Testes diagnósticos no contexto da avaliação de tecnologias em saúde: abordagens, métodos e interpretação

Diagnostic tests in the context of the health technologies assessment: approaches, methods, and interpretation

Altacílio A. Nunes[1]; Edson Zangiacomi Martinez[1]; Lauro Wichert Ana[1]; Antonio Pazin-Filho[1]; Eduardo Barbosa Coelho[1]; Luane Marques de Mello[1]

RESUMO

As Avaliações de Tecnologias em Saúde (ATS) no mundo todo, predominantemente tem sido focadas em medicamentos, dispositivos médicos terapêuticos e procedimentos, sobretudo, os cirúrgicos. Apesar de sua inquestionável importância na história natural de um grande número de doenças e do impacto econômico associado ao seu uso, os testes e exames diagnósticos (TED), considerando-se suas qualidades e consequentemente a acurácia dos mesmos, tem sido pouco avaliados no contexto da ATS. Há nítida escassez de estudos que avaliam os TED tanto do ponto de vista clínico e de segurança para o paciente, quanto do econômico. O propósito desse artigo é apresentar e discutir os conceitos inerentes ao uso dos TED, as abordagens para seu emprego, as metodologias de avaliação de suas propriedades e acurácia, bem como a interpretação de resultados dos TED, sejam eles realizados individualmente, ou sob a forma de síntese de estudos de acurácia. Espera-se que esse texto possa contribuir para melhor compreensão das especificidades encontradas nos estudos dos TED e estimular sua inclusão nas ATS.

Palavras-chave: Estudos de Acurácia**.** Teste de Diagnóstico. Avaliação de Tecnologias de Saúde. Revisão Sistemática. Rastreamento. Tomada de Decisões.

1. Docente. Faculdade de Medicina de Ribeirão Preto - Universidade de São Paulo, Brasil. Núcleo de Avaliação de Tecnologias em Saúde do Hospital das Clínicas da FMRP-USP.

Correspondência:
Prof. Altacílio A. Nunes
Departamento de Medicina Social – FMRP/USP
Av. Bandeirantes, 3900 – Monte Alegre
CEP: 14.049-900 - Ribeirão Preto/SPBrasil
altacilio@fmrp.usp.br

Artigo recebido em 16/06/2014
Aprovado para publicação em 14/08/2014

 DOI: http://dx.doi.org/10.11606/issn.2176-7262.v48i1p8-18

ABSTRACT

The Health Technology Assessments (HTA) worldwide has been predominantly focused on drugs, medical devices and therapeutic procedures, above all, the surgeries. Despite its unquestionable importance in the natural history of a large number of diseases and of the economic impact associated with its use, the diagnostic exam and tests (DET), considering their qualities and hence the accuracy thereof, has been not evaluated in the context of the HTA. There is a clear shortage of studies that evaluate the DET, both clinician and patient safety, and economical. The purpose of this article is to present and discuss the concepts inherent in the use of DET, the approaches to your employ, the methodologies of evaluation of their properties and accuracy, as well as the interpretation of results DET studies, whether performed individually or in the form of synthesis of studies of accuracy. It is hoped that this text may contribute to better understanding of the specifics found in studies of DET and encourage their inclusion in the HTA.

Key-Words: Accuracy Studies**.** Diagnostic Tests. Health Technology Assessment. Review, Systematic. Screening. Decision Making.

Introdução

Quanto aos seus propósitos, as tecnologias da saúde podem ser classificadas como: preventivas, de triagem, de diagnóstico, terapêuticas e de reabilitação.[1] No entanto**,** apesar das três primeiras se sustentarem basicamente por exames ou testes diagnósticos (ETD), e as duas últimas também delas dependerem em alguma fase da doença, tradicionalmente a agenda das Avaliações de Tecnologias em Saúde (ATS) em todo o mundo tem se pautado mais com as tecnologias ligadas a tratamentos (medicamentos) e outros procedimentos terapêuticos (sobretudo cirurgias)[2-6], deixando para segundo plano as avaliações de testes e equipamentos diagnósticos. Seguindo essa tendência, a capacitação de recursos humanos em ATS, sobretudo no Brasil tem se voltado exclusivamente para avaliação de fármacos e procedimentos.[7] Independente das razões desse cenário, essa lacuna necessita ser preenchida, pois, os ETD estão presentes nas atividades de prevenção primária como, por exemplo, no *screening* neonatal, no rastreamento de câncer de colo uterino através da citologia oncótica, etc., bem como na prevenção secundária, considerando-se que um diagnóstico confiável é parte crucial nos principais desfechos de qualquer doença específica. O diagnóstico correto e precoce, sobretudo das enfermidades graves, interfere de forma decisiva na história natural da doença, determinando condutas igualmente adequadas e consequentemente, maiores chances de desfechos favoráveis com menores custos financeiros e sociais. Desse modo, de mesma importância que avaliar a eficácia, efetividade, segurança e custos de fármacos e procedimentos médicos, é estudar as propriedades e acurácia dos ETD.

Sempre que empregados de maneira adequada, com base em evidências, os ETD são de inquestionável valor para seus usuários diretos (médicos e pacientes), trazendo grande contribuição, não raramente decisiva, na conduta médica. Desde confirmação diagnóstica de doenças agudas e graves como no caso das meningites bacterianas, até condições silenciosas e crônicas como diabetes, hipertensão arterial e HIV/SIDA, os testes diagnósticos desempenham papel de grande relevância, não sendo menos importantes nos casos de pacientes que já estão em tratamento, pela avaliação da gravidade da doença, seleção do tratamento e dos resultados da terapia instituída (estabilização, progressão, melhora ou recidiva do quadro), estimativa de prognóstico, etc. Apesar da relevância dos ETD na prática clínica, ao se lançar mão de um deles como auxílio, muitas vezes indispensável, o solicitante deve sempre contrabalançar os benefícios conhecidos ou potenciais com os riscos diretos e indiretos associados. [8-11] É consenso que além de avaliar as consequências benéficas ou não das ETD, os custos totais envolvidos devem sempre ser considerados, possibilitando a realização de avaliações econômicas como estimativas e cálculos de custo-efetividade, custo-utilidade e impacto orçamentário.[11] Assim o objetivo desse artigo é apresentar os principais conceitos relacionados à avaliação das propriedades dos testes e exames diagnósticos visando contribuir para melhor compreensão das especificidades encontradas nos estudos dos TED e estimular sua inclusão nas ATS.

