição, pelo diferente grau de calibração do mesmo ou pelo efeito de fatores ambientais como a temperatura ou a iluminação;
- a variabilidade intraobservador, por modificar o grau de atenção do profissional encarregado de tomar as medidas em razão de cansaço ou condições ambientais, tanto físicas como psicológicas (pressão diante de uma situação de necessidade de diagnóstico urgente);
- a variabilidade interobservador, devido ao grau de conhecimento prévio diferente, maior ou menor acuidade visual (para discriminar uma imagem radiológica) ou auditiva

(para auscultar um sopro), maior ou menor experiência ou capacidade para integrar o restante dos sinais e sintomas ao resultado do exame que se tenta medir, ou um diferente nível de motivação, entre outros possíveis fatores;

- a variabilidade biológica inerente à característica que se deseja medir (o colesterol plasmático e a pressão arterial variam no decorrer do dia, a primeiro de forma particular em resposta à última refeição realizada e a segunda em função da postura (sentado, deitado), nível de estresse ao qual o indivíduo está submetido no momento da medição, como na hipertensão do jaleco branco, provocada como resposta ao simples fato de que um profissional de saúde esteja examinando um paciente;
- e, por último, a variabilidade aleatória, que representa o grau de variação que não pode ser atribuída a nenhum dos fatores anteriores e que, como consequência, atribuímos ao acaso.

A única forma de aumentar a reprodutibilidade das medições consiste em reduzir ao máximo possível à variabilidade inerente a elas, atuando sobre cada uma das fontes de variação mencionadas. Para isso é necessário padronizar o método de aferição e as suas condições de aplicação.

O "*Seventh Report of the Joint National Committee on Detection, Evaluation, and Treatment of High Blood Pressure*",<sup>9</sup> por exemplo, estabeleceu um método metuculoso para aferição da pressão arterial com esfigmomanômetro (Quadro 1).

**Quadro 1. *Accurate Blood Pressure Measurement in the Office***

A reprodutibilidade de um método reflete o grau de concordância que ele é capaz de proporcionar ao se repetir a sua aplicação em condições similares.

*The accurate measurement of BP is the sine qua non for successful management. The equipment—whether aneroid, mercury, or electronic—should be regularly inspected and validated. The operator should be trained and regularly retrained in the standardized technique, and the patient must be properly prepared and positioned. The auscultatory method of BP measurement should be used. Persons should be seated quietly for at least 5 minutes in a chair (rather than on an exam table), with feet on the floor, and arm supported at heart level. Caffeine, exercise, and smoking should be avoided for at least 30 minutes prior to measurement. Measurement of BP in the standing position is indicated periodically, especially in those at risk for postural hypotension, prior to necessary drug dose or adding a drug, and in those who report symptoms consistent with reduced BP upon standing. An appropriately sized cuff (cuff bladder encircling at least 80 percent of the arm) should be used to ensure accuracy. At least two measurements should be made and the average recorded. For manual determinations, palpated radial pulse obliteration pressure should be used to estimate SBP—the cuff should then be inflated 20–30 mmHg above this level for the auscultatory determinations; the cuff deflation rate for auscultatory readings should be 2 mmHg per second. SBP is the point at which the first of two or more Korotkoff sounds is heard (onset of phase 1), and the disappearance of Korotkoff sound (onset of phase 5) is used to define DBP. Clinicians should provide to patients, verbally and in writing, their specific BP numbers and the BP goal of their treatment.*

Retirado de *National Heart, Lung, and Blood Institute (US); 2004. Report No. 04-5230<sup>8</sup>.*

Para avaliar a reprodutibilidade de um teste diagnóstico ou de qualquer outro instrumento de aferição, como por exemplo, um questionário, é preciso medir a variável em questão em duas ou mais ocasiões em cada indivíduo e posteriormente comparar qual é a diferença nos resultados dessas medições repetidas. Se o que se pretende medir é o grau de estabilidade ou consistência da medição, deve-se manter constantes ou submetidos ao mínimo grau de variação possível o restante dos fatores mencionados anteriormente que influem na variabilidade do resultado. O mesmo princípio é aplicável se o que se pretende medir é a reprodutibilidade entre diferentes observadores ou entre diferentes instrumentos de aferição; ou seja, quando se pretende medir o grau de concordância ou equivalência dos resultados obtidos por eles. Os índices ou estatísticas utilizados para medir a reprodutibilidade dependem do caráter da variável que está sendo medida. Quando se tratar

de variáveis categóricas (“doente-saudável”, “leve-moderado-grave”, “positivo-negativo”) utiliza-se o Índice de Concordância Global e o Índice Kappa; enquanto que no caso de variáveis quantitativas aplica-se o Coeficiente de Correlação Intraclasse.

### 3.1.1 Índice de Concordância Global (IC)

O IC representa o percentual de concordância entre os observadores (ou entre medições repetidas de um único observador) sobre o total de observações realizadas. A título de exemplo, um painel de especialistas foi convidado a revisar os históricos médicos e demais informações relevantes para avaliar a causa da morte dos participantes de um ensaio clínico com estatinas e, em particular, o número de mortes decorrentes da doença cardíaca isquêmica. Os resultados da comparação da avaliação feita pelo painel de especialistas com as do pesquisador do estudo são mostrados na tabela 2 (os dados entre parênteses são os valores esperados devido ao acaso, calculados conforme descrito na seção seguinte sobre o Índice Kappa).

Tabela 2. Análise da concordância entre observadores.

		Painel de Especialistas		
		DCI (Ce)	Outros (Ce)	Total
Pesquisador	DCI	184 (108)	36 (112)	220
	Outros	23 (99)	180 (104)	203
	Total	207	216	423

DCI: Doença Cardíaca Isquêmica; Ce: Concordância esperada

Dos 220 óbitos atribuídos à doença cardíaca isquêmica pelo pesquisador, em 36 casos o painel de especialistas opinou que a causa da morte era outra, enquanto que das 203 mortes atribuídas pelo pesquisador a outras causas, 23 foram atribuídas a doença cardíaca isquêmica segundo o painel de especialistas. Portanto,  $IC = (184 + 180) / 423 = 0,86$ ; o que representa uma concordância de 86%.

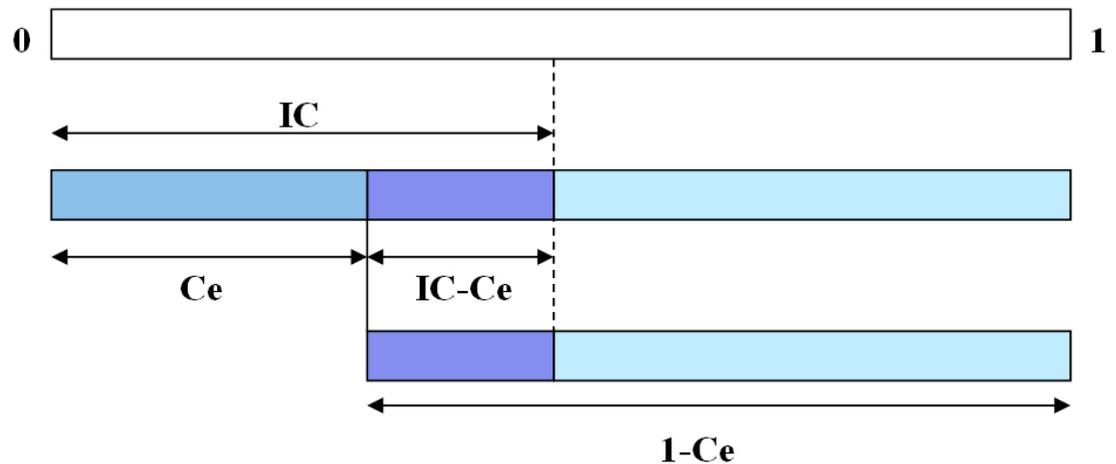
### 3.1.2 Índice Kappa (IK)

No exemplo anterior, se em vez de um painel de especialistas tivéssemos lançado uma moeda para cima para estabelecer o diagnóstico da causa do óbito (cara=doença cardíaca isquêmica

e coroa=outras causas), em algumas ocasiões o resultado teria coincidido com o do pesquisador por puro acaso. O IK ajusta a porcentagem de concordância observada, ou seja, o IC, pela porcentagem de concordância esperada exclusivamente pelo acaso ( $C_e$ ), por meio da seguinte fórmula:

$$\mathbf{IK = (IC - C_e) / (1 - C_e)}$$

Esta fórmula representa, tal como se pode ver na figura 2, a concordância existente além da que se esperaria somente pelo acaso. Como regra geral, a reprodutibilidade é considerada pequena se o IK for menor do que 0,4; boa se estiver entre 0,4 e 0,75 e excelente se for superior a 0,75.



