L'iniziativa STARD per la produzione di studi di accuratezza diagnostica: spiegazione e commenti*

P.M. Bossuyt[1], J.B. Reitsma[1], D.E. Bruns[2,3], C.A. Gatsonis[4], P.P. Glasziou[5], L.M. Irwig[6], D. Moher[7], D. Rennie[8,9], H.C.W. De Vet[10], J.G. Lijmer[1] per lo STARD Group

[1] Department of Clinical Epidemiology and Biostatistics, Academic Medical Center University of Amsterdam, 1100 DE Amsterdam, The Netherlands
[2] Department of Pathology, University of Virginia, Charlottesville, VA 22903
[3] Clinical Chemistry, Washington, DC 20037
[4] Centre for Statistical Sciences, Brown University, Providence, RI 02912
[5] Centre for General Practice, University of Queensland, Herston QLD 4006, Australia
[6] Department of Public Health & Community Medicine, University of Sydney, Sydney NSW 2006, Australia
[7] Chalmers Research Group, Ottawa, Ontario, K1N 6M4 Canada
[8] Institute for Health Policy Studies, University of California, San Francisco, San Francisco, CA 94118
[9] Journal of the American Medical Association, Chicago, IL 60610
[10] Institute for Research in Extramural Medicine, Free University, 1081 BT Amsterdam, The Netherlands

***Il documento è stato tradotto con il permesso del Direttore di Clinical Chemistry David E. Bruns da Romolo M. Dorizzi**

La qualità degli articoli dedicati alla accuratezza diagnostica è modesta. Articoli completi ed accurati sono necessari per consentire ai lettori di valutare il potenziale per la presenza di bias nello studio e la generalizzabilità dei risultati.
Un gruppo di ricercatori e di direttori di giornali hanno sviluppato il documento STARD (Standards for Reporting of Diagnostic Accuracy; Standard per la produzione d'articoli riguardanti l'Accuratezza Diagnostica) per migliorare la qualità degli articoli concernenti l'accuratezza diagnostica. Il documento consiste in una lista di controllo di 25 punti ed un diagramma di flusso che gli autori possono usare per assicurare che tutte le informazioni rilevanti siano presenti. Questo documento di spiegazione ha lo scopo di facilitare l'uso, la comprensione e la diffusione della lista di controllo. Il documento contiene una spiegazione del significato, del razionale e dell'uso ottimale di ogni punto della lista di controllo e di un breve riassunto delle prove disponibili su bias ed applicabilità.
Il documento STARD, la lista di controllo, il diagramma di flusso e questa spiegazione dovrebbero essere utili per migliorare gli articoli dedicati all'accuratezza diagnostica. Articoli completi ed informativi possono solo migliorare le decisioni in sanità.

Introduzione

Negli studi d'accuratezza diagnostica, i risultati prodotti da uno o più esami sono confrontati con i risultati ottenuti negli stessi soggetti con l'esame di riferimento. Tali studi di accuratezza costituiscono una tappa fondamentale nella valutazione di tecnologie diagnostiche nuove e tecnologie esistenti.
Molti fattori minacciano la validità interna ed esterna di uno studio d'accuratezza diagnostica[3-8]. Alcuni di questi fattori hanno a che fare con il disegno degli studi, altri con la selezione dei pazienti, l'esecuzione degli esami o l'analisi dei dati. In uno studio che ha coinvolto molte metanalisi è stato dimostrato che numerosi difetti di progettazione erano correlati a stime dell'accuratezza diagnostica eccessivamente ottimistiche[9].
Risultati esagerati ottenuti con studi diagnostici progettati male possono portare alla prematura adozione di esami diagnostici e portare i clinici ad iniziative terapeutiche errate relativamente ai singoli pazienti. I revisori dei giornali e gli altri lettori di studi diagnostici devono quindi conoscere tale potenziale di bias e tale potenziale mancanza di applicabilità.
Un'indagine sugli studi di accuratezza diagnostica pubblicati tra il 1978 e il 1993 in quattro dei più importanti giornali medici - ha rivelato che la qualità metodologica era, nel migliore dei casi, mediocre[8]. Questa rassegna ha anche mostrato che non erano fornite informazioni su elementi chiave del disegno, dell'esecuzione e dell'interpretazione degli studi diagnostici.
Per migliorare la qualità degli articoli dedicati all'accuratezza diagnostica è nata l'iniziativa *Standards for Reporting Diagnostic Accuracy*

(STARD). L'obiettivo dell'iniziativa STARD è quello di migliorare la qualità degli articoli dedicati all'accuratezza diagnostica. Un articolo completo ed accurato permette al lettore di individuare i potenziali di *bias* nello studio e di valutare la possibilità di generalizzare e applicare i risultati. A questo scopo il gruppo del progetto STARD ha prodotto una lista di controllo di una sola pagina. Tutte le volte che è stato possibile il gruppo STARD ha basato la decisione di inserire un punto sulle prove che legano il punto a stime a bias, variabilità nei risultati o limitazioni dell'applicabilità dei risultati ad altri contesti. La lista di controllo può essere usata per verificare che tutti gli elementi essenziali sono compresi nell'articolo.
Questo documento di spiegazione ha lo scopo di facilitare l'uso, la comprensione e la diffusione della lista di controllo. Il documento contiene una spiegazione del significato, del razionale e dell'uso ottimale di ogni punto sulla lista di controllo e un breve riassunto sulle prove disponibili relative a bias e applicabilità.
La prima parte di questo documento contiene un riassunto del disegno e della terminologia degli studi d'accuratezza diagnostica. La seconda parte contiene una discussione punto per punto con esempi.

Studi di accuratezza diagnostica

Gli studi di accuratezza diagnostica hanno una struttura di base comune[10]. Sono valutati uno o più esami, allo scopo di identificare, o di predire la presenza in una condizione bersaglio. La condizione bersaglio può riferirsi ad una malattia particolare, ad uno stadio di malattia, ad uno stato di salute o ad ogni altra condizione identificabile del paziente, come la stadiazione di una malattia già conosciuta o altre condizioni che dovrebbero originare degli interventi clinici come l'inizio, la modifica o l'interruzione di una terapia.
Il termine *"esame (test)"* si riferisce a qualsiasi metodo per ottenere informazioni aggiuntive sullo stato di salute di un soggetto. Comprende esami di laboratorio, esami di diagnostica per immagini, esami funzionali ed istologici, anamnesi e visita.
In uno studio di accuratezza diagnostica, l'esame in valutazione- l'esame indice- è eseguito in una serie di soggetti. I risultati ottenuti con l'esame indice sono confrontati con i risultati ottenuti con lo standard di riferimento ottenuti negli stessi soggetti. In questo ambito, lo *standard di riferimento* è considerato il miglior metodo disponibile per stabilire la presenza o l'assenza della condizione di interesse. Lo standard di riferimento può essere un singolo esame o una combinazione di esami e tecniche comprese il monitoraggio dei soggetti nel tempo (*follow-up*).
Il termine *accuratezza* si riferisce al grado di concordanza tra l'informazione ottenuta dall'esame in valutazione e lo standard di riferimento. L'accuratezza diagnostica può essere espressa in molti modi: coppia sensibilità -specificità, quoziente di probabilità (*likelihood ratio), odds ratio* diagnostici o come area sottostante una curva ROC[11-12].