Trata-se de uma revisão com foco didático-pedagógico e de teor metodológico voltado para avaliação e interpretação de estudos de acurácia com aplicabilidade em avaliação de tecnologias em saúde. Para melhor compreensão dos termos aqui empregados, quanto ao resultado existem duas categorias de ETD: aqueles com resultados dicotômicos (positivo/negativo ou alterado/normal) e aqueles com resultados originados de variáveis contínuas, cuja definição de valores de referência e ponto de corte são essenciais. O primeiro conceito importante é o de validade, que é definida como a capacidade do exame/teste em distinguir pessoas que apresentam a doença em questão daquelas saudáveis. É de se esperar que entre as pessoas com a doença de interesse o resultado do teste seja considerado anormal ou alterado ou positivo, e o contrário nas pessoas sem a doença. A validade de ETD é composta por duas importantes propriedades: a sensibilidade (S) e a especificidade (E). A sensibilidade de um ETD é a sua capacidade de identificar corretamente indivíduos que apresentam a doença sob investigação, ou seja, se o teste for realizado em quem realmente possui a doença ou condição em estudo, o resultado será positivo, enquanto que a especificidade é a capacidade do exame/teste em identificar corretamente quem não tem a doença investigada, isto é, quando realizado em pessoas sem a condição ou doença, o resultado é negativo. Para exemplificarmos a aplicabilidade desses dois conceitos e seus complementos, suponha uma amostra populacional hipotética de 5000 pessoas, das quais 1000 apresentam determinada doença e o restante não a possui. Na tentativa de se verificar a validade de um novo teste diagnóstico, toda a amostra de indivíduos é submetida a esse novo teste (Tabela 1). Observando os resultados apresentados, quantas pessoas com a doença (S) e saudáveis (E) foram corretamente identificadas pelo teste?

Ao se analisar a tabela 1, pode-se notar que na primeira coluna, formada por 1000 indivíduos com a doença (casos) segundo o padrão-ouro*, o novo teste apresentou resultado positivo em 900 pessoas, ou seja, foi capaz de identificar adequadamente 90% dos doentes, logo, a "S" foi 900/1000 = 0,9 (90%), sendo estes resultados considerados "verdadeiros positivos" (VP). Os outros 100 resultados em indivíduos (doentes) em que o teste foi negativo, quando deveria ser positivo, são denominados "falsos negativos" (FN). Por outro lado, ao considerar-se a coluna dos 4000 sadios (controles ou sem a doença), o resultado do teste foi negativo em 3900 pessoas, ou seja, a "E" foi de 97,5% ou 3900/4000 = 0,975 (97,5%), sendo, portanto, capaz de identificar corretamente a quase totalidade dos não doentes como tal. Esses resultados são denominados "verdadeiros negativos" (VN), enquanto que os 100 resultados positivos são denominados de "falsos positivos" (FP). Em testes com resultados contínuos e naqueles com apresentação dicotômica resultante de categorização, com frequência, há uma relação inversa entre "S e E", ou seja, quanto mais sensível é um teste, menos específico ele tende a ser, sobretudo, quando há um ponto de corte estabelecido a partir de resultados contínuos. As variáveis e condições que interferem nesse equilíbrio serão discutidas posteriormente. ETD com altos percentuais de resultados FN e FP não são adequados, considerando-se que nos casos FN, doenças potencialmente graves e fatais (passíveis de tratamento) como, por exemplo, os cânceres, deixarão de ser diagnosticadas e tratadas, acarretando prejuízos irreparáveis aos pacientes e à sociedade, além do possível aumento de custos decorrente da necessidade da repetição dos testes ou

Tabela 1: Exemplo hipotético de uma amostra de 5000 indivíduos submetidos a um novo teste diagnóstico destinado a identificar pessoas com e sem determinada doença.

	Grupos (baseando-se em padrão-ouro*)		
Novo teste	***Doentes (Casos)***	***Sadios (Controles)***	***Total***
Positivo	900 (VP)	100 (FP)	**1000**
Negativo	100 (FN)	3900 (VN)	**4000**
Total	**1000**	**4000**	**5000**

* Padrão-ouro: exame ou teste diagnóstico único ou combinado capaz de representar com fidelidade o real estado (doente ou sadio) para a doença investigada (p. ex: biópsia de próstata nos casos suspeitos de adenocarcinoma prostático) .

realização de novos ETD confirmatórios, entre outras coisas. Por outro lado, resultados FP inevitavelmente levam a transtornos não menos danosos à população e aos pacientes, ao, sobretudo, rotulá-los (quase sempre pelo resto de suas vidas) como doentes quando na verdade não são, induzindo a distúrbios emocionais, incapacitação e aumentos de custos, quase sempre associados a tratamentos desnecessários. Ainda utilizando o exemplo da tabela 1, é importante citar mais algumas propriedades dos ETD: Valor preditivo negativo (VPN), definido como a probabilidade de que dado o resultado do novo teste foi negativo, a doença realmente não exista. O VPN pode ser calculado com os dados da tabela empregando-se como numerador os resultados VN e como denominador a soma dos FN + VN, assim no exemplo dado, VPN= 3900 (VN)/4000 (FN + VN) = 0,975 (97,5%), ou seja, a probabilidade de que pessoas do grupo de saudáveis realmente não terem a doença, quando o resultado do teste for negativo é de 97,5%. Valor preditivo positivo (VPP) que é definido como grau de probabilidade de que diante de um teste positivo, realmente a doença exista, essa propriedade é também denominada de probabilidade pós-teste ou prevalência. O VPP sofre grande influência da prevalência populacional (real) da doença sob investigação e, portanto, na maioria das vezes não pode ser calculado com dados obtidos de amostras, como o exposto na tabela 1, pois, como se pode observar a prevalência em tais casos é muito maior do que a prevalência na população. No exemplo dado, a prevalência da doença é de 20% (1000/5000), e ao calcular-se o VPP com tal cifra temos VPP = VP/VP+FP (900/900+100) = 0,9 (90%), significando que entre o grupo de doentes (casos), a probabilidade de o indivíduo estar realmente doente dado que o resultado do teste foi positivo é de 90%. No entanto, como mencionado anteriormente, há uma relação direta entre prevalência e VPP, ou seja, quanto maior a prevalência da doença na população maior o VPP (Tabela 2), sendo que com o VPN tal fato não é tão evidente. [12]