Para se calcular os valores esperados pelo acaso em cada célula, multiplicam-se o valor total da linha com o valor total da coluna correspondente à célula. Este valor dividido pelo total de observações representa o valor esperado pelo acaso para aquela célula. Em nosso exemplo, os resultados esperados são os seguintes:  $220 \times 207 / 423 = 108$ ;  $220 \times 216 / 423 = 112$ ;  $203 \times 207 / 423 = 99$  e  $203 \times 216 / 423 = 104$ . Uma vez obtidos os valores esperados, a porcentagem de concordância esperada devida ao acaso ( $C_e$ ) é obtida do mesmo modo que o IC, em nosso exemplo, na tabela 2,  $(104 + 108) / 423 = 0.5012$ ; ou seja, 50,12%. Considerando que o IC era de 86%, o Índice Kappa é de:

$$\mathbf{IK = (0.86 - 0.5012) / (1 - 0.5012) = 0.72}$$

Ou seja, uma vez eliminado o acaso, a porcentagem de concordância no diagnóstico da causa do óbito entre o investigador do estudo e o painel de especialistas seria de 72%.

---

### **1.1.3 Coeficiente de correlação intraclasse (CCI)**

---

Para se estimar a reprodutibilidade das medições associadas a variáveis quantitativas, o teste estatístico que se deve utilizar é o CCI. Ele não deve ser confundido com os coeficientes de correlação ( $R^2$ ) tais como o de Pearson, o de Spearman e o de Tau, que refletem unicamente o grau de associação existente entre duas variáveis numéricas. Suponhamos que se multiplicarmos por 10 os valores de pressão arterial obtidos em uma amostra de indivíduos, os novos valores resultantes seriam absurdos já que, por exemplo, ninguém poderia ter uma medida de pressão arterial sistólica de 1.400 mmHg. O grau de equivalência entre os dois valores seria virtualmente nulo e, entretanto, a associação entre as duas variáveis seria perfeita, já que com qualquer modelo de regressão linear simples poderíamos prever os resultados de uma variável a partir dos resultados da outra variável e vice-versa, simplesmente multiplicando-se ou dividindo-se seus valores por 10. Quer dizer, que enquanto que o coeficiente de correlação ( $R^2$ ) seria igual a 1 (o máximo possível), o CCI seria virtualmente igual a 0. Os coeficientes de correlação ignoram o único aspecto que tem importância ao se avaliar a reprodutibilidade: a diferença absoluta existente entre os dois valores, já que, quanto menor ela for, maior é o grau de concordância entre os observadores ou entre medidas repetidas.

No exemplo da tabela 3, dois médicos tomaram a pressão arterial sistólica de 10 pacientes de forma consecutiva e com o mesmo esfigmomanômetro. Como se pode observar, as medições dos dois médicos são um pouco diferentes. Além disso, esta diferença é sistemática, pois os valores registrados pelo médico B são em todos os casos superiores aos registrados pelo médico A e, além disso, a diferença é quase sempre da mesma magnitude (em 8 das 10 ocasiões a diferença é de 5 mmHg, enquanto que nas duas restantes é de 10 mmHg). O fato de que as diferenças sejam no mesmo sentido e de magnitude similar faz com que o valor do coeficiente de correlação seja muito alto, concretamente de 0,95; entretanto, essas mesmas diferenças fazem com que o valor do CCI seja muito menor, concretamente 0,64.

Tabela 3. Medição da pressão arterial sistólica.

Paciente	Médico A (mmHg)	Médico B (mmHg)	Diferença (B – A)
1	135	140	5
2	140	145	5
3	130	135	5
4	145	150	5
5	140	145	5
6	150	160	10
7	140	145	5
8	135	140	5
9	140	145	5
10	135	145	10
Média (DP)	139 (5,68)	145 (6,67)	6 (2,11)

O cálculo do CCI baseia-se em um modelo de análise da variância (de medidas repetidas). De forma simplificada, a ideia consiste em decompor a variação total observada (variância) nos seguintes componentes: a variabilidade devido às diferenças entre pacientes ou indivíduos observados (Variância entre indivíduos =  $s_E$ ); a variabilidade devido às diferenças entre diferentes aparelhos de medição ou entre medições repetidas realizadas pelo mesmo observador ou, como neste caso, entre diferentes observadores (Variância intraindivíduos =  $s_I$ ) e por último a variabilidade residual, que é a parte da variabilidade inexplicável e que atribuímos ao acaso (Variância residual =  $s_R$ ). O CCI é definido como o quociente entre a variabilidade entre os pacientes e a variabilidade total.

$$\text{CCI} = s_E / (s_E + s_I + s_R)$$

Quando um aparelho de medição for confiável ou quando a concordância entre dois observadores for alta, os resultados das medições repetidas sobre um mesmo indivíduo serão muito parecidos e, como consequência, a variância intraindivíduos será muito baixa e o valor do quociente anterior aumentará. Por ser uma proporção, os valores do CCI podem variar de 0 a 1 onde, 0 indica ausência de concordância e 1 concordância ou reprodutibilidade absoluta. Como norma geral, valores inferiores a 0,4 indicam baixa reprodutibilidade, entre 0,4 e 0,75 a reprodutibilidade é boa e valores superiores a 0,75 refletem um grau excelente de reprodutibilidade.

## **Avaliação Da Acurácia De Testes Diagnósticos**

Os testes diagnósticos incluem desde a anamnese e o exame físico até as técnicas mais modernas de diagnóstico por imagem (TC, RMN) e rastreamento, passando pelos exames de laboratório e outros tipos de procedimentos como o eletrocardiograma. A priori, pode parecer que os testes diagnósticos que medem parâmetros biológicos são mais seguros e confiáveis do que aqueles baseados na anamnese, já que estes últimos estariam sujeitos a uma avaliação mais subjetiva, observador dependente, enquanto que os primeiros teriam um caráter mais objetivo. Entretanto, isto é pura ilusão, já que ambos os tipos de exames estão sujeitos à interpretação. Com frequência os valores dos parâmetros biológicos são obtidos a partir de uma imagem ou um som que deverão ser interpretados pela pessoa encarregada de realizar o exame diagnóstico.

### **4.1 Validade de testes diagnósticos**

A validade de um teste diagnóstico é o grau em que mede o que realmente procura medir; ou seja, a capacidade de classificar corretamente os indivíduos em função de seu status de saúde-doença. Um teste válido é aquele que carece de erros sistemáticos ou vieses (ex.: um esfigmomanômetro que tenha um defeito de fabricação e por isso sempre mede uma pressão arterial 10 mmHg abaixo do valor real). Para estudar a validade comparamos os resultados do teste em estudo com o verdadeiro status do indivíduo, ou em sua ausência, com os resultados do melhor instrumento disponível de medição do fenômeno em estudo. Este exame de referência é denominado padrão ouro. Assim, um teste será válido quando classificar adequadamente tanto doentes como pessoas saudáveis ou quando seus resultados coincidam ou tenham um alto grau de concordância com os do padrão ouro.

*A validade de um teste diagnóstico é o grau em que classifica corretamente os indivíduos em função de seu status de saúde-doença.*

### **4.2 Análise da validade (risco de viés) de estudos seccionais para avaliação da acurácia de testes diagnósticos**

Os estudos seccionais são os mais frequentemente utilizados para análise da acurácia de um teste diagnóstico. Os critérios

de validade de um estudo seccional para avaliação de um teste diagnóstico segundo o Users Guides,<sup>10</sup> são os seguintes:

- ✓ Os pacientes incluídos no estudo estão dentro de um espectro de incerteza diagnóstica? Se não estão, o estudo pode apresentar o que se chama de "viés de espectro", quando os pacientes apresentam baixa ou alta probabilidade pré-teste de doença, ao invés de probabilidade pré-teste intermediária.
- ✓ O teste foi comparado a um critério de referência ou "padrão ouro" adequado? Se não foi, o estudo pode apresentar o que se chama de "viés de incorporação". Este viés acontece quando o teste em questão é incorporado ao critério de referência para estabelecer se o paciente tem ou não a condição sob investigação.
- ✓ Aqueles interpretando os resultados do teste em questão estavam cegos para os resultados do critério de referência? Se não estavam, o estudo pode apresentar o que se chama de "viés de revisão". Este viés acontece quando aquele que emite o laudo do exame conhece os resultados do critério de referência.
- ✓ O critério de referência foi aplicado da mesma forma em todos os pacientes independentemente dos resultados do teste sob investigação? Se não foi, o estudo pode apresentar o que se chama de "viés de verificação". Este viés acontece quando o critério de referência é aplicado no paciente dependendo do resultado do teste em questão, se positivo ou negativo.