QUESITO, DISEGNO E POTENZIALE DI BIAS DELLO STUDIO

All'inizio della valutazione di un esame, l'autore può semplicemente volere conoscere se un esame è in grado di discriminare due condizioni. La prima domanda corretta può essere "I risultati nell'esame nei pazienti con la malattia bersaglio sono diversi da quelli nei sani?" Se la risposta a questa domanda negli studi preliminari è affermativa, la domanda successiva è "I pazienti con specifici risultati degli esami hanno più probabilità di avere la patologa bersaglio rispetto a pazienti simili con altri risultati dell'esame?" Il disegno di studio usato per rispondere a questa domanda è quello di applicare l'esame indice e lo standard di riferimento a molti pazienti sospettati di presentare la condizione bersaglio.
Alcuni disegni di studio sono più soggetti al bias e hanno una applicabilità minore di altri. In questo articolo, il termine "bias" si riferisce alla differenza tra risultati osservati e risultati veri. Nessun singolo disegno è sicuramente fattibile ed in grado di fornire risposte valide, informative ed importanti a tutte le domande dello studio. Il lettore deve valutare per ogni studio la rilevanza, il potenziale per i bias e le limitazioni alla applicabilità; diventa quindi critico riportare i risultati in modo completo e trasparente. Per questo i punti della lista di controllo fanno riferimento al quesito della ricerca che ha originato lo studio della accuratezza diagnostica e chiede una descrizione esplicita e completa del disegno e dei risultati dello studio.

VARIABILITÀ

Le misure di accuratezza dell'esame possono essere diverse negli studi. La variabilità può riflettere differenze nei gruppi di pazienti, differenze di sede, differenze nella definizione della condizione bersaglio, differenze nei protocolli di esami o nei criteri di positività dell'esame[13].
Per esempio, si può verificare un bias se un esame è valutato in circostanze diverse da quelli del quesito della ricerca. Esempi sono la valutazione di un esame di screening per la malattia precoce in pazienti con malattia in fase avanzata o valutare uno strumento per analisi al letto del paziente in un dipartimento specialistico di un ospedale universitario.
Numerosi punti della lista di controllo hanno lo scopo di assicurare che un articolo contenga una descrizione chiara dei criteri di inclusione per i pazienti, i

protocolli di analisi e i criteri di positività insieme ad una descrizione adeguata dei soggetti inseriti nello studio e dei loro risultati. Questi punti consentiranno al lettore di giudicare se i risultati dello studio si applicano alle circostanze in cui operano.

Punti nella lista di controllo

La sezione seguente contiene una discussione punto per punto della lista di controllo. L'ordine dei punti corrisponde alla sequenza usata in molti articoli dedicati a studi di accuratezza diagnostica. Requisiti specifici dei giornali possono modificare l'ordine.

PUNTO 1. IDENTIFICARE L'ARTICOLO COME STUDIO DELL'ACCURATEZZA DIAGNOSTICA (RACCOMANDATO IL MeSH HEADING 'SENSITIVITY AND SPECIFICITY')

Esempio (estratto da un riassunto strutturato)
Scopo: Determinare sensibilità e specificità della tomografia computerizzata del colon per diagnosi differenziale tra polipo e cancro del colon-retto usando la colonscopia come standard di riferimento[14]

Le banche dati elettroniche sono diventate degli strumenti indispensabili per identificare gli studi. Per facilitare l'individuazione del loro studio, gli autori devono identificarlo in modo esplicito come un articolo dedicato alla accuratezza diagnostica. Raccomandiamo l'uso del termine "accuratezza diagnostica" nel titolo o nel riassunto di un articolo che confronta i risultati di uno o più esami indice con i risultati di uno standard di riferimento. Nel 1991 la banca dati MEDLINE della National Library of Medicine ha introdotto la parola chiave specifica (MeSH Heading) "Sensitivity and specificity" per gli studi diagnostici. L'uso di questa parola chiave per cercare gli studi di accuratezza diagnostica rimane problematico[15-19]. Negli articoli pubblicati da un gruppo selezionato di giornali indicizzati da Medline tra il 1992 ed il 1995 l'uso del MeSH Heading "Sensitivity and Specificity" ha identificato solo il 51% di tutti gli studi che si occupano di accuratezza diagnostica e ha identificato in modo non corretto molti articoli che non si occupavano di accuratezza diagnostica[18].
Nell'esempio, gli autori hanno impiegato il termine più generale "Performance Characteristics of CT colonography" nel titolo. La sezione scopo del riassunto strutturato menziona in modo specifico Sensitivity and Specificity. La registrazione MEDLINE di questo articolo contiene il MeSH "Sensitivity and Specificiy".

PUNTO 2. DEFINIRE I QUESITI O GLI OBIETTIVI DELLO STUDIO, AD ESEMPIO, STIMA DELL'ACCURATEZZA DIAGNOSTICA O CONFRONTO DELL'ACCURATEZZA TRA ESAMI O GRUPPI DI PARTECIPANTI

Esempio
L'angiografia coronarica invasiva rimane il gold standard per la diagnosi di una malattia coronarica clinicamente significativa.(….) Sarebbe desiderabile disporre di un esame non invasivo. Lo studio angiografico- risonanza magnetica eseguito mentre il paziente respira regolarmente ha raggiunto una maturità tecnica sufficiente per consentire una applicazione più diffusa con un protocollo standardizzato. Abbiamo quindi condotto uno studio per determinare l'[accuratezza] della angiografia- risonanza magnetica per diagnosticare la malattia coronarica[20].

La Dichiarazione di Helsinki afferma che la ricerca biomedica che coinvolge persone deve essere basata su una conoscenza approfondita della letteratura scientifica[21]. Nella introduzione degli articoli gli autori descrivono l'ambito scientifico, gli articoli pubblicati in questo settore, gli aspetti che rimangono non chiariti ed il razionale dello studio.
Specificare in modo chiaro i quesiti dello studio aiuta il lettore a giudicare l'appropriatezza del disegno dello studio e l'analisi dei dati. Una singola descrizione generale come "diagnostic value (valore diagnostico) o clinical usefulness (utilità clinica) non è di solito utile ai lettori.
Nell'esempio gli autori usano l'introduzione del loro articolo per descrivere il potenziale dell'angiografia coronarica-risonanza magnetica come alternativa non-invasiva alla angiografia tradizionale nella diagnosi di stenosi coronarica clinicamente significativa. Questa descrizione aiuta il lettore a giudicare l'appropriatezza della sezione criteri, la scelta dello standard di riferimento ed i metodi statistici usati per riassumere ed analizzare i dati.