Por tal razão o cálculo do VPP requer o conhecimento da verdadeira prevalência da doença em estudo na população de interesse, para que se possa empregar outra forma de cálculo, o teorema de Bayes:

$$VPP = \frac{\text{Sensibilidade} \times \text{Prevalência}}{(\text{Sensibilidade} \times \text{Prevalência}) + (1 - \text{Especificidade}) \times (1 - \text{Prevalência})}$$

Tabela 2 – Relação entre Prevalência real de determinada doença na população e os valores preditivo positivo e preditivo negativo de um determinado ETD, mantendo-se constantes os valores de sensibilidade e especificidade.

Prevalência	*VPP(%)*	*VPN (%)*
1/100.000	0,5	100,00
1/10.000	4,5	100,00
1/1.000	32,2	99,99
1/500	48,8	99,99
1/200	70,5	99,99
1/100	82,7	99,99
1/50	99,7	99,89

No exemplo anterior, se aplicarmos essa fórmula com os valores da prevalência da doença naquela amostra (20%), teremos:

$$VPP = \frac{0,9 \times 0,2}{(0,9 \times 0,2) + (1 - 0,975) \times (1 - 0,2)} = 0,9\ (90\%)$$

ou seja, o mesmo valor do que o calculado na tabela, no entanto, supondo-se a mesma sensibilidade e especificidade encontradas para o novo teste, porém com uma prevalência real da doença na população de 5% (0,05), teremos:

$$VPP = \frac{0,9 \times 0,05}{(0,9 \times 0,05) + (1 - 0,975) \times (1 - 0,05)} = 0,65\ (65\%)$$

claramente nota-se que com uma prevalência quatro vezes menor, o VPP caiu de 90% para 65%.

Outras duas importantes propriedades na avaliação da acurácia dos testes diagnósticos são a razão de verossimilhança positiva (RVS +) ou "Positive Likelihood Ratio" e a razão de verossimilhança negativa (RVS -) ou "Negative Likelihood Ratio.[13,14] A primeira é definida como a razão de um resultado positivo de um ETD entre pacientes com a doença e sem a doença. Pode ser mensurada pela seguinte expressão, tendo-se como base o exemplo da tabela 1, RVS + = Sensibili-

dade/(1-Especificidade), assim no exemplo citado a RVS+ para o novo teste seria = 0,9/(1 – 0,975) = 36, ou seja, no caso exemplificado o teste tem 36 vezes mais chance de apresentar resultados positivos em pessoas doentes (casos) em relação a pessoas sem a doença (controles). A RVS - é a razão de um resultado negativo do ETD entre pacientes com a doença e sem a doença. Pode ser mensurada pela seguinte expressão, tendo-se como base o exemplo da tabela 1, RVS - = 1-Especificidade/Sensibilidade, desse modo usando o exemplo anteriormente citado, tem-se 0,025/0,9 = 0,028, isto é, o resultado negativo desse ETD está associado com a ausência da doença. Há ainda uma propriedade mensurável que sintetiza a qualidade global de um ETD, a acurácia - definida como sendo o número ou proporção de resultados do teste diagnóstico avaliado, que são corretamente classificados (verdadeiros positivos [VP] e verdadeiros negativos [VN]), [15,16] podendo ser calculada pela expressão: Acurácia = VN + VP/(VP+FN+VN+FP). No exemplo da tabela 1, a acurácia do teste diagnóstico avaliado seria = 3900 (VN) + 900 (VP)/(900[VP] + 100[FN] + 3900[VN] + 100[FP]) = 4800/5000 = 0,96 (96%), ou seja, o teste é capaz de fornecer 96% de resultados corretos (positivos e negativos). Os valores da razão de verossimilhança podem variar de zero ao infinito, sendo que resultados > 1 indicam associação positiva do ETD com diagnóstico da doença de interesse, isto é, quanto maior o valor encontrado, maior a probabilidade diagnóstica do teste. Por outro lado valores entre 0 e 1, depõem contra o diagnóstico, sendo que em achados <1, o resultado do ETD associa-se negativamente com o diagnóstico da doença e quanto mais próximo de zero, menor a probabilidade de diagnóstico da doença. Valor igual a 1 significa que o ETD avaliado apresenta a mesma proporção de resultados positivos entre doentes e sadios. A razão de verossimilhança é usada para avaliar o papel de um determinado ETD (isolado ou combinado) no diagnóstico de uma enfermidade em investigação, com vantagens sobre a "S e a E", pois, é menos dependente da prevalência real da doença, pode ser calculada para diferentes níveis de gravidade da patologia, ou pontos de corte diferentes, permite a combinação de resultados de testes diagnósticos múltiplos e, por último, possibilita o cálculo da probabilidade pós-teste da doença, sendo esta importante na tomada de decisão clínica diante do resultado encontrado. De posse dos valores da prevalência real da doença e da RVS + é possível se conhecer a probabilidade pós-teste da doença, em função do resultado do ETD encontrado, usando-se o nomograma de Fagan, construído a partir de cálculos pelo método Bayesiano [17], que pode ser acessado livremente através do seguinte endereço eletrônico: http://araw.mede.uic.edu/cgi-bin/testcalc.pl,[18] onde também são possíveis os cálculos de sensibilidade, especificidade e RVS+ e RVS -