---

### 4.3 Apresentação dos Resultados

---

Os resultados de um estudo seccional para avaliação do desempenho de um teste diagnóstico tem dois componentes principais, a sensibilidade (capacidade de diagnosticar corretamente as pessoas que padecem da doença em estudo) e a especificidade (capacidade de diagnosticar corretamente as pessoas que não padecem da doença em estudo). Vamos ilustrar estes conceitos com um exemplo. O serviço médico de uma empresa de transporte de substâncias perigosas está avaliando a implantação de um novo teste para o diagnóstico rápido em urina do possível consumo de cocaína entre seus motoristas. Para isso, realizaram um estudo de acurácia deste

teste diagnóstico sobre uma amostra de 1500 indivíduos, 45 dos quais eram consumidores de cocaína, com os resultados que são apresentados a seguir:

Tabela 4. Tabela 2x2 para a avaliação dos resultados de testes diagnósticos

		Consumo de cocaína		Total
		SIM	NÃO	
Resultado do teste	Positivo	23 (VP)	44 (FP)	67
	Negativo	22 (FN)	1411 (VN)	1433
Total		45	1455	1500

FN: Falso Negativo; FP: Falso Positivo; VN: Verdadeiro Negativo; VP: Verdadeiro Positivo.

Como se pode ver na tabela, podemos encontrar quatro possíveis situações: Verdadeiro Positivo (VP), quando o indivíduo padece da condição sob estudo (consumidor de cocaína) e o exame dá positivo; Falso Negativo (FN), quando o indivíduo padece da condição e a prova dá negativa; Verdadeiro Negativo (VN), quando o indivíduo não padece da condição e a prova dá negativa; e Falso Positivo (FP), quando o indivíduo não padece da condição e a prova dá positiva.

### 4.3.1 Sensibilidade (S)

A sensibilidade é a probabilidade de que um sujeito doente seja classificado como tal e é calculada como a percentagem dos sujeitos doentes que dão positivo quando é aplicado o teste diagnóstico. Quanto mais sensível o teste, menor a probabilidade de ele perder alguém doente, ou seja, menor a probabilidade de um falso negativo. Em nosso exemplo,  $S = 23 / 45 = 0,511$ ; ou seja, 51% dos consumidores são detectados pelo novo teste, mas 49% não o são.

### 4.3.2 Especificidade (E)

A especificidade é a probabilidade de que um sujeito não doente seja classificado como tal e é calculada como a percentagem de sujeitos não doentes, que dão negativo, ao lhes ser aplicado o exame diagnóstico. Quanto mais específico o teste menor a probabilidade dele classificar equivocadamente uma pessoa

saudável como doente, ou seja, menor a probabilidade de um falso positivo. Em nosso exemplo,  $E = 1411 / 1455 = 0,9698$ ; ou seja, 97% dos que não consomem cocaína serão verdadeiros negativos no novo teste de detecção.

### 4.3.3 Razão de Verossimilhança (RV)

A RV de um teste divide-se em duas medidas. A RV positiva (RV+) aplicada quando o resultado do teste for positivo e a RV negativa (RV-) aplicada quando o resultado do teste for negativo.

Matematicamente, A RV+ é a probabilidade de um resultado positivo nos doentes, dividido pela probabilidade de um resultado positivo nos não doentes: Sensibilidade / (1 - Especificidade). E a RV de um resultado negativo (RV-) é a probabilidade de um resultado negativo nos doentes dividido pela probabilidade de um resultado negativo nos não doentes: (1 - Sensibilidade) / Especificidade.

Tabela 5. Tabela 2x2 para o cálculo da razão de verossimilhança

	Doente	Não Doente	RV
Teste +	%VP	%FP	$\%VP / \%FP = S / (1 - E)$
Teste -	%FN	%VN	$\%FN / \%VN = (1 - S) / E$
	100%	100%	

E: Especificidade; FN: Falso Negativo; FP: Falso Positivo; S: Sensibilidade; VN: Verdadeiro Negativo; VP: Verdadeiro Positivo.

A RV+ indica quanto mais provável é um resultado positivo em um indivíduo doente do que em um não doente. Se o teste não tiver capacidade de discriminação alguma, a probabilidade do resultado positivo é a mesma em doentes e não doentes, e o valor do RV+ é igual a 1. Quanto maior do que 1 for o valor da RV+, maior a capacidade de discriminação do exame. Já a RV- indica quanto menos provável é um resultado negativo em um indivíduo doente do que em um não doente. Quanto menor do que 1 for o valor da RV-, maior a capacidade de discriminação do exame.

### 4.3.4 Curvas ROC

A determinação da sensibilidade e da especificidade de um teste diagnóstico requer que trabalhem com variáveis dicotômicas (doente-não doente e positivo-negativo). Quando trabalhamos

com variáveis quantitativas (ex.: mmHg, glicemia, colesterol, BNP, etc) é preciso definir um ponto de corte a partir do qual vamos considerar que estamos diante de uma situação patológica e, portanto, vamos outorgar um resultado positivo ao exame diagnóstico e considerar esse indivíduo como doente. Entretanto, em geral não existe um número mágico que permita separar com precisão e exatidão os indivíduos doentes dos não doentes e, ainda que este número existisse, os erros de medição (sistemáticos e aleatórios) provocariam uma atribuição errônea (falsos positivos ou negativos) em uma determinada porcentagem de casos. Como exemplo, suponhamos um ponto de corte para o diagnóstico de diabetes de 100 mg/dl de glicose no sangue.

Neste caso, a sensibilidade será muito alta: todos os diabéticos serão diagnosticados, já que é muito pouco provável que um diabético tenha um nível de glicose inferior a este limite. Entretanto, a especificidade será muito baixa: muitos indivíduos não diabéticos podem ter glicemias superiores a 100 mg/dl, e erroneamente classificados como diabéticos (falsos positivos). A situação oposta ocorreria no caso de escolhermos como ponto de corte 140 mg/dl.

Posto que não existe um ponto de corte de sensibilidade e especificidade de 100%, para selecionar o melhor ponto de corte possível foi desenvolvido um método denominado análise de curvas de desempenho diagnóstico (curva ROC). Tal como se observa na Figura 3, para representar uma curva ROC utilizamos um gráfico de coordenadas, com a sensibilidade no eixo Y e o complementar da especificidade no eixo X.

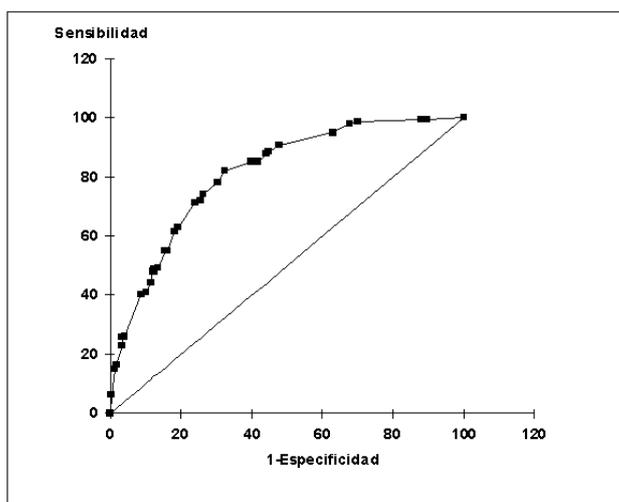


Figura 6. Curva ROC

Um teste diagnóstico inacurado é aquele cujo traçado coincide com a diagonal. A  $RV+$  e a  $RV-$  em tais pontos é igual a 1: é tão provável um resultado positivo em doentes quanto em não doentes e o mesmo pode-se dizer de um resultado negativo. Na área acima da diagonal, a  $RV+$  é maior de 1 e a  $RV-$  é menor de 1. A acurácia é tanto maior quanto maior é a área sob a curva; ou seja, quanto mais a curva se aproxima de um L invertido (ângulo superior esquerdo) melhor a capacidade de discriminação do teste.