PUNTO 3. DESCRIVERE LA POPOLAZIONE DELLO STUDIO: CRITERI D'INCLUSIONE ED ESCLUSIONE E SEDE IN CUI SONO STATI RACCOLTI I DATI

Esempio
Popolazione dei pazienti. Le pazienti seguite presso le cliniche di pianificazione familiare negli stati di Washington ed Oregon nel periodo 1992-1993 erano considerate per l'arruolamento nello studio. Per valutare l'arruolamento nello studio sono stati usati i criteri di screening pubblicati precedentemente dalla "Region X Chlamidya Project". Questi criteri comprendevano uno qualunque dei seguenti: I) cervicite mucupurulenta, malattia infiammatoria pelvica, cervicite e sanguinamenti; II) partner con segni e/o sintomi suggestivi di uretrite; III) richiesta della paziente, IV) violenza sessuale nei due mesi prece-

denti; V) valutazione preliminare all'inserimemento di anticoncezionale meccanico VI) test di gravidanza positivo e visita ginecologica. In alternativa i criteri di arruolamento comprendevano due o più dei punti seguenti; I) età inferiore ai 24 anni ed attività sessuale; II) nuovo partner nei due mesi precedenti; III) partner multipli nei due mesi precedenti; e IV) uso di metodi anti-concezionali senza barriera (contraccettivi orali, dispositivi meccanici, sterilizzazione, tutti i metodi naturali) o non impiego di metodi anticoncezionali[22].

Poiché l'accuratezza diagnostica descrive il comportamento di un esame in circostanze particolari, un articolo deve comprendere anche una descrizione della popolazione considerata. I criteri di selezione descrivono la popolazione di pazienti bersaglio, compresi ulteriori criteri di esclusione usati per ragioni di sicurezza e fattibilità.
I lettori devono sapere se lo studio escludeva i pazienti con una condizione specifica che si conosce interessare in maniera avversa i risultati dell'esame e che avrebbe aumentato l'accuratezza dell'esame (bias di applicazione limitata)[23]. Esempi sono rappresentati dall'esclusione di pazienti in terapia con beta-bloccanti in studi dell'elettrocardiografia da sforzo e l'esclusione di pazienti con patologia polmonare preesistente in studi di scintigrafia ventilatoria periferica[24,25].
I risultati degli esami possono essere diversi in ambito di assistenza primaria, secondaria e terziaria. L'esame si può comportare in modo diverso se l'esame è usato nello screening piuttosto che per conferma o per sospetto diagnostico. Lo spettro della malattia bersaglio e l'intervallo delle altre condizioni che compaiono nei pazienti sospettati di presentare la malattia bersaglio possono variare nelle diverse sedi a seconda delle modalità di ricovero in quella sede[25-28]. L'articolo deve quindi comprendere una precisa descrizione della sede in cui i pazienti sono stati reclutati e dove sono stati eseguiti l'esame in valutazione e l'esame di riferimento.

PUNTO 4. DESCRIVERE IL RECLUTAMENTO DEI PARTECIPANTI: È STATO BASATO SULLA SINTOMATOLOGIA ALLA PRESENTAZIONE, SUI RISULTATI DI ESAMI PRECEDENTI, O SUL FATTO CHE I PARTECIPANTI ERANO STATI SOTTOPOSTI ALL'ESAME IN VALUTAZIONE O A QUELLO DI RIFERIMENTO?

Un elemento importante della descrizione è l'indicazione di come sono stati identificati i pazienti da studiare. Il reclutamento dei partecipanti agli studi diagnostici può cominciare in modo diverso[10]. Spesso lo studio arruola in modo consecutivo i pazienti che sono sospettati clinicamente di presentare la malattia bersaglio a causa dei sintomi al momento della presentazione o perché inviati da altro specialista. Questi pazienti sono poi sottoposti all'esame (i) indice ed al gold standard.
Sono possibili altri disegni[2]. In alcuni studi i pazienti sono identificati dopo essere stati sottoposti all'esame indice. Altri studi iniziano dai pazienti in cui il gold standard ha stabilito o ha escluso la presenza della condizione bersaglio. Questi pazienti sono sottoposti successivamente all'esame indice. Altri studi ancora comprendono sia pazienti già diagnosticati con la malattia bersaglio sia partecipanti in cui la condizione è stata esclusa. Alcuni studi, spesso con la raccolta di dati retrospettivi, comprendono pazienti individuati attraverso il sistema informativo dell'ospedale per conoscere se sono stati o meno sottoposti allo standard di riferimento, all'esame indice o ad entrambi[29].
E' probabile che questi disegni di studio alternativi influenzino lo spettro di malattia nei pazienti compresi e lo spettro e la frequenza relativa delle condizioni alternative in pazienti che non presentano la malattia bersaglio. Nell'esempio presentato al punto 3, le ragioni dell'accesso alla clinica di pianificazione familiare non sono state dichiarate in modo esplicito.

PUNTO 5. DESCRIVERE IL RECLUTAMENTO DEI PARTECIPANTI: È AVVENUTO SU UNA SERIE CONSECUTIVA DI PAZIENTI DEFINITI DAI CRITERI DI SELEZIONE INDICATI AI PUNTI 3 E 4? SE NO, SPECIFICARE COME SONO STATI ULTERIORMENTE SELEZIONATI I PAZIENTI

Esempio
I pazienti sono stati arruolati in modo prospettico nei periodi in cui i ricercatori o gli associati allo studio erano disponibili[30].
Per definizione, la popolazione dello studio considerato comprende tutti i pazienti che soddisfano il criteri di inclusione e non sono eliminati da uno o più criteri di esclusione. I pazienti inclusi (quelli i cui risultati comprendono i risultati dello studio) possono essere o una serie consecutiva di pazienti che accedono al centro dello studio o una sottoselezione. La sottoselezione può essere e meno casuale (per esempio usando una tabella di numeri casuali).
E' importante per i lettori conoscere lo schema di campionamento, poiché può essere utile conoscere la generalizzabilità delle conclusioni dello studio.

PUNTO 6. DESCRIVERE LA RACCOLTA DEI DATI: LA RACCOLTA DEI DATI È STATA PIANIFICATA PRIMA CHE L'ESAME IN STUDIO E IL RIFERIMENTO FOSSERO ESEGUITI (STUDIO PROSPETTICO) O DOPO (STUDIO RETROSPETTIVO)?