Na prática clínica os ETD resultantes de variáveis contínuas como o nível de colesterol, glicose, creatinina e tantos outros elementos no sangue ou outros fluidos corporais, são mensurados e, portanto, suas propriedades, também devem ser rigorosamente avaliadas. Como os resultados obtidos, a princípio, não são dicotômicos torna-se necessário o estabelecimento de um nível ou ponto de corte, contido dentro de determinados valores ou uma faixa de referência, em torno da qual o achado é considerado normal ou anormal e, portanto, os indivíduos submetidos ao teste diagnóstico são considerados doentes ou não doentes para a enfermidade em estudo. Da mesma forma que para os ETD dicotômicos, após a definição do ponto de corte, a qualidade dos ETD contínuos pode ser avaliada através do cálculo da sensibilidade, especificidade, VPP (pelo teorema de Bayes), VPN, RVS +, RVS – e acurácia.

A faixa de referência (FR) é um intervalo numérico definido e específico para cada ETD e respectiva técnica, o que significa que para cada teste diagnóstico (p. ex. glicemia de jejum), existe uma FR correspondente e que para o mesmo teste, se a técnica de realização diferir, a FR geralmente também difere. Na prática clínica a FR é a representação de uma variação possível dos resultados encontrados e tidos como normais ou aceitáveis em 95% de uma amostra populacional considerada sem a doença de interesse, assim se presume que 5% desses indivíduos apresentarão resultados alterados (anormais), consequentemente, resultados levemente alterados (fora da FR) devem ser interpretados com cautela, pois, podem ser verdadeira ou falsamente anormais. Como as FR são construídas a partir de uma amostra de indivíduos sem a doença em estudo, os seus limites (inferior e superior) situam-se geralmente em um intervalo de confiança de 95%, sendo a faixa de normalidade compreendida entre dois desvios padrões para baixo e para cima em torno da média do resultado encontrado, considerando-se uma distribuição normal, entretanto, muitos elementos e substâncias de interesse, presentes nos fluidos corporais e mensurados pelos ETD podem apresentar distribuição não normal, dificultan-

do a adoção de valores referenciais confiáveis. Do mesmo modo, não há plena garantia de que resultados dentro da FR descartem a presença de doença, pois, os valores lá contidos não foram estabelecidos em amostras de pacientes com a doença em investigação.[19,20] A determinação da "S e da E" para ETD com resultados contínuos depende de se estabelecer um limiar que seja capaz de distinguir valores normais dos anormais, sendo este denominado nível ou ponto de corte. [21,22] Didaticamente vamos considerar o exemplo a seguir (Figura 1), onde se pode observar a relação entre um determinado ponto de corte, a "S e a E" para um determinado teste.

Pode-se claramente observar que à medida que o ponto de corte diminui (deslocamento para a esquerda), a "S" aumenta à custa de diminuição da "E", enquanto que com seu aumento (deslocamento para a direita), há aumento da "E" com diminuição da "S". Diante dessa interpendência entre "S e a E", o ponto de corte escolhido situa-se onde tais propriedades se encontram em níveis intermediários. Para se estabelecer um ponto de corte considerado como ótimo para que um ETD seja capaz de render resultados confiáveis, o uso da curva ROC (Receiver Operating Characteristic)[22] é uma ferramenta importante (Figura 2), ao possibilitar a visualização dos valores encontrados confrontando-os com valores de "S e E" correspondentes, permitindo inclusive a comparação entre dois ou mais ETD no diagnóstico da doença sob investigação.

Como se pode observar, através da curva ROC é possível determinar um ponto de corte ideal, onde a "S e E" se equilibram e apresentam maior performance dada pela maior área sob a curva (AUC). Na comparação de dois ou mais ETD em uma mesma curva, o teste cujo ponto de corte se aproximar mais do canto superior esquerdo, portanto, com maior AUC, deve ser visto como aquele que apresenta a melhor qualidade, não se esquecendo de que o julgamento e decisão clínica para sua adoção não dependem apenas disso. Além das qualidades anteriormente citadas, há outras propriedades dos ETD, entre elas a considerada mais importante é a reprodutibilidade [23], definida como sendo a capacidade de um teste em apresentar resultados semelhantes quando repetidos sob as mesmas condições. Diante do exposto, considera-se que um teste diagnóstico útil, minimamente apresente as seguintes características: a) metodologia adequada e bem descrita de modo a permitir que novos estudos sobre o teste possam ser confiavelmente e