A escolha do ponto de corte mais apropriado para decidir a partir de onde considerar o resultado do teste positivo depende das consequências decorrentes do diagnóstico de falsos negativos e positivos. Se as consequências de se diagnosticar um falso positivo resultarem em elevado dano emocional, físico ou financeiro para o paciente interessa uma maior especificidade e o ponto de corte se situará na parte baixa da curva. Pelo contrário, se os prejuízos forem grandes por se perder um diagnóstico, como o de condições graves e passíveis de cura, principalmente quando o diagnóstico é precoce, procuraremos máxima sensibilidade e colocaremos o ponto de corte na parte alta da curva. Nesta decisão também será preciso levar em consideração a prevalência da doença.<sup>8</sup>

Embora teoricamente a sensibilidade e a especificidade de um teste não sejam afetadas pela prevalência da doença, na prática diversas características dos pacientes tais como estágio e gravidade, podem afetar a sensibilidade e a especificidade do teste. Se a prevalência for baixa, por exemplo, na população preponderam os indivíduos não doentes ou em estágios muito iniciais da doença. Neste caso a sensibilidade do teste, conhecida a partir de estudos clínicos, tende a diminuir e a especificidade a aumentar. O ideal é escolher um teste diagnóstico com máxima sensibilidade, já que esta estará subestimada naquela população e poderá resultar em um número maior de falsos negativos. A chance de perder indivíduos em estágios iniciais da doença em uma população com baixa prevalência aumenta. Portanto, escolheremos um ponto de corte na parte alta da curva. Isso se aplica principalmente aos testes de *screening*. Pelo contrário, se a prevalência da doença for alta, a sensibilidade do teste tende a aumentar e a especificidade a diminuir. Neste caso será necessário um teste com elevada especificidade e situaremos o ponto de corte na parte baixa da curva, já que será necessário um teste com alta capacidade de afastar com certeza o diagnóstico quando o resultado for negativo.

## 4.4 Utilização dos testes diagnósticos

---

Um teste diagnóstico é tão mais eficaz quanto maior a mudança que provoca na probabilidade de padecer da doença, considerando-se que a magnitude desta mudança é que determina a possível mudança em nossa atitude terapêutica (tratar ou não) ou diagnóstica (confirmar ou não a presença da condição). Uma vez que um teste diagnóstico sempre supõe custos (econômicos, efeitos secundários, incômodos para o paciente e seu entorno, etc.) só deve ser aplicado se a informação que ele acrescenta puder mudar a atitude terapêutica diante do paciente. Por outro lado, na prática clínica, não existe nenhum exame com sensibilidade e/ou especificidade de 100%, e, portanto, o médico sempre se vê obrigado a tomar decisões com certo grau de incerteza. Um erro no diagnóstico é seguido por um erro no tratamento, seja por tratar um indivíduo não doente (FP), seja por deixar de administrar o tratamento a um indivíduo doente (FN). Por isso, na hora de decidir se aplicamos ou não um determinado exame diagnóstico, ou qual procedimento diagnóstico indicar entre vários outros, temos que levar em conta a gravidade das consequências decorrentes tanto de diagnosticar e tratar um indivíduo sem a doença como de não tratar um indivíduo com a doença.

Neste contexto e como regra geral, os exames de alta sensibilidade são especialmente indicados nas seguintes situações:

- O preço a ser pago por não detectar a doença é muito elevado. Isto ocorre diante da suspeita de processos perigosos, mas para os quais existe um tratamento eficaz, como a tuberculose, a sífilis ou o linfoma de Hodking.
- Para descartar hipóteses diagnósticas nas primeiras etapas de um estudo. Um exame muito sensível raramente é negativo em presença da doença (dá poucos FN), por isso um resultado negativo virtualmente elimina a possibilidade de que se padeça da doença.
- Para detectar os casos em que a probabilidade da doença é relativamente baixa. Este é o caso dos programas de rastreamento populacional, onde o propósito do exame é descobrir a doença em pessoas sem sintomas.

De modo similar, é conveniente utilizar exames de alta especificidade nas seguintes situações:

- Quando é necessário verificar a presença de uma doença (confirmação de um diagnóstico), que foi sugerida por outros dados. Isto se deve a que uma prova muito específica raramente é positiva em ausência da doença (dá poucos FP), por isso um resultado positivo virtualmente nos garante que o paciente padece da doença.
- Quando o diagnóstico e/ou tratamento incorretos (FP) têm graves consequências para o paciente, seja porque se trata de doenças estigmatizantes, como AIDS, e/ou de mau prognóstico, seja porque o tratamento tem importantes efeitos secundários sobre a saúde ou sobre o bem-estar econômico, emocional e/ou social do paciente, como no caso da quimioterapia para o câncer.

O ideal seria contar com um teste diagnóstico com sensibilidade e especificidade elevadas, mas isto nem sempre é possível. A opção por um teste ou outro deve se basear em uma análise de riscos e benefícios derivados de diagnosticar e tratar a uma porcentagem de indivíduos sem doença e vice-versa; levando em consideração que para uma sensibilidade e especificidade dadas, a probabilidade prévia da doença será a que irá determinar o maior ou menor número de FP e FN que vamos obter.

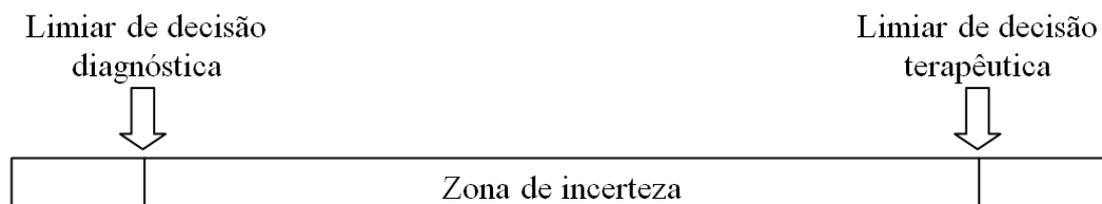
---

#### **4.4.1 Limiar de decisão**

---

Na prática clínica diária, os testes diagnósticos são utilizados como uma ferramenta de ajuda na decisão terapêutica. Quando aplicamos um teste diagnóstico em um paciente estamos tentando confirmar ou descartar uma suspeita diagnóstica ou, melhor dizendo, tentamos aumentar ou diminuir essa suspeita em um grau que nos permita interromper a investigação (limiar de decisão diagnóstica) ou iniciar um tratamento (limiar de decisão terapêutica). Por exemplo, ao avaliarmos um indivíduo com dor torácica frequentemente suspeitaremos da possibilidade de uma obstrução coronariana como causa dos sintomas. A suspeita deste diagnóstico será maior ou menor de acordo com as características do indivíduo. Se a suspeita for forte o suficiente para indicarmos uma coronariografia, que é um exame invasivo,

poderemos solicitar diretamente este exame para confirmação do diagnóstico e início do tratamento sem necessidade de solicitação de qualquer outro teste. Neste caso estaríamos acima do limiar terapêutico e não solicitaríamos outros testes antes da indicação do teste invasivo e início do tratamento. Se a suspeita for fraca o suficiente, poderemos simplesmente considerar outras hipóteses e interrompermos a investigação de doença coronariana antes mesmo de inicia-la. Neste caso estaríamos abaixo do limiar de decisão diagnóstica e a investigação seria interrompida. A zona existente entre o limiar de decisão diagnóstica e o limiar de decisão terapêutica é a zona de incerteza onde reside grande parte das situações clínicas e ela é definida pela probabilidade pré-teste da doença. O limiar de decisão é frequentemente definido de forma arbitrária, mas em algumas situações ele se baseia em estudo de avaliação econômica. No exemplo da dor torácica, a diretriz do NICE para dor torácica de início recente recomenda que o limiar diagnóstico seja em 10% e o limiar terapêutico em 90%.<sup>11</sup>



#### 4.4.2 Probabilidade pré-teste

A suspeita diagnóstica se baseia na estimativa do risco de padecer da doença apresentado por um sujeito que vai à consulta em função de seu histórico clínico, os sinais e sintomas que refere e o resultado de quantos testes diagnósticos tenha feito até esse momento relacionados com a possível doença em estudo. Este risco é o que se denomina probabilidade pré-teste, ou seja, antes da realização de novos testes diagnósticos que se esteja pensando aplicar a seguir. Se a probabilidade pré-teste for muito alta (como na HAS de grau 3), o médico terá a suficiente certeza sobre o risco de padecer da doença e, em consequência, decidirá aplicar o tratamento sem necessidade de realizar novos procedimentos diagnósticos. No caso contrário (HAS de grau 1), aplicará novos testes diagnósticos cujo resultado permitirá modificar a estimativa da probabilidade de sofrer da doença em grau suficiente para poder tomar uma decisão terapêutica ou interromper a investigação. Esta probabilidade posterior ao

resultado do teste é o que denominamos de probabilidade pós-teste, ou seja, probabilidade da doença após o resultado do novo teste diagnóstico que acabamos de aplicar.