Esempio
Abbiamo esaminato le cartelle cliniche di 251 pa-

zienti sottoposti da artrografia-TAC spirale del ginocchio. La popolazione dello studio consisteva di 50 pazienti selezionati consecutivamente sottoposti ad artografia-TAC spirale e successiva artroscopia presso la nostra istituzione ma non precedente atroscopia in quel ginocchio. Gli altri 201 pazienti comprendevano 12 che erano stati sottoposti prima ad artroscopia del ginocchio e dopo artroscopia, 69 che erano stati inviati da medici che operavano fuori dalla istituzione e 120 che non avevano subito artroscopia[31]

Se gli autori definiscono il quesito dello studio prima di identificare i pazienti e di raccogliere i dati, possono indirizzare la raccolta dei dati dello studio ai pazienti arruolati, usando dei moduli speciali di cartella clinica e moduli personalizzati di raccolta dati. La raccolta di dati prospettici e dedicati ha molti vantaggi; un migliore controllo di dati, ulteriori controlli sulla integrità e sulla congruità dei dati ed un livello di dettaglio clinico appropriato al problema[32]. Come risultato avremo un numero minore di dati mancanti o non interpretabili.
In alternativa, la raccolta dei dati può iniziare dopo che i pazienti sono stati sottoposti all'esame indice e al gold standard. La raccolta retrospettiva di dati è spesso basata sulla revisione delle cartelle cliniche. Gli studi con una raccolta retrospettiva dei dati può riflettere la pratica clinica routinaria meglio di uno studio prospettico, ma può anche non riuscire ad identificare tutti i pazienti selezionabili o non fornire dati di elevata qualità[29].

PUNTO 7. DESCRIVERE LO STANDARD DI RIFERIMENTO E LA SUA LOGICA

Esempio
L'allele 4-e del gene che codifica per l'apolipoproteina E (apoE) è associata fortemente alla malattia di Alzheimer, ma il suo valore nella diagnosi rimane incerto. (…) Abbiamo confrontato la sensibilità e la specificità della diagnosi clinica di malattia di Alzheimer, del genotipo ApoE e della diagnosi clinica e del genotipo ApoE misurati sequenzialmente usando la diagnosi istologica di malattia di Alzheimer come gold standard[33].

Negli studi di accuratezza diagnostica, si usa lo standard di riferimento per distinguere pazienti con la condizione bersaglio da quelli senza di essa. Alcune condizioni bersaglio non possono essere definite senza ambiguità. A seconda del quesito dello studio la condizione bersaglio può essere definita da rilevanza clinica, decisioni gestionali, prognosi o diagnosi istologica[10].
Quando non è possibile sottoporre tutti i pazienti al gold standard per ragioni pratiche o etiche, gli autori usano spesso un gold standard "multiplo". I diversi componenti di questo "gold standard" possono riflettere definizioni diverse delle diverse strategie per fare diagnosi della condizione bersaglio. Un esempio viene dall'impiego della translucenza nucale nel primo trimestre di gravidanza come marcatore della sindrome di Down[34]. In molti di questi studi, i risultati positivi sono stati verificati mediante l'esecuzione del cariotipo, mentre i risultati negativi sono stati verificati solo al parto. Gli studi in cui la decisione di eseguire il cariotipo del feto dipendeva dal risultato della translucenza nucale sovrastimavano in maniera considerevole la sensibilità della translucenza nucale[34].
Gli autori devono definire chiaramente lo standard di riferimento e come la scelta dello standard di riferimento si correla al gold standard dello studio in questione. Nell'esempio gli autori usano la diagnosi istologica dopo esame autoptico come gold standard nei pazienti inviati al centro per la malattia di Alzheimer per la valutazione di una demenza. Anche se la valutazione istologica è considerata il gold standard per la diagnosi di malattia di Alzheimer, la correlazione dei dati clinici ed istologici non è perfetta. Gli stessi anatomo-patologi non danno la stessa diagnosi di una determinata serie di preparati istologici[35].

PUNTO 8. DESCRIVERE LE SPECIFICHE TECNICHE DEI MATERIALI E DEI METODI IMPIEGATI COMPRESO COME E QUANDO SONO STATE ESEGUITE LE DETERMINAZIONI (E/O INDICAZIONE DEI RIFERIMENTI PER L'ESAME IN STUDIO E IL GOLD STANDARD)

Esempio
La concentrazione dell'antigene prostatico specifico (PSA) è stata misurata con i metodi che impiegano anticorpi monoclonali Tandem Total PSA e PSA libero (Hybritech). Un nuovo metodo immunofluorimetrico Time –resolved, recentemente sviluppato presso il nostro laboratorio è stato usato per misurare le concentrazioni di hK2. Brevemente, il metodo hK2 usa un anticorpo monoclonale di topo di cattura (codice G586 fornito da Hybritech (San Diego, CA) ottenuto contro hK2 ricombinante, un anticorpo monoclonale di topo biotinilato di rivelazione (codice 8311 Diagnostic Systems Laboratories) e streptavidina marcata con fosfatasi alcalina. Abbiamo misurato l'attività di fosfatasi alcalina aggiungendo il substrato diflunisal fosfato; incubando per 10 minuti e poi aggiungendo una soluzione che sviluppava Tb^{3+} EDTA. La fluorescenza è stata misurata con un immunoanalizzatore Cyberfluor 615 (MDS Nordion). Il dosaggio della hK2 ha un limite di rivelabilità di 0.006 μg/L ed una cross-reattività per il PSA inferiore allo 0.2%. Una descrizione completa e la valutazione del metodo è stata data altrove[36].

Gli autori devono descrivere i metodi impiegati nell'esecuzione dell'esame indice e del gold standard in un dettaglio sufficiente a permettere agli altri ricercatori di replicare lo studio e di permettere ai lettori di giudicare la fattibilità dell'esame indice presso di loro. Differenze nell'esecuzione dell'esame indice e del gold standard costituiscono una fonte potenziale di diversità nell'accuratezza diagnostica[13,24].
La descrizione deve comprendere l'intero protocollo dell'esame comprese le specifiche dei materiali e degli strumenti insieme alle loro istruzioni per l'uso e specifiche misure (preparazione) nei partecipanti (ad esempio digiuno prima del prelievo del sangue, sede anatomica della misura). Se non sono disponibili descrizioni, devono essere forniti i dettagli nel testo. Una variabilità tra studi nelle misure di accuratezza dell'esame causata da differenze nel protocolli d'esame è stata documentata per molti esami, compresi l'impiego della iperventilazione prima dell'elettrocardiografia da sforzo e l'uso del tomografia per la scintigrafia al tallio dopo sforzo[23,24]

PUNTO 9. DESCRIVERE DEFINIZIONE E RAZIONALE DELLE UNITÀ DI MISURA, DEI VALORI DI CUTOFF E/O DELLE CATEGORIE DEI RISULTATI DELL'ESAME IN STUDIO E DI QUELLO DI RIFERIMENTO

Esempio
Abbiamo scelto tre punti di cut-off del peptide natriuretico di tipo B per raggiungere valori di sensibilità di almeno il 90%, l'80% ed il 70% (37).