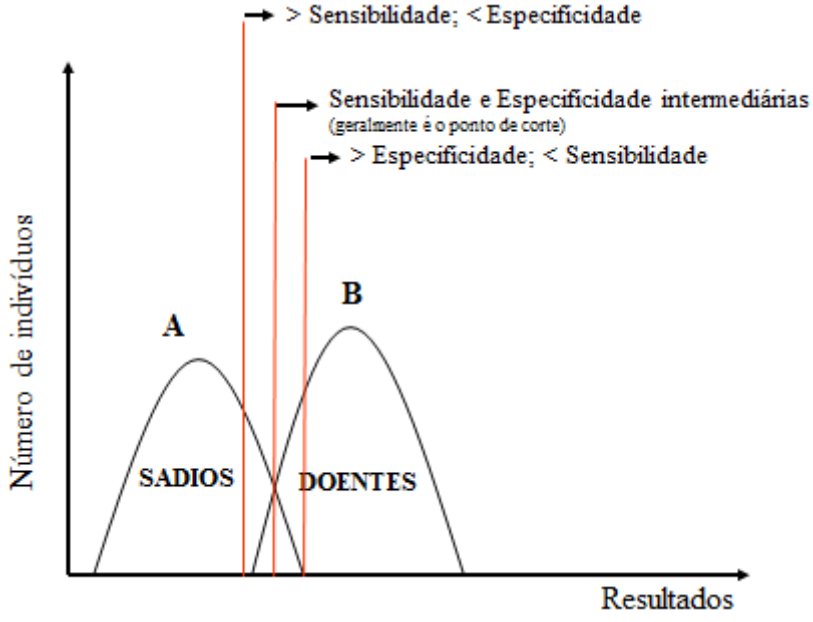
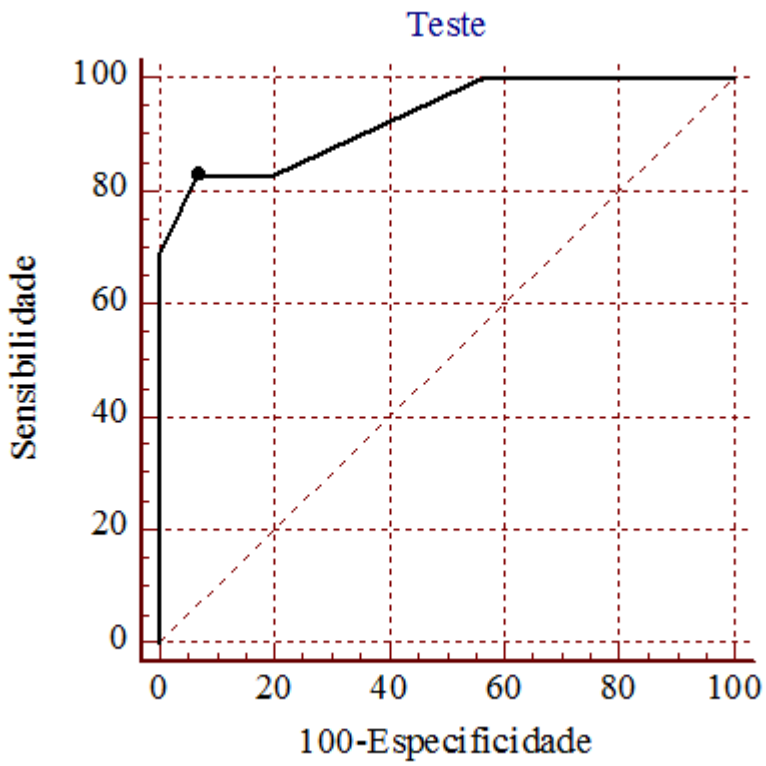


Figura 1 – Distribuição de duas populações hipotéticas (A – sem a doença; B – com a doença) e o posicionamento de três diferentes pontos de corte para um ETD hipotético.



Curva ROC

Prevalência da doença (%)								49,2
Área sob a curva ROC								0,929
Erro padrão								0,035
Intervalo de confiança a 95%								0,831 - 0,979
Nível de significância P (Área=0.5)								0,0001
Critério	Sensibilidade	IC95%	Especificidade	IC95%	RVS +	RVS -	VPP	VPN
< 40	0,00	0,0 - 12,1	100,00	88,3 - 100,0		1,00		50,8
<=40	3,45	0,6 - 17,8	100,00	88,3 - 100,0		0,97	100,0	51,7
<=56	10,34	2,3 - 27,4	100,00	88,3 - 100,0		0,90	100,0	53,6
<=60	17,24	5,9 - 35,8	100,00	88,3 - 100,0		0,83	100,0	55,6
<=64	44,83	26,5 - 64,3	100,00	88,3 - 100,0		0,55	100,0	65,2
<=68	68,97	49,2 - 84,7	100,00	88,3 - 100,0		0,31	100,0	76,9
<=72 *	**82,76**	**64,2 - 94,1**	**93,33**	**77,9 - 99,0**	**12,41**	**0,18**	**92,3**	**84,8**
<=74	82,76	64,2 - 94,1	90,00	73,4 - 97,8	8,28	0,19	88,9	84,4
<=76	82,76	64,2 - 94,1	80,00	61,4 - 92,2	4,14	0,22	80,0	82,8
<=80	100,00	87,9 - 100,0	43,33	25,5 - 62,6	1,76	0,00	63,0	100,0
<=84	100,00	87,9 - 100,0	3,33	0,6 - 17,3	1,03	0,00	50,0	100,0
<=88	100,00	87,9 - 100,0	0,00	0,0 - 11,7	1,00		49,2	

Figura 2 – Exemplo hipotético de um teste diagnóstico plotado em curva ROC. O ponto escuro representa o valor que alcançou o melhor balanço entre sensibilidade (82,76%) e especificidade (93,33%).

acuradamente reproduzidos; b) acurácia e precisão devidamente avaliadas e estabelecidas; c) "S e E" estabelecidas com base em comparação com um "padrão-ouro" para a doença de interesse, segundo diretrizes e protocolos recomendados; d) na avaliação do teste foram selecionadas pessoas com a doença e sem a doença de interesse com características semelhantes quanto à gravidade da enfermidade, tratamento, faixa etária, sexo de modo a minimizar viés de espectro e permitir generalização apropriada dos resultados. Assim como para ensaios clínicos[24] e estudos observacionais[25], onde a síntese de resultados é possível através de revisões sistemáticas com e sem metanálises, para estudos de avaliação de testes diagnósticos [26], isso também é possível e bastante desejável.[27] Revisões sistemáticas, principalmente com metanálises, dos estudos de acurácia, fornecem bons níveis de evidência científica e clínica para a avaliação dos ETD e as razões para isso são as mesmas que para as revisões dos ensaios clínicos [28], no entanto, publicações e diretrizes metodológicas sobre o tema ainda são escassas, sobretudo, voltadas para a ATS. A metodologia aplicada na realização de revisões sistemáticas dos estudos de acurácia de ETD, em seus pontos básicos, não difere muito daquelas dos ensaios clínicos: O acrônimo "PICO" pode também ser perfeitamente aplicado; na busca de literatura pode-se lançar mão dos mesmos bancos de dados (Pubmed, Scopus, Embase, Cochrane, etc), no entanto, as diferenças residem nos instrumentos de avaliação da qualidade dos estudos selecionados, nas ferramentas estatísticas analíticas e na interpretação dos resultados, aspectos que serão brevemente comentados a seguir.