A probabilidade pré-teste na prática clínica é frequentemente calculada de forma intuitiva ou arbitrária com base no conhecimento de base que o profissional de saúde tem a respeito daquela condição clínica. Entretanto, em muitas situações esta probabilidade pode ser estimada de forma objetiva, a partir de regras de predição clínica, como por exemplo, os critérios de Duke para doença coronariana.<sup>12</sup>

---

### 4.4.3 Probabilidade pós-teste

---

Podemos estimar a probabilidade pós-teste a partir da probabilidade pré-teste e da razão de verossimilhança. Se, por exemplo, estamos diante de um paciente com dor torácica, com uma probabilidade pré-teste de 50% estimada a partir dos critérios de Duke,<sup>12</sup> que considera o sexo, a idade, o tipo de dor torácica e os fatores de risco, precisaremos de um teste diagnóstico que nos permita aumentar ou diminuir essa probabilidade pré-teste até o nível do limiar de decisão terapêutica ou de decisão diagnóstica, respectivamente. Se solicitarmos um teste com uma RV+ de 10 e uma RV- de 0.1, poderemos estimar quanto será a probabilidade pós-teste da doença coronariana se o resultado do teste vier positivo ou negativo, respectivamente. Este cálculo pode ser feito manualmente, sabendo-se que a  $RV = \text{Chance pré-teste} \times \text{Chance pós-teste}$ . E que a chance pode ser calculada a partir da probabilidade ( $\text{Chance} = \frac{\text{Probabilidade}}{100 - \text{Probabilidade}}$ ). Ou podemos obter esta estimativa rapidamente a partir do nomograma de Fagan,<sup>13</sup> que correlaciona probabilidade pré-teste, RV e probabilidade pós-teste. Neste caso, portanto, se o resultado do teste vier positivo, a probabilidade pós-teste seria de aproximadamente 95% e, portanto, acima do limiar terapêutico. A decisão provavelmente seria iniciar o tratamento. Se vier negativo, a probabilidade pós-teste seria de aproximadamente 5% e, portanto, abaixo do limiar diagnóstico. A decisão provavelmente seria interromper a investigação e buscar outras causas para os sintomas apresentados pelo paciente.

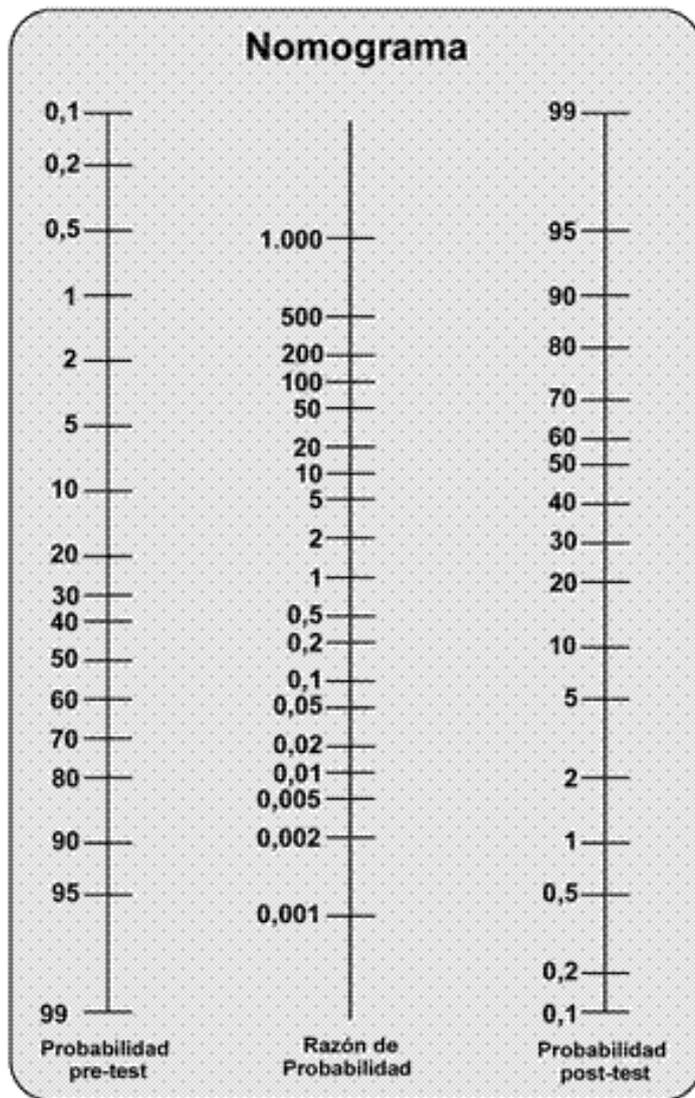


Figura 9. Nomograma de Fagan.

## 5. Ensaio Clínico Randomizado

Um ensaio clínico é uma experiência planejada na qual, de maneira prospectiva, comparam-se duas ou mais intervenções (ou intervenção e placebo) preventivas, curativas ou reabilitadoras, que são atribuídas de forma individualizada e aleatória a um grupo de pacientes para determinar sua eficácia.<sup>14</sup> Além disso, tanto a seleção dos sujeitos como os períodos de tratamento e acompanhamento têm que ter lugar simultaneamente em todos os grupos. Uma vez que os participantes de um ensaio clínico são, por definição, sujeitos doentes (pacientes), as intervenções

*O ensaio clínico controlado e randomizado constitui a ferramenta padrão para avaliar a eficácia das intervenções sejam elas terapêuticas, diagnósticas ou prognósticas.*

preventivas às quais nos referimos são secundárias, dirigidas a evitar as sequelas da doença, incluindo o óbito. Por exemplo, uma dieta rica em fibra para prevenir a recorrência de pólipos adenomatosos de cólon ou o uso de estatinas para reduzir a mortalidade pós-enfarte de miocárdio. Não serão objeto de nosso estudo neste tema os ensaios cuja finalidade consiste em avaliar a eficácia de intervenções preventivas em sujeitos saudáveis (ensaios de prevenção primária), denominados ensaios de campo, quando a intervenção for aplicada em nível individual ou ensaios de intervenção comunitária, quando a intervenção for aplicada em nível grupal (escola, empresa, família, etc.).

Os ensaios clínicos avaliam as intervenções em condições muito diferentes às de sua aplicação habitual, uma vez que os pacientes são submetidos a um estreito acompanhamento e vigilância médica e não costumam ser representativos de todo o espectro de possíveis sujeitos que padecem da doença em questão. Portanto, um ensaio está avaliando o efeito da intervenção em condições ideais (eficácia). Entretanto, não pode avaliar como se comportará essa mesma intervenção em condições reais de uso (efetividade) nem sua relação custo-benefício (eficiência), aspectos que são considerados no marco da investigação de resultados. Os pacientes selecionados para realizar um ensaio representam habitualmente o grupo mais acessível à equipe pesquisadora, com menos patologias associadas e com maiores probabilidades de cumprir o protocolo do estudo. Além disso, os pacientes que aceitam participar são em geral sujeitos motivados e mais conscientizados com seu estado de saúde, por isso é provável que o grau de adesão a uma terapia concreta ou a uma modificação dietética seja maior do que aquele que seria observado nos demais pacientes. Por outro lado, a estreita vigilância a que estão submetidos estes pacientes enquanto dura o estudo faz que os descumprimentos ou abandonos do tratamento sejam menos prováveis do que em condições habituais.

Os ensaios clínicos randomizados são estudos longitudinais de grupos paralelos. Neste modelo, cada paciente é alocado de forma aleatória em um grupo do estudo e depois de um período de acompanhamento variável, realiza-se a medição da resposta (Figura 7).