I risultati dell'esame possono essere realmente dicotomici (ad esempio, presente o assente), avere categorie multiple o essere continui. I lettori devono sapere come gli autori hanno espresso i risultati dell'esame indice o del gold standard.
Se gli autori hanno definito molte categorie di risultati, i lettori devono conoscere come e quando hanno definito i limiti delle categorie e se li hanno definiti prima dello studio o dopo che hanno ottenuto i risultati. Nell'ultimo caso aumenta la probabilità che gli autori abbiano adottato un cut-off per ottimizzare una particolare caratteristica dell'esame e diminuisce la probabilità che un altro studio possa replicare i risultati[38,39]. Nell'esempio, gli autori sono espliciti circa la selezione dei cut-off per il peptide natriuretico di tipo B nella diagnosi di disfunzione sistolica del ventricolo sinistro. Hanno stabilito questi cut-off post hoc per ottenere determinate sensibilità.

PUNTO 10. DESCRIVERE IL NUMERO, L'ADDESTRAMENTO E L'ESPERIENZA DELLE PERSONE CHE ESEGUONO ED INTERPRETANO L'ESAME IN STUDIO E QUELLO DI RIFERIMENTO

Esempio
Soggetti sono stati classificati come forti bevitori sulla base delle loro risposte al Self Administered Alcohol Screening Test (SAAST) ed al questionario Khavari relativo alla quantità di alcol assunto nel corso dell'anno precedente. I due questionari sono stati somministrati da un associato allo studio che era stato addestrato da una psicologa specialista specializzata nella terapia dell'alcolismo. Ha discusso con lei come dare i punteggi alle risposte e come chiarire le risposte ambigue e ha seguito la stessa psicologa mentre somministrava ella stessa più di 10 questionari[40].

La variabilità nella manipolazione, la processazione e l'interpretazione dello standard di riferimento influenzerà le misure di accuratezza diagnostica[41,42]. Molti studi hanno dimostrato la variabilità di interpretazione soprattutto nel campo della diagnostica per immagini[43,44]. L'addestramento di chi interpreta l'esame aiuta il lettore a giudicare se risultati simili possono essere ottenuti nella propria sede, dove chi interpreta l'esame può essere meno esperto.
La preparazione di base, l'esperienza e l'addestramento precedente nel migliorare l'interpretazione e nel ridurre la diversità tra osservatori influenzano la qualità dell'interpretazione[45,46]. E' più facile che esami (soggettivi) siano interpretati come anormali in sedi con la prevalenza della condizione bersaglio più alta; questa tendenza è definita bias di contesto[47].
L'esempio descrive il gold standard in uno studio di un modello che usa i risultati di esami di laboratorio eseguiti comunemente per identificare forti bevitori.

PUNTO 11. DESCRIVERE SE CHI INTERPRETAVA L'ESAME IN STUDIO E IL GOLD STANDARD ERA ALL'OSCURO DEI RISULTATI DELL'ALTRO ESAME (IN CIECO); INDICARE QUALSIASI ALTRA INFORMAZIONE CLINICA RESA DISPONIBILE A CHI INTERPRETA L'ESAME

Esempio
Tutte le immagini sono state interpretate indipendentemente sulla stazione computerizzata da due radiologi (JK, RKH) e successivamente è stata eseguita una lettura consensuale. I radiologi non conoscevano l'anamnesi del paziente e non sapevano se il paziente era stato reclutato per screening, o per la presenza di sintomi e non conoscevano i risultati della colonscopia standard e dell'esame istologico[14].

Conoscere i risultati dell'esame di riferimento può influenzare l'interpretazione dei risultati dell'esame indice e viceversa. E' probabile che tale conoscenza aumenti l'accordo tra i risultati dell'esame indice e quelli del gold standard, migliorando i valori di accuratezza diagnostica. La distorsione delle misure di

accuratezza diagnostica causata dalla conoscenza dei risultati dello standard di riferimento mentre si interpretano i risultati dell'esame indice è conosciuto come bias della revisione dell'esame[23]. Conoscere il risultato dell'esame indice mentre si interpretano i risultati del gold standard è stato denominato bias della revisione diagnostica[23]. L'osservazione che le interpretazioni diventano più accurate fornendo informazioni cliniche aggiuntive a chi interpreta è conosciuto come bias di revisione clinica[6,48,49].
Non dare informazioni a chi interpreta l'esame è definito come "far operare in cieco" (blinding) o mascherare. Chi interpreta i risultati può essere "in cieco" rispetto a risultati di altri esami e anche tutte le informazioni relative al paziente.
E' importante che chi interpreta gli esami operi "in cieco". In una metananalisi di un ampio spettro di esami, il bias di revisione dell'esame ha prodotto un moderato aumento delle misure di accuratezza diagnostica[9]. Singoli studi hanno mostrato un effetto rilevante di inappropriato mascheramento[24].
L'esempio mostra come i lettori delle colongrafie CT per la diagnosi differenziale tra polipo e cancro del colon-retto erano in cieco rispetto ad ulteriori informazioni cliniche ed ai risultati della colonscopia, il gold standard.

PUNTO 12. DESCRIVERE I METODI PER CALCOLARE O CONFRONTARE I PARAMETRI DI ACCURATEZZA DIAGNOSTICA E I METODI STATISTICI IMPIEGATI PER QUANTIFICARE L'INCERTEZZA (AD ESEMPIO, INTERVALLI DI CONFIDENZA AL 95%)

Esempio
La significatività statistica delle differenze di sensibilità tra angiografia-risonanza magnetica (MRA) ed ecografia-duplex è stata valutata con il test di McNemar[50].

Le misure di accuratezza diagnostica sono numerose[12]. Gli autori devono descrivere in sufficiente dettaglio i metodi usati per calcolare le misure che ritengono appropriate.
La stima dell'accuratezza diagnostica è soggetta a variazioni casuali, e gli studi più ampi sono di solito più precisi. Gli autori devono quindi quantificare l'incertezza statistica del valore osservato[51]. Esistono articoli che descrivono i metodi per calcolare la precisione intorno alle misure di accuratezza diagnostica[12].
In alternativa possono essere usate tecniche statistiche per verificare ipotesi più specifiche, come la superiorità di un esame rispetto ad un altro o l'ipotesi che una specifica misura di accuratezza diagnostica sia superiore ad un valore pre-specificato.
Nell'esempio gli autori hanno usato il test statistico di McNemar per respingere l'ipotesi nulla che l'angiografia-risonanza magnetica ha la stessa sensibilità della ecografia duplex per la diagnosi della malattia renovascolare.