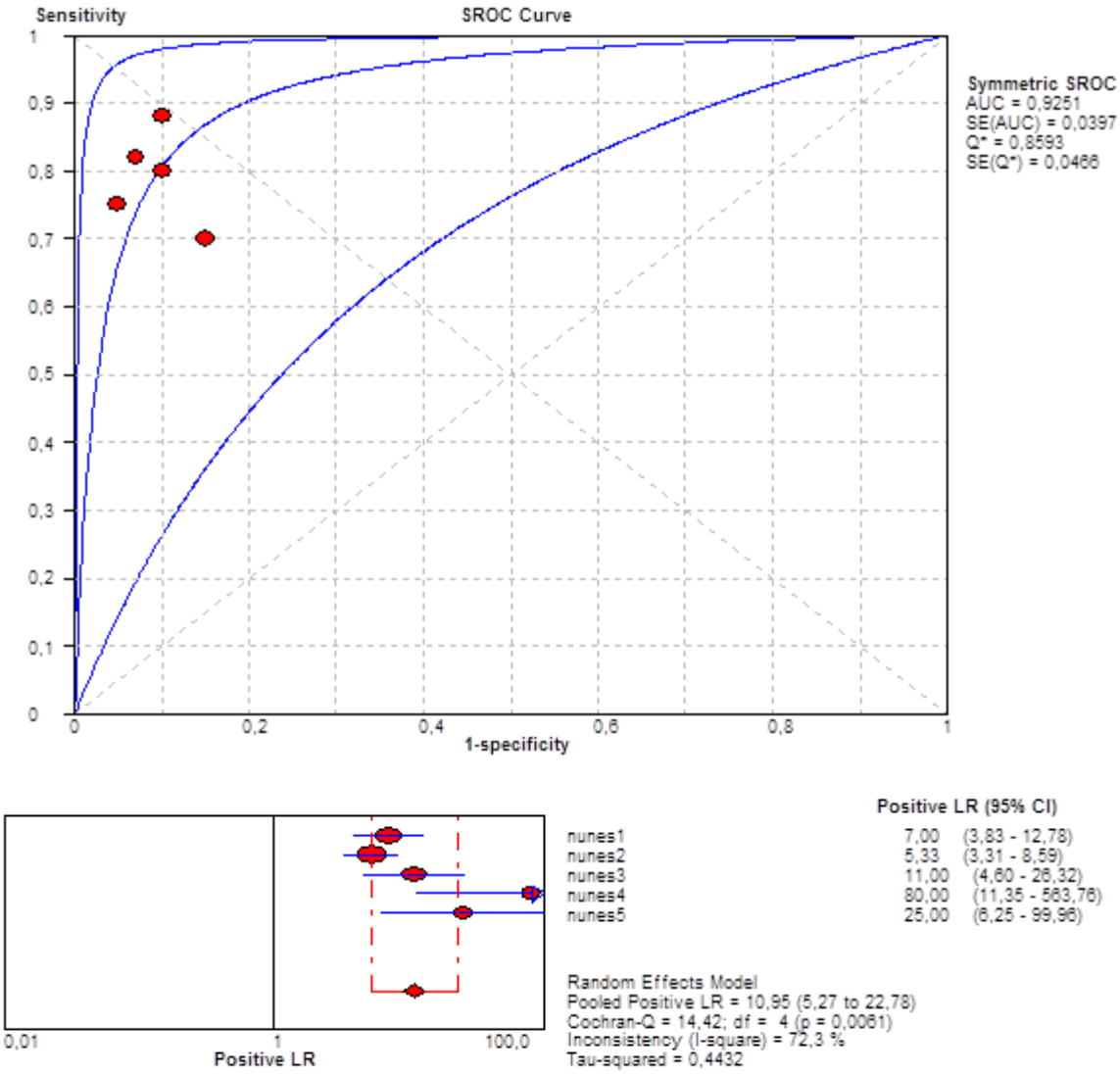
O passo inicial na avaliação dos estudos selecionados, visando a inclusão na revisão sistemática é a aplicação de guias como "*The Standards for Reporting of Diagnostic Accuracy*" (STARD) [29], um instrumento que lista 25 recomendações a serem observadas sobre aspectos metodológicos dos estudos. Atualmente os bons periódicos científicos adotam e exigem o STARD para a publicação de estudos de ETD. Outro importante instrumento para analisar a qualidade de estudos diagnósticos é o "*Quality Assessment of Diagnostic Ac-curacy Studies*" (QUADAS) [30], composto por 14 perguntas objetivas, cujas respostas são sim/não/pouco claro, sendo que cada resposta negativa ou pouco clara diminui a qualidade do estudo.[31] Importante salientar que por ter recebido críticas e sugestões de melhoria os autores desenvolveram o QUADAS-2 [32], uma versão mais elaborada que engloba quatro abrangentes domínios avaliados quanto ao risco de ocorrência de viés e considerações sobre aplicabilidade do teste.