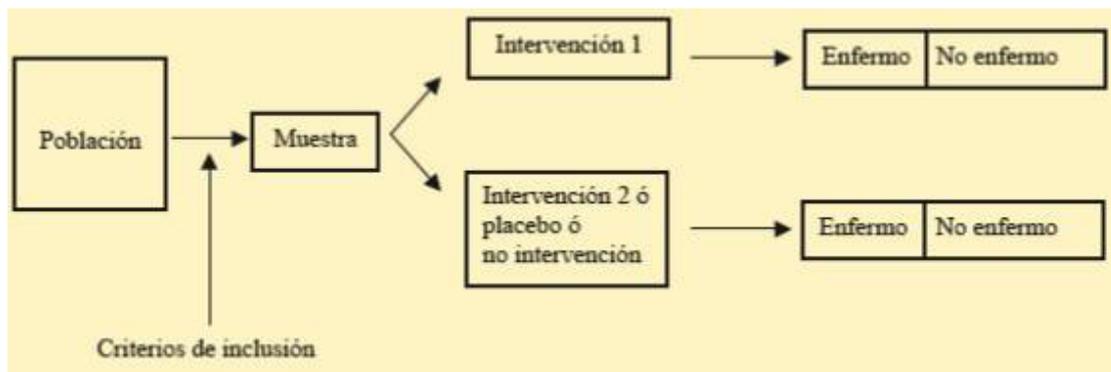


Figura 7. Ensaio clínico randomizado

Às vezes este desenho pode considerar um período de pré-inclusão anterior à randomização, que tem como objetivo diminuir as taxas de não cumprimento da intervenção ou de abandono do estudo e, por vezes, estabilizar os valores das variáveis através das quais a resposta será medida. Durante o período de pré-inclusão, todos os indivíduos são submetidos à mesma intervenção, habitualmente a do grupo controle, para identificação e exclusão de sujeitos não aderentes. Mensink e cols. realizaram um ensaio clínico randomizado para comparar o efeito dos ácidos gordurosos monoinsaturados e poliinsaturados sobre o perfil lipídico plasmático. Para isso, estabeleceram um período de pré-inclusão durante o qual os sujeitos receberam uma dieta rica em ácidos gordurosos saturados, permitindo avaliar o grau de cumprimento da dieta e estabilizar os lipídios sanguíneos. Uma vez finalizado este período, os sujeitos aderentes foram incluídos e alocados de forma aleatória a uma dieta rica em ácidos gordurosos mono ou poliinsaturados, respectivamente.<sup>15</sup>

## 5.1 Análise da validade (risco de viés) de ensaios clínicos randomizados

### 5.1.1 Randomização

A randomização consiste na alocação das unidades experimentais (sujeitos participantes) a duas ou mais intervenções utilizando um mecanismo randômico, de forma que nem o pesquisador nem o sujeito investigado conheçam a intervenção que vai ser alocada no momento da inclusão deste último no estudo, uma vez obtido seu consentimento informado para participar do estudo.<sup>16</sup> Embora um estudo randomizado seja necessariamente controlado, o

grupo controle poderá receber uma intervenção ativa, que habitualmente será aquela utilizada em condições normais para a condição sob estudo, inativa e inócua (placebo) ou simplesmente não receber intervenção alguma. A randomização elimina o viés do pesquisador na alocação dos sujeitos, garante a validade dos níveis de significação das provas estatísticas que serão aplicadas na análise dos dados e tende a produzir grupos comparáveis em relação a características conhecidas e desconhecidas que poderiam afetar o resultado de interesse (fatores prognósticos). Frente a outros métodos de controle de fatores de confusão utilizados nas fases de desenho (amostragem estratificada, pareamento) ou análise (ajuste de taxas, regressão múltipla), a randomização apresenta a grande vantagem de controlar tanto fatores conhecidos como desconhecidos.

*A randomização garante que os grupos sejam semelhantes no início do estudo em relação aos fatores prognósticos conhecidos e desconhecidos.*

A eficácia da randomização depende do tamanho da amostra do estudo. Quando este não é suficientemente grande - a título de orientação estabeleceremos o ponto de corte em 100 sujeitos, embora isto também dependa do número de grupos analisados - não se pode garantir o equilíbrio no tamanho dos grupos nem a distribuição homogênea dos fatores de confusão. Por isso, na hora de apresentar os resultados de um ensaio, o primeiro passo deve consistir em mostrar uma tabela com a distribuição dos fatores prognósticos conhecidos nos grupos de intervenção e controle. Uma distribuição homogênea do número de sujeitos e dos fatores conhecidos em ambos os grupos constitui uma prova de que a randomização foi eficaz e nos indica que, com toda probabilidade, os grupos também serão comparáveis em relação a possíveis fatores prognósticos que, por serem desconhecidos, não foi possível medir, permitindo atribuir com segurança o efeito observado à intervenção sob estudo.

Um bom método de randomização deve cumprir os seguintes requisitos para garantir a qualidade do mesmo e a eficácia do processo:

- a. A randomização tem que ser cega, ou seja, desconhecida pelo pesquisador até o momento em que o sujeito dá sua autorização para participar do estudo.
- b. A sequência de alocação deve ser imprevisível, ou seja, o pesquisador deve ser incapaz de adivinhar com antecedência a intervenção a qual vai ser designado o próximo sujeito.

- c. O processo deve ser baseado em propriedades matemáticas conhecidas e ser reprodutível.

---

### 5.1.2 Cegamento

---

O cegamento em um ensaio clínico, é o conjunto de medidas ou precauções que são tomadas com a finalidade de que o sujeito de pesquisa, aqueles que intervêm no mesmo (médico, avaliador, etc.), aqueles que coletam dados e aqueles que os interpretam desconheçam o grupo ao qual o paciente foi alocado, uma vez que este conhecimento poderia introduzir vieses no desenvolvimento, análise e avaliação de um ensaio clínico. O objetivo do cegamento é garantir que os grupos permanecem semelhantes ao longo do estudo em relação às características prognósticas conhecidas e desconhecidas e, dessa forma, permaneçam comparáveis até o final do estudo. A apresentação das intervenções medicamentosas, por exemplo, deverá ser idêntica, diferenciando-se unicamente pela substância ativa que contêm. Neste caso, é preciso considerar diversos aspectos que precisam ser idênticos em os grupos: via e pauta de administração, forma, tamanho, cor, sabor, textura, etc. O cegamento do responsável por avaliar a resposta (paciente, médico ou um terceiro) à intervenção é particularmente importante quando a variável resposta é subjetiva (ex.: avaliação da intensidade da dor, análise de uma radiografia, etc.), já que o conhecimento da intervenção pode afetar a percepção sobre a variável resposta.

Mesmo os resultados mais objetivos, como óbito, podem sofrer influência deste tipo de viés. Sabe-se que a administração de uma intervenção produz um efeito terapêutico sobre o indivíduo, denominado efeito placebo, independentemente do princípio ativo que contenha. Ou seja, inclusive a administração de um princípio inativo inócuo produz um efeito mensurável. O simples fato de ser observado, de participar de um estudo, produz uma modificação no estado clínico dos pacientes, denominada "efeito Hawthorne". A isto é preciso acrescentar o poder de sugestão derivado do fato de receber uma intervenção, mesmo que esta seja farmacologicamente inativa, e o da presença de um médico, ou pesquisador que a aplica. O sugestionamento também interfere no comportamento do profissional de saúde. O simples fato de tanto o profissional de saúde quanto o paciente, saber o grupo de alocação, pode produzir efeitos no comportamento de ambos

e conseqüentemente nos desfechos clínicos que estiverem sendo mensurados no estudo, mesmo que eles sejam objetivos.

Infelizmente, nem sempre é possível aplicar o cegamento, por exemplo, no caso em que a intervenção é uma dieta ou uma intervenção cirúrgica, ou quando um tratamento pode apresentar efeitos secundários bem conhecidos que podem permitir ao paciente realizar a identificação do mesmo. Por ex.: a rifampicina tingem de vermelho a urina. Mas mesmo quando não é possível mascarar os cuidadores e os pacientes para o grupo alocado, é possível cegar os avaliadores dos desfechos, como por exemplo, nos comitês independentes que fazem a adjudicação dos desfechos clínicos.

Além disso, nunca os grupos nunca devem ser mascarados se os pacientes venham a ser submetidos a algum tipo de risco. Em todas estas situações o pesquisador deve descrever não só o método de cegamento e as violações do mesmo, mas também os aspectos comentados anteriormente, em relação à impossibilidade de aplicá-lo ou os possíveis riscos de que não seja eficaz.