PUNTO 13. DESCRIVERE I METODI PER CALCOLARE LA RIPRODUCIBILITÀ, SE IMPIEGATI

Esempio
La variabilità tra osservatori nella interpretazione della angiografia convenzionale e angiografia-risonanza magnetica (MRA) è stata calcolata usando la statistica k compresi gli intervalli di confidenza al 95%[50].

L'esame indice ed il gold standard sono di rado perfetti. La loro riproducibilità varia e una ridotta riproducibilità influenza in maniera negativa l'accuratezza diagnostica[52].
La variabilità dell'osservatore si può verificare negli esami di diagnostica per immagine in cui l'osservatore deve riassumere delle osservazioni visive in un'affermazione circa la presenza di malattia. Si presenta anche durante la classificazione, quando l'osservatore deve usare i dati per classificare i pazienti in categorie diagnostiche[41]. La variabilità strumentale interessa la variazione che si manifesta durante l'uso di dispositivi o sistemi, come le analisi automatiche di laboratorio. Altri termini di questa forma di variazione comprendono imprecisione, variazione metodologica analitica o rumore di fondo analitico (errore). Una scarsa riproducibilità influenza in maniera negativa l'accuratezza diagnostica. Se possibile, gli autori devono valutare la riproducibilità dei metodi usati nel loro studio e descrivere come lo hanno fatto.
Per i metodi quantitativi, è utile riportare l'imprecisione come coefficiente di variazione a due o più valori medi specificati più vicini a punti decisionali clinici ottenuti ripetendo l'esame in un numero definito di giorni diversi. I coefficienti di variazione intra-dosaggio sono appropriati se tutti i campioni dei pazienti sono analizzati in una singola seduta analitica.
Nell'esempio gli autori hanno usato la statistica kappa per esprimere la variabilità tra osservatori per angiografia tradizionale e MRA nella diagnosi di malattia renovascolare.

PUNTO 14. INDICARE QUANDO È STATO CONDOTTO LO STUDIO COMPRESE LE DATE DI INIZIO E DI FINE DEL RECLUTAMENTO

Esempio
Abbiamo sottoposto retrospettivamente a screening tutte le emocolture provenienti da pazienti ricoverati nel reparto di Oncologia del New England Medical Center, un ospedale universitario di 300 letti per as-

sistenza terziaria tra l'agosto 1994 e il giugno 1996[53].

La tecnologia coinvolta in molti esami progredisce continuamente, migliorando l'accuratezza diagnostica. Le data in cui è stato eseguito lo studio e quella in cui sono stati pubblicati i risultati possono essere molto lontane. I lettori vogliono quindi conoscere la data in cui è stato condotto lo studio. Questa informazione può dare anche una indicazione circa la velocità di reclutamento.

PUNTO 15. INDICARE LE CARATTERISTICHE CLINICHE E DEMOGRAFICHE (AD ESEMPIO ETÀ, SESSO, SPETTRO DEI SINTOMI, PATOLOGIA INTERCORRENTE, TRATTAMENTI IN CORSO, CENTRI DI RECLUTAMENTO)

Esempio
Caratteristiche demografiche, cliniche e angiografiche dei 109 pazienti dello studio[20] *(vedi tabella).*

Una descrizione adeguata delle caratteristiche demografiche e cliniche dei partecipanti permette al lettore di giudicare l'applicabilità dei risultati dello studio ad un'altra popolazione. La maggior parte degli autori presenta le caratteristiche demografiche e cliniche del gruppo studiato in una tabella.

PUNTO 16. INDICARE IL NUMERO DI PARTECIPANTI CHE SODDISFANO I CRITERI D'INCLUSIONE CHE NON SONO STATI SOTTOPOSTI ALL'ESAME IN VALUTAZIONE E/O ALL'ESAME DI RIFERIMENTO; INDICARE PERCHÉ ALCUNI PARTECIPANTI SONO STATI EVENTUALMENTE ESCLUSI DALL'UNO O DALL'ALTRO (L'USO DI UN DIAGRAMMA DI FLUSSO È MOLTO RACCOMANDATO)

Esempio 1
Durante lo studio 272 pazienti con sospetto di trombosi venosa profonda (TVP) sono stati inviati ai centri partecipanti allo studio. Di questi 28 sono stati esclusi per le seguenti ragioni: precedente TVP[21]*, allergia da contrasto*[1]*, insufficienza renale*[1]*, non concessione del consenso da parte del paziente*[5]*. 25 dei rimanenti 244 pazienti sono stati esclusi dall'analisi perché la venografia era risultata inadeguata o era fallita e 5 perché la pletismografia ad impedenza era risultata inadeguata o era fallita*[54].

Esempio 2. (vedi Figura 1)
Lo studio deve presentare il numero dei partecipanti che erano valutati per la selezione, se disponibili. Questo numero è un indicatore utile di quanto la popolazione dello studio è strettamente simile alla popolazione dei pazienti.
Il diagramma di flusso fornisce il numero esatto di pazienti ad ogni fase dello studio e quindi il denominatore corretto per calcolare le percentuali e le proporzioni. Mostra anche il numero di soggetti che non è stato sottoposto all'esame indice e/o al gold standard.
Le misure di accuratezza diagnostica avranno un bias se i risultati dell'esame indice influenzano la decisione di richiedere l'esecuzione del gold standard[56-63]. I termini usati per descrivere questo effetto comprendono il bias di verifica (parziale), il bias di valutazione (work-up), bias di selezione (primario), bias di richiesta sequenziale e bias di verifica (il termine più generale). Il bias di verifica è presente in

Caratteristiche	Valore
Sesso femminile - n (%)	34 (31)
Età – anni	
Media ± DS	59 ± 10
Intervallo	27-75
Dolore toracico - n (%)	89 (79)
Precedente infarto del miocardio - n (%)	26 (24)
Storia di ipertensione sistemica- n (%)	54 (50)
Fumatore (attuale o precedente) - n (%)	58 (53)
Colesterolo > 200 mg/dL- n (%)	67 (61)
Diabete- n (%)	19 (17)
Anamnesi familiare positiva di malattia coronarica prematura*- n (%)	43 (39)
Reperti agiografici - n (%)	
Malattia di un vaso	31 (28)
Malattia di due vasi	20 (18)
Malattia di tre vasi	13 (12)

***Anamnesi familiare positiva è stata definita come una storia positiva di infarto del miocardio e angina in un parente di primo grado prima dell'età di 65 anni**

Figura 1. Esempio di diagramma di flusso di uno studio di accuratezza diagnostica[55].