A acurácia combinada dos estudos incluídos pode ser avaliada de algumas maneiras[33]: através de construção de uma curva ROC "combinada" ou *Summary Receiver Operating Characteristic* (S-ROC), onde os resultados de todos os estudos são plotados e há um resultado global de todos eles, demonstrando a "S e E" final daquele teste, sendo que a interpretação do resultado é semelhante à descrita anteriormente para a curva ROC "simples". Outra maneira de se avaliar a combinação dos resultados é o cálculo do Odds Ratio Diagnóstico (ORD ou DOR) que é uma medida de poder discriminatório de um ETD, sendo expresso por: ORD = S/(1- E)/1- E)/E. Para verificar a homogeneidade dos estudos a estatística Q de Cochran é empregada, enquanto que para análise de consistência da homogeneidade utiliza-se a estatística I^2. A apresentação dos resultados pode ser pela demonstração da curva SROC, gráfico de floresta, tabela com dados dos ORD individuais e total, tabela com dados de sensibilidade e especificidade individuais e totais. Notar que na curva SROC (Figura 3) o teste com melhor desempenho está apontado pela seta (mais próximo do canto superior esquerdo), enquanto que no gráfico de floresta há uma RVS+ combinada de 10,95, significando que resultados positivos do teste em análise estão bastante associados à presença da doença. No entanto, ao analisar-se o resultado do I^2 e da estatística Q, se deduz que há grande variabilidade e inconsistência entre os cinco estudos selecionados, portanto, apesar de estimadores favoráveis, os resultados devem ser avaliados com cautela.

ETD e tomada de decisão

Apesar de muito úteis, os métodos e estudos tem limitações por não levarem em consideração os inúmeros desfechos clínicos que podem ocorrer, bem como, os valores subjetivos que os pacientes, seus familiares e médicos depositam nas possíveis consequências clínicas, assim, a análise de decisão pode e deve ser usada para que tais fatores sejam adequadamente considerados.[33] Aconselha-se modelar as opções baseadas na decisão médica diante do resultado de determinado teste diagnóstico atribuindo probabilidades às alternativas existentes, assim como considerar os possíveis valores envolvidos como os desfechos clínicos, medidas de utilidade como anos de vida ajustados pela qualidade e custos, além de proceder

Figura 3:



Summary Diagnostic Odds Ratio (Random effects model)

Study	DOR	[95% Conf.	Iterval.]	% Weight
nunes1	21,000	9,618	- 45,851	26,45
nunes2	22,667	10,860	- 47,308	27,45
nunes3	23,222	8,700	- 61,986	22,24
nunes4	396,00	52,016	- 3014,8	9,19
nunes5	49,000	11,450	- 209,69	14,66
(REM) pooled DOR	**32,528**	**16,099**	**- 65,724**	

Heterogeneity chi-squared = 8,81 (d.f.= 4) p = 0,066
Inconsistency (I-square) = 54,6 %
Estimate of between-study variance (Tau-squared) = 0,3281
No. studies = 5.

Summary Sensitivity

Study	Sen	[95% Conf. Iterval.]	TP/(TP+FN)	TN/(TN+FP)
Barsam1	0,800	0,708 - 0,873	80/100	90/100
nunes2	0,750	0,653 - 0,831	75/100	95/100
Barsam3	0,700	0,600 - 0,788	70/100	85/100
nunes4	0,880	0,800 - 0,936	88/100	90/100
nunes5	0,820	0,731 - 0,890	82/100	93/100
Pooled Sen	**0,790**	**0,752 - 0,825**		

Heterogeneity chi-squared = 11,57 (d.f.= 4) p = 0,021
Inconsistency (I-square) = 65,4 %
No. studies = 5.

Summary Specificity

Study	Spe	[95% Conf. Iterval.]	TP/(TP+FN)	TN/(TN+FP)
Barsam1	0,900	0,824 - 0,951	80/100	90/100
nunes2	0,950	0,887 - 0,984	75/100	95/100
Barsam3	0,850	0,765 - 0,914	70/100	85/100
nunes4	0,900	0,824 - 0,951	88/100	90/100
nunes5	0,930	0,861 - 0,971	82/100	93/100
Pooled Spe	**0,906**	**0,877 - 0,930**		

Heterogeneity chi-squared = 6,69 (d.f.= 4) p = 0,153
Inconsistency (I-square) = 40,2 %
No. studies = 5.

Figura 3: Formas de apresentação de resultados de metanálises de estudos de diagnóstico (exemplo hipotético, empregando-se o Software Metadisc 1.4®).

aos cálculos que possibilitarão visualizar qual decisão oferece os melhores resultados para todos os envolvidos. O modelo mais apropriado para se avaliar custos e consequências de ETD é a árvore de decisão, onde são comparadas as alternativas de teste diagnóstico, bem como, as opções entre tratar e não tratar o paciente, diante do resultado encontrado, portanto, é perfeitamente possível e desejável a realização de avaliações econômicas dos ETD.

Considerações finais

O avanço na área de diagnóstico com auxílio de ETD de alta densidade tecnológica é notório, rapidamente progressivo e irreversível, resultando em elevação dos custos em saúde, tanto no setor público quanto no suplementar e privado. Assim sendo, compreender os instrumentos de ATS nesse contexto é fundamental para subsidiar o raciocínio clínico, direcionar os profissionais às boas práticas em saúde que resultem em benefício e segurança aos pacientes, além de racionalizar os custos para os sistemas de saúde e auxiliar os gestores da saúde na tomada de decisão com base em informações compreensíveis e confiáveis.

Referências

1. Goodman CS. Introduction to health care technology assessment:

ten basic steps. 1998. Disponível em: http://www.nlm.nih.gov/nichsr/hta101/hta101.pdf. Acesso em: 21/01/2013.

2. Taylor RS, Drummond MF, Sullivan SD. Inclusion of cost effectiveness in licensing requirements of new drugs: the fourth hurdle. BMJ. 2004; 329: 972-5.

3. Jönsson B. Technology assessment for new oncology drugs. Clin Cancer Res 2013; 19:6-11.

4. Rawlins M. De Testimonio: On the Evidence for Decisions About the Use of Therapeutic Interventions, The Harveian Oration of 2008. London: Royal College of Physicians; 2008.

5. National Institute for Health and Clinical Excellence. NICE guidance. NICE guidance research recommendations. Disponível em: http://www.nice.org.uk/guidance/GuidanceResearchReccommendations.jsp. Acesso em 08/01/2013.