---

### **5.1.3 Perda de seguimento**

---

A perda de seguimento dos pacientes incluídos no ensaio clínico também pode comprometer significativamente a validade do estudo. Recomenda-se que diante de uma perda de seguimento no estudo considere-se o pior e o melhor cenários para aquelas perdas, ou seja, estima-se os extremos do resultado considerando que todos os pacientes perdidos no grupo intervenção tiveram o evento, enquanto que no grupo controle não (pior cenário), e vice-versa (melhor cenário). Se a conclusão do estudo muda com os extremos do resultado, então a perda é considerada significativa e o risco de viés alto. Entretanto, apesar disso, a decisão final quanto ao risco de viés nesta situação é subjetiva e também depende do julgamento clínico já que o pior e o melhor cenários na vida real são improváveis.<sup>10</sup>

---

### **5.1.4 Intenção de tratar**

---

Em geral, a análise principal de um ensaio clínico deve ser baseada no princípio de intenção de tratar, que inclui cada

paciente em seu grupo de randomização independentemente das mudanças que possam ocorrer *a posteriori*: descumprimento da intervenção, mudança de tratamento, etc.. Embora este tipo de análise tenha uma maior tendência a obter resultados nulos, aumentando as dificuldades para demonstrar a eficácia de qualquer nova intervenção, existem duas poderosas razões para levá-lo a cabo. Por um lado, com este tipo de análise se pretende preservar o que foi obtido com a randomização, que os grupos sejam comparáveis, o que constitui o fundamento para poder obter conclusões válidas (não enviesadas) sobre a eficácia de uma intervenção. Por outro lado, trata-se de avaliar o efeito da intervenção em condições as mais parecidas possíveis com as da vida real (efetividade) e não só tendo em conta as condições ideais (eficácia) proporcionadas pelo subgrupo dos pacientes que seguiram o protocolo de forma estrita.

*A análise por intenção de tratar proporciona uma medida não distorcida da eficácia da intervenção em condições reais.*

---

### **5.1.5 Interrupção precoce por benefício**

---

A interrupção precoce de um ensaio clínico acarreta no risco de se superestimar o resultado do estudo, principalmente quando o número de eventos relatados for inferior a 200.<sup>17</sup>

---

## **5.2 Apresentação dos Resultados**

---

Tradicionalmente, a medição do efeito de uma intervenção poderia ser realizada por meio de medidas absolutas, como a redução do risco absoluto, ou relativas, como a redução do risco relativo. De forma mais recente surgiram outras medidas de carácter complementar, que enfatizam a finalidade de facilitar a tarefa do clínico no momento de orientar o tratamento de seus pacientes. Estas medidas de carácter mais pragmático incluem a porcentagem de pacientes livres da condição sob estudo e o número necessário de pacientes a tratar (NNT). O NNT constitui o número de pacientes que teríamos que tratar para evitar que se produza um caso com a doença ou condição estudada, e é calculado como o inverso da redução do risco atribuível.

A apresentação dos resultados mediante um único indicador, como a redução percentual do risco relativo, pode produzir uma impressão ampliada sobre o efeito do tratamento, por isso atualmente recomenda-se que, ao apresentar os resultados

de um ensaio clínico, seja utilizada a redução percentual do risco relativo, acompanhada de alguma magnitude que meça a redução do risco em termos absolutos, particularmente o NNT. Para ilustrar esta última afirmação, nos basearemos no exemplo descrito por Cook e Sacket, a partir de uma revisão de ensaios clínicos sobre tratamentos farmacológicos para a hipertensão arterial.<sup>18</sup> Os estudos foram divididos em dois grupos mutuamente exclusivos: aqueles nos quais todos os pacientes tinham, no início do ensaio, uma pressão diastólica menor do que 110 mmHg (hipertensão ligeira) e aqueles nos quais todos os pacientes tinham, no início do estudo, uma pressão diastólica menor do que 115 mmHg (hipertensão moderada). Como se pode observar na tabela 5, a redução percentual do risco relativo de sofrer um AVC durante os 5 anos seguintes com tratamento farmacológico foi similar nos dois grupos, de 40%, o que poderia levar a crer que a estratégia terapêutica deveria ser igual em ambos os grupos. Entretanto, a redução absoluta do risco foi muito menor nos sujeitos com hipertensão leve, uma vez que este grupo de pacientes tinha uma taxa de AVC em ausência de tratamento muito mais baixa do que o grupo com hipertensão moderada. Traduzido em termos de NNT, enquanto em um caso era necessário tratar 13 pacientes para evitar um AVC no outro era preciso tratar 167 pacientes para obter o mesmo resultado. Como consequência, parece que a estratégia terapêutica deveria ser diferente nesses dois grupos de pacientes.

Hipertensión	ACV en 5 años		Reducción del RR (Pc - Pt) / Pt	Reducción del RA Pc - Pt	NNT 1 / (Pc - Pt)
	Grupo control	Grupo tratamiento			
<b>Moderada</b>					
Tasa de ACV (P)	0,20	0,12	0,40	0,08	13
N	16.778	16.898			
<b>Ligeira</b>					
Tasa de ACV (P)	0,015	0,009	0,40	0,006	167
N	15.165	15.238			

Tabla 5. Cálculo da redução do risco e do NNT para evitar um caso de AVC em pacientes com hipertensão. Retirado de BMJ 1995; 310: 452-454.<sup>18</sup>

Por fim, a importância de uma adequada apresentação dos resultados dos ensaios clínicos levou à elaboração da "Declaração CONSORT", um relatório com recomendações para melhorar a qualidade da apresentação dos resultados de ensaios clínicos, elaborado por pesquisadores, estatísticos, epidemiologistas e editores biomédicos, e apoiado por numerosas revistas biomédicas e grupos editoriais. A declaração abrange os aspectos (22 itens) que têm que ser considerados e a forma de

apresentá-los nas diferentes seções (título e resumo, introdução, métodos, resultados e discussão) do relatório sobre um ensaio clínico de dois grupos paralelos. A declaração, juntamente com as extensões da mesma para os diferentes tipos de ensaios clínicos (em função de dados, intervenções ou desenhos alternativos) pode ser consultada no seguinte endereço: <http://www.consort-statement.org/consort-statement>.

---

## 5.3 Utilização de ensaios clínicos randomizados

---

Os ensaios clínicos randomizados são comumente utilizados para avaliar a eficácia de intervenções terapêuticas, e cada vez mais frequentemente também têm sido utilizados para avaliar a eficácia de intervenções diagnósticas e prognósticas na redução de desfechos clínicos.

---

### 5.3.1 Aplicabilidade dos resultados

---

A seleção dos sujeitos que participarão do estudo deve ser realizada de forma cuidadosa. Para isso, será preciso começar por definir de forma precisa o perfil dos sujeitos suscetíveis de participar do estudo, por ex.: sujeitos de ambos os sexos, de 25 a 40 anos de idade, com sobrepeso (IMC entre 25 e 30) e residentes em Madri. A seguir devem ser definidos os critérios de exclusão, em função de aspectos que possam comprometer a participação no estudo (ex.: doenças incapacitantes física ou mentalmente) ou que possam interagir com a intervenção sob estudo e afetar a variável de interesse. Em todos os casos, precisamos tentar ser o menos restritivos possível na hora de estabelecer os critérios de inclusão e exclusão, com a finalidade de evitar que tenhamos problemas para recrutar um número apropriado de sujeitos.

---

### 5.3.2 Rastreamento

---

O rastreamento consiste no uso de testes diagnósticos para a detecção precoce de doenças em pacientes assintomáticos; ou seja, em pacientes que ainda não desenvolveram a fase clínica da doença. O critério básico para aplicar um teste diagnóstico neste caso é o mesmo de quando nos encontramos diante de um

paciente sintomático: só deve ser aplicado se a informação que ele acrescenta pode mudar a atitude terapêutica diante do paciente. Além disso, para se realizar um "programa de rastreamento" é preciso que seja cumprida uma série de condições que afetam a doença, o teste diagnóstico e o próprio programa. Existem duas estratégias principais de rastreamento:

- a. Rastreamento populacional: Neste tipo de programa é convidada toda a população suscetível de submeter-se ao teste diagnóstico em questão. Ex.: rastreamento de câncer de mama através de mamografia em mulheres maiores de 50 anos.
- b. Busca ativa de casos: Aproveitando a visita ao médico por outras causas, convida-se o paciente a submeter-se a determinados testes diagnósticos destinados a detectar patologias não suspeitadas pela clínica. Ex.: busca ativa de indivíduos com hipertensão arterial ou diabetes não diagnosticadas.