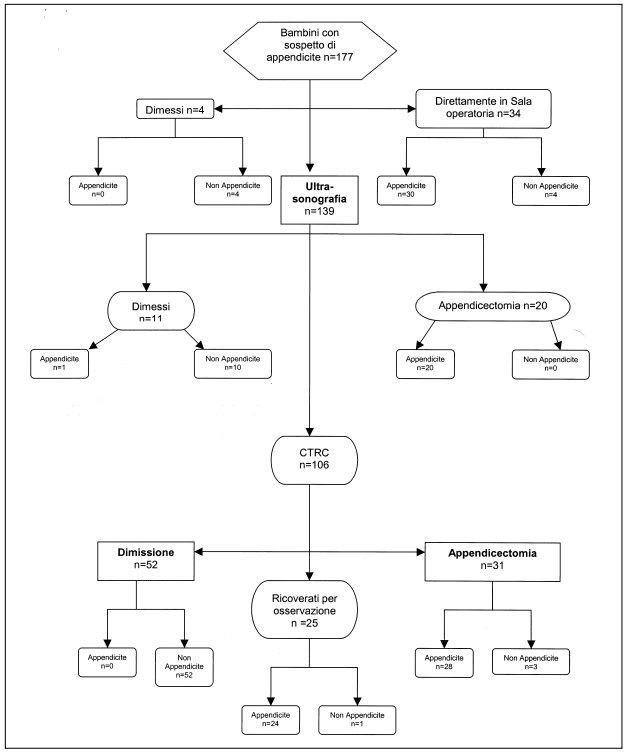
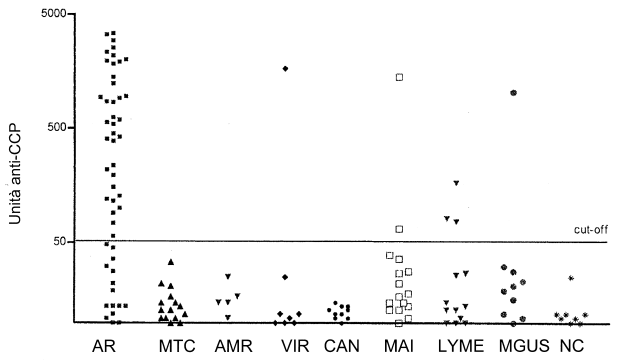


Figura 2. Esempio su scala logaritmica dei risultati di un esame in unità anti-CCP per gruppi di pazienti diversi.

Un valore di cut-off di 50 unità garantisce una buona specificità perché tutti tranne 7 dei pazienti non AR presentano una concentrazione di anticorpo al di sotto della soglia (AR: Artrite Reumatoide; MTC: Malattie tessutali connettivali; AMR: Altre malattie rematiche; VIR: Malattie virali; CAN: Cancro; MAI: Malattie autoimmuni tessutali; LYME: Malattia di Lyme; MGUS: Gammapatia monoclonale di incerto significato; NC: Controlli)



un numero di esami diagnostici che arriva al 26% ed è particolarmente comune quando lo standard di riferimento è una procedura invasiva[60].
Raccomandiamo molto l'uso del diagramma di flusso per illustrare il disegno dello studio e fornire il numero esatto di partecipanti ad ogni stadio dello studio. Un diagramma di flusso può trasmettere in modo trasparente gli elementi chiave del disegno dello studio. Un diagramma di flusso è stato una utile aggiunta agli articoli dedicati ai trial clinici randomizzati[64].

PUNTO 17. INDICARE L'INTERVALLO TRA ESECUZIONE DELL'ESAME IN VALUTAZIONE E DELL'ESAME DI RIFERIMENTO E FORNIRE NOTIZIE SU EVENTUALI TRATTAMENTI SOMMINISTRATI NEL FRATTEMPO

Esempio
Era programmato che i pazienti fossero sottoposti a colongrafia-TAC prima della colonscopia convenzionale ed entrambe sono state eseguite nello stesso giorno[14].
In termini epidemiologici gli studi dedicati alla accuratezza diagnostica sono cross-sezionali. I risultati dell'esame indice ed il gold standard sono eseguiti negli stessi pazienti e nello stesso momento[10].
Quando si verifica un ritardo tra il momento in cui si esegue l'esame indice ed il momento in cui si esegue il gold standard, la condizione del paziente può cambiare, portando ad un peggioramento o un miglioramento delle condizioni bersaglio o delle condizioni rispetto alle quali si deve fare diagnosi differenziale.
Le preoccupazioni sono simili se s'inizia una terapia dopo che è stato eseguito l'esame indice ma prima del gold standard.

PUNTO 18. INDICARE LO SPETTRO DI GRAVITÀ DELLA MALATTIA INDAGATA (DEFINIRNE I CRITERI) IN COLORO CHE NE SONO AFFETTI; DESCRIVERE LE ALTRE PATOLOGIE PRESENTI NEI PARTECIPANTI CHE NON NE SONO AFFETTI.

Le caratteristiche demografiche e cliniche della popolazione studiata possono influenzare le misure di accuratezza diagnostica. Questa variabilità è conosciuta come bias di spettro[56]. L'effetto di spettro comprende la gravità della patologia bersaglio, le caratteristiche demografiche e le patologie coesistenti. Tutti questi elementi hanno causato variabilità nelle misure di accuratezza dell'esame, ma gli esempi più rilevanti hanno interessato differenze nella gravità della condizione bersaglio[65-70].
Molte patologie bersaglio non sono condizioni puramente dicotomiche ma coprono un continuum che va da patologie iniziali a patologie avanzate. La sensibilità dell'esame è spesso più alta in studi con una percentuale più alta di pazienti in uno stadio della malattia bersaglio più avanzato[56]. D'altra parte, in presenza di altre patologie si possono verificare più frequentemente risultati falsi-positivi o falsi-negativi[25,56,71].
E' importante quindi descrivere la gravità della malattia nel gruppo studiato.

PUNTO 19. PRESENTARE UN TABELLA DI CONFRONTO DEI RISULTATI DELL'ESAME IN STUDIO E DI QUELLI DELL'ESAME DI RIFERIMEN-

	Nondiagn	*Normale*	*Atipia*	*Neopl. Fol.*	*Sospetto*	*Maligno*
PAP	12	30	5	17	18	18
FOL	18	31	3	40	5	3
MID	15	15	4	11	28	27
ANAPL	18	12	5	5	13	47
Totale	14	28	4	23	14	17

Nondiagn: non diagnostico; Neopl. Fol.: Neoplasia Follicolare; PAP: Carcinoma papillare; FOL: carcinoma follicolare; MID carcinoma midollare; ANAPL Carcinoma anaplastico[72].