6. Supporting the Medical Services Advisory Committee. Disponível em: http://www.msac.gov.au/internet/msac/publishing.nsf/Content/B8E1F7C44BE7E25BCA257A7D002477C3/$File/Short%20guidelines%20Sept-commentable.pdf. Acesso em: 08/01/2013.

7. Monteiro PHN, Alves OSF, Ianni AMZ, Salum e Morais ML. A Gestão da Incorporação Tecnológica no SUS: desafios para a formação de gestores. Boletim do Instituto de Saúde (BIS) 2007; 42: 29-31.

8. Ortiz de la Tabla González R, Martínez Navas A, Echevarría Moreno M. Neurologic complications of central neuraxial blocks. Rev Esp Anestesiol Reanim. 2011; 58:434-43.

9. Amorim JA, Remígio DS, Damázio Filho O, de Barros MA, Carvalho VN, Valença MM. Intracranial subdural hematoma post-spinal anesthesia: report of two cases and review of 33 cases in the literature. Rev Bras Anestesiol. 2010; 60:620-9, 344-9.

10. Lavi R, Rowe JM, Avivi I. Lumbar puncture: it is time to change the needle. Eur Neurol. 2010; 64:108-13.

11. Van Den Bruel A, Cleemput I, Aertgeerts B, Ramaekers D, Buntinx F. The evaluation of diagnostic tests: evidence on tecnhical and diagnostic accuracy , impact on patient outcome and cost-effectiveness is need. J Clin Epidemiol 2007; 60: 1116-22.

12. Leeflang MM, Bossuyt PM, Irwig L. Diagnostic test accuracy may vary with prevalence: implications for evidence-based diagnosis. J Clin Epidemiol 2009; 62:

5-12.

13. McGee S. Simplifying Likelihood Ratios. J Gen Intern Med 2002; 17: 647–50.

14. Altman DG, David M, Bryant TN, Gardner M. Statistics with confidence: confidence intervals and statistical guidelines. Second Edition. London: BMJ Books; 2000.

15. Steurer J, Fischer JE, Bachmann LM, Koller M, ter Riet G. Communicating accuracy of tests to general practitioners: a controlled study. BMJ 324 2002; (7341): 824–6.

16. WeinsteinS, Obuchowski NA, Lieber ML. Clinical Evaluation of Diagnostic Tests. Am J Roentgenol 2005; 184: 14-19.

17. Fagan TJ. Nomogram for Bayes' theorem. N Engl J Med 1975; 293:257.

18. Schwartz A. Diagnostic test calculator (version 2010042101). Available: http://araw.mede.uic.edu/cgi-bin/testcalc.pl?DT=&Dt=&dT=&dt=&2x2=Compute. Access: 28/01/2013.

19. Jung B, Adeli K. Clinical laboratory reference intervals in pediatrics: the CALIPER initiative. Clin Biochem 2009; 42:1589-95.

20. Hess AS, Shardell M, Johnson JK, Thom KA, Strassle P, Netzer G, et al. Methods and recommendations for evaluating and reporting a new diagnostic test. Eur J Clin Microbiol Infect Dis 2012; 31:2111-6.

21. Edler L, Ittrich C. Biostatistical methods for the validation of alternative methods for in vitro toxicity testing. Altern Lab Anim 2003; 31 Suppl 1:5-41.

22. Akobeng AK. Understanding diagnostic tests 3: Receiver operating characteristic curves. Acta Paediatr 2007; 96:644-7.

23. Tighe P, Negm O, Todd I, Fairclough L. Utility, reliability and reproducibility of immunoassay multiplex kits. Methods 2013; doi: 10.1016/j.ymeth.2013.01.003 (in press).

24. Martinez EZ. Metanálise de ensaios clínicos controlados aleatorizados: aspectos quantitativos. Medicina (Ribeirão Preto) 2007; 40: 223-35.

25. Porat S, Amsalem H, Shah PS, Murphy KE. Transabdominal amnioinfusion for preterm premature rupture of membranes: a systematic reviewand metaanalysis of randomized and observational studies. Am J Obstet Gynecol 2012; 207:393.e1-11.

26. Deeks JJ. Systematic reviews in health care: Systematic reviews of evaluations of diagnostic and screening tests. BMJ 2001; 323 (7305):157-62.

27. Leeflang MM, Deeks JJ, Gatsonis C, Bossuyt PM. Systematic reviews of diagnostic test accuracy. Ann Intern Med 2008; 149:889-97.

28. Lared W, Valente O. Revisões sistemáticas de estudos de acurácia. Diagn Tratamento 2009;14:85-8.

29. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, Moher D, Rennie D, de Vet HC, Lijmer JG; Standards for Reporting of Diagnostic Accuracy. The STARD statement for reporting studies of diagnostic accuracy: explanation and elaboration. Ann Intern Med 2003;138:W1-12.

30. Whiting P, Rutjes AW, Reitsma JB, Bossuyt PM, Kleijnen J. The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. BMC Med Res Methodol 2003; 3:25.

31. Reitsma JB, Rutjes AWS, Whiting P, Vlassov VV, Leeflang MMG, Deeks JJ. In: Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0.0. The Cochrane Collaboration. Deeks JJ, Bossuyt PM, Gatsonis C, editor. 2009. Chapter 9: Assessing methodological quality. Disponível em: http://srdta.cochrane.org. Acesso em 30/01/2013.

32. Whiting PF, Rutjes AW, Westwood ME, Mallett S, Deeks JJ, Reitsma JB, Leeflang MM, Sterne JA, Bossuyt PM; QUADAS-2 Group. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med 2011;155:529-36.

33. Devillé WL, Buntinx F, Bouter LM, Montori VM, de Vet HC, van der Windt DA, Bezemer PD. Conducting systematic reviews of diagnostic studies: didactic guidelines. BMC Med Res Methodol 2002; 2:9.