A melhor evidência para apoiar o rastreamento é fornecida pelos ensaios clínicos randomizados nos quais a intervenção é o rastreamento e as variáveis de resultado são a diminuição da morbimortalidade atribuída à condição que pretende rastrear. No caso do câncer de mama, diversos ensaios clínicos mostraram uma redução de 20 a 30% da mortalidade por câncer de mama em mulheres de 50 a 64 anos de idade. Não obstante, dois destes ensaios, selecionados em função de critérios de qualidade, observaram um efeito nulo da mamografia. No que diz respeito às evidências sobre a efetividade do rastreamento de câncer pela mamografia nos países onde foram aplicados estes programas, estas são em alguns casos promissoras, mas ainda muito limitadas, embora se reconheça que os programas de rastreamento tiveram um impacto positivo na organização global da assistência ao câncer de mama. Por outro lado, a mamografia implica certos incômodos e riscos por radiação. Além disso, apenas entre 5 e 20% das mamografias positivas são confirmadas como câncer de mama, por isso entre 80 e 95% das mulheres com mamografia positiva são submetidas a testes diagnósticos adicionais (mamografias adicionais, ecografia, citologia, punção, biópsia) que as expõem a novos riscos, e finalmente não conduzem ao diagnóstico de câncer de mama. Também foi detectado super diagnóstico e tratamento com

cirurgia e/ou radioterapia desnecessário.<sup>5</sup> Em resumo, o conjunto das evidências científicas é com frequência incapaz de dar uma resposta consistente a todas as interrogações apresentadas quando se avalia a introdução de um programa de rastreamento populacional e, no caso do rastreamento do câncer de mama com mamografia, amplamente aceito em mulheres > 50 anos e implantado em muitos países de nosso entorno, a controvérsia continua.<sup>19</sup>

O objetivo de um "programa de rastreamento" consiste em reduzir a morbimortalidade e aumentar a qualidade de vida dos pacientes através de um diagnóstico e tratamento precoces da doença. Para que um programa de rastreamento possa cumprir seu objetivo, a doença deve ser suscetível de ser abordada de forma efetiva, para o qual o programa deve cumprir uma série de requisitos. Em primeiro lugar, é preciso que exista uma fase pré-clínica durante a qual a doença é detectável. Em segundo lugar, é necessário que exista um tratamento eficaz para a mesma. E em terceiro lugar, é imprescindível que exista um ponto crítico no curso da doença, situado entre o diagnóstico precoce possível e o diagnóstico clínico habitual; de modo que o tratamento antes de alcançar este ponto crítico afete de forma favorável o curso do processo, por ser mais fácil de aplicar e/ou mais eficaz do que durante a fase clínica da doença.

Além do mais, quanto maiores forem a gravidade e a prevalência da doença em fase pré-clínica detectável, maior será o desempenho do exame diagnóstico e mais eficiente o programa de rastreamento. No que diz respeito ao exame diagnóstico deve ser válido, confiável, fácil de aplicar, barato, com poucos efeitos secundários, aceitável pela população à qual é dirigido e com uma sensibilidade muito alta (próxima ou igual a 100%), de forma que permita detectar praticamente todos os casos existentes em fase pré-clínica. Uma estratégia utilizada com frequência nos programas de rastreamento consiste na combinação de testes diagnósticos. Combinando dois exames em série, primeiro aplicaremos um exame muito sensível que nos permitirá detectar quase todos os casos presentes na população, embora às custas de muitos falsos positivos. Os que obtiverem um resultado negativo com este primeiro exame serão declarados saudáveis. A seguir aplicaremos um exame muito específico com o qual conseguiremos descartar os falsos positivos, que agora darão majoritariamente um resultado negativo, e serão declarados saudáveis, restando os verdadeiros doentes aos quais aplicaremos o tratamento, aqueles

*Os "programas de rastreamento" são desenhados, para tratar de reduzir a morbimortalidade da população, mediante o diagnóstico e tratamento precoces, daquelas doenças que cumprirem os requisitos de aplicação.*

que também deram positivo neste segundo exame. Por último, o próprio programa de rastreamento também precisa cumprir uma série de requisitos: deve ser factível de todos os pontos de vista, tanto economicamente (custo-benefício) como por questões de acessibilidade da população alvo, ou quaisquer outros aspectos que possam pôr em risco seu adequado desenvolvimento. É particularmente importante que exista disposição e recursos para atender a todos os pacientes que sejam diagnosticados por meio do programa. Entre os programas de rastreamento populacional que atualmente são aplicados em nosso entorno vale destacar o programa de detecção da fenilcetonúria e do câncer de cervix uterino.

## Referências bibliográficas

1. Spitzer WO. *Clinical epidemiology. J Chronic Dis.* 1986;39(6):411-415.
2. Fletcher RH, Fletcher SW, Wagner EH. *Epidemiología Clínica.* 2nd ed. Madri: Consulta SA; 1989.
3. Depetris AR. *La tercera revolución epidemiológica y la promoción de salud. Conferencia internacional de promoción de salud;* 1992; Santa Fe de Bogotá.
4. Feinstein AR. *Ciencia, medicina clínica y espectro de enfermedades.* In: Cecil-Loeb, ed. *Tratado de Medicina Interna.* 14 ed. Madri: Importécnica; 1978.
5. Imaz I, González J. *Epidemiología clínica.* In: Royo MA, Damián J, eds. *Método Epidemiológico.* Madri: ENS-ISCIH; 2009:169-186.
6. Sackett DL, Richardson WS, Rosenberg W, Haynes RB. *Medicina basada en la evidencia. Cómo ejercer y enseñar la MBE.* Madri: ADS Printing SA; 1997.
7. Sackett DL, Haynes RB, Tugweell P. *Epidemiología clínica. Una ciencia básica para la medicina clínica.* Madrid: Díaz de Santos; 1989.
8. Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology.* 3rd ed. Baltimore, Maryland, USA: Williams & Wilkins; 1996.
9. *The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation, and Treatment of High Blood Pressure.* 04-5230 ed. Bethesda (MD): U.S. Department of Health and Human Services. National Institute of Health. National Heart, Lung, and Blood Institute; 2004.
10. *Users' Guides to the Medical Literature: A Manual for Evidence-Based Clinical Practice.* 3rd ed. USA: McGraw-Hill Education; 2015.

11. Skinner JS, Smeeth L, Kendall JM, Adams PC, Timmis A, Chest Pain Guideline Development G. NICE guidance. Chest pain of recent onset: assessment and diagnosis of recent onset chest pain or discomfort of suspected cardiac origin. *Heart*. 2010;96(12):974-978.
12. Gibbons RJ, Balady GJ, Bricker JT, et al. ACC/AHA 2002 guideline update for exercise testing: summary article. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Committee to Update the 1997 Exercise Testing Guidelines). *J Am Coll Cardiol*. 2002;40(8):1531-1540.
13. Caraguel CG, Vanderstichel R. The two-step Fagan's nomogram: ad hoc interpretation of a diagnostic test result without calculation. *Evid Based Med*. 2013;18(4):125-128.
14. Pocock SJ. *Clinical Trials: A practical approach*. West Sussex, England: John Wiley & Sons Ltd; 2013.
15. Mensink RP, Katan MB. Effect of a diet enriched with mono-unsaturated or polyunsaturated fatty acids on levels of low-density and high-density lipoprotein cholesterol in healthy women and men. *N Engl J Med*. 1989;321(7):436-441.
16. Guallar E, Royo-Bordonada MA. Directrices del "Comitee propietary medicinal products": Estadística. II Reunião de atualização de ensaios clínicos com medicamentos. Madri: Ergón; 1996.
17. Bassler D, Briel M, Montori VM, et al. Stopping randomized trials early for benefit and estimation of treatment effects: systematic review and meta-regression analysis. *JAMA*. 2010;303(12):1180-1187.
18. Cook RJ, Sackett DL. The number needed to treat: a clinically useful measure of treatment effect. *BMJ*. 1995;310(6977):452-454.
19. Baines CJ. Rational and irrational issues in breast cancer screening. *Cancers (Basel)*. 2011;3(1):252-266.