TO (COMPRESI I RISULTATI INDETERMINATI E MANCANTI); PER RISULTATI CONTINUI PRESENTARE LA DISTRIBUZIONE DEI RISULTATI DELL'ESAME IN VALUTAZIONE RISPETTO A QUELLI DELL'ESAME DI RIFERIMENTO

Esempio 1
Distribuzione degli outcome citologici all'interno di ogni tipo istologico di cancro della tiroide
Esempio 2 (Vedi Figura 2)
I ricercatori vogliono verificare i risultati importanti e quindi la ripetizione dell'analisi è un aspetto importante del metodo scientifico. Per facilitare questo processo, gli autori devono presentare i risultati come numeri assoluti. La tabulazione incrociata dei risultati dell'esame in categorie ed i grafici di distribuzione dei risultati continui sono essenziali per permettere ai colleghi di (ri)calcolare l'accuratezza diagnostica o di eseguire altre analisi, comprese le metanalisi. Gli autori devono presentare tutti i risultati, compresi quelli indeterminati ottenuti nell'esame indice e nel gold standard.
Un esempio con poche categorie di risultati dell'esame è ricavato da uno studio di citologia ad ago sottile in casi di carcinoma della tiroide diagnosticati istologicamente; il secondo esempio mostra la distribuzione della concentrazione degli anticorpi anti-citrullina in pazienti con la patologia bersaglio (artrite reumatoide) e in pazienti con diagnosi diversa.

PUNTO 20. SEGNALARE OGNI EVENTO AVVERSO CAUSATO DALLA ESECUZIONE DELL'ESAME IN VALUTAZIONE O DELL'ESAME DI RIFERIMENTO

Esempio
Un tempo medio di 15 minuti era sufficiente per una indagine completa della cavità uterina. La tolleranza media del dolore su una scala del dolore da 0 a 10 era 1. Tuttavia la ecoisterografia non è stata tollerata una volta (indicazione di dolore pari a 10). La paziente ha provato dolore pelvico che è receduto dopo somministrazione con fosfoglucinolo e dopo 20 minuti di riposo in posizione distesa. Si è verificata una sola complicazione, una endometrite in una paziente con diabete scompensato nei tre giorni successivi l'esame. Un trattamento antibiotico con ampicillina ha risolto completamente il quadro[74].

Non tutti gli esami sono sicuri. Misurare e segnalare gli eventi collaterali negli studi di accuratezza diagnostica può fornire informazioni supplementari circa l'utilità clinica di un particolare esame. Il requisito di segnalare eventi collaterali si applica sia alla ricerca in ambito diagnostico sia alla ricerca in ambito terapeutico[75].
Può essere importante anche avere informazioni circa l'invasività ed i rischi del gold standard usato. Per esempio, se nella valutazione di un risultato positivo dello screening dell'Hemocult, sono usati colonscopia, sigmoidoscopia e clisma opaco a doppio contrasto si può prevedere una complicanza (perforazione o emorragia) ogni 300-900 esami[76].
L'esempio proviene dalla prima parte della sezione risultati di uno studio di ecoisterografia per la diagnosi di anomalie intrauterine, usando come gold standard composti l'istologia e l'outcome clinico.

PUNTO 21. RIPORTARE LE STIME DELL'ACCURATEZZA DIAGNOSTICA E MISURE DELL'INCERTEZZA STATISTICA (AD ESEMPIO INTERVALLI DI CONFIDENZA AL 95%)

Esempio
Curve ROC che confrontano i valori CDTect con la percentuale di valori di CDT indipendentemente per uomini e donne sono presentate in Fig.2 (…) Le aree sotto le curve (con intervallo di confidenza al 95%) sono rispettivamente 0.88 (0.85-0.91) e 0.89 (0.86-0.92) negli uomini (P = 0.67) e 0.72 (0.68-0.76) e 0.76 (0.72-0.81) nelle donne (P = 0.26)[77].

Lo scopo finale di uno studio di accuratezza diagnostica è l'espressione di quanto bene i risultati dell'esame si correlano con la presenza o l'assenza della condizione bersaglio, come stabilita dal gold standard. I valori presentati nell'articolo devono essere considerati come una stima. A causa delle variazioni casuali nei pazienti sottoposti agli esami ed ad altri fattori, è probabile che i risultati cambino se lo studio è replicato anche nella stessa popolazione[51]. Indicare la precisione mostrerà al lettore l'ambito di

probabilità intorno alla accuratezza diagnostica.
Molti giornali richiedono o raccomandano fortemente l'uso degli intervalli di confidenza come misure di precisione. Convenzionalmente si considera un intervallo di confidenza del 95%. Solo il 50% degli articoli sulle valutazioni diagnostiche pubblicati nel 1996 o nel 1997 nel British Medical Journal conteneva la precisione per la stima dell'accuratezza diagnostica[78].

PUNTO 22. INDICARE COME SONO STATI GESTITI I RISULTATI INDETERMINATI, LE RISPOSTE MANCANTI E GLI OUTLIER DELL'ESAME IN VALUTAZIONE

Risultati dell'esame non interpretabili, indeterminati ed intermedi pongono un problema nella valutazione di un esame diagnostico[71,79,80]. La stessa frequenza di questi risultati rappresenta un indicatore importante dell'utilità dell'esame. Inoltre, ignorare questi risultati può portare a stime non corrette dell'accuratezza diagnostica se questi risultati si verificano più frequentemente nei pazienti con la condizione bersaglio che in quelli che non la presentano o viceversa.
Le cause di risultati dell'esame non interpretabili, indeterminati ed intermedi possono essere numerose[79]. E' possibile non ottenere un risultato dell'esame per cause tecniche o per scarsità del campione (per esempio, mancanza di cellule nel preparato di una biopsia con ago sottile da un cancro, risultato non interpretabile)[45,81,82]. Un risultato di un esame può essere invalidato da una condizione medica concomitante o da una terapia che influenza l'esame, per esempio l'effetto di farmaci beta-bloccanti sulla risposta cardiaca durante un test da sforzo (risultato non determinato)[24].
La frequenza di risultati dell'esame non interpretabili, indeterminati ed intermedi varia nei diversi esami, ma la frequenza può arrivare al 40%[79]. Risultati dell'esame intermedi (non chiaramente positivi o negativi) possono avere un valore diagnostico, come nel caso della scintigrafia a ventilazione-perfusione che non sono né normali né indicano con alta probabilità una di embolia polmonare[83]. L'inserimento di tali risultati nel processo decisionale clinico è variabile[80].
La Tabella presentata dopo il punto 19 comprende correttamente i risultati non diagnostici dell'esame.

PUNTO 23. INDICARE LE STIME DELLA VARIABILITÀ DELL'ACCURATEZZA DIAGNOSTICA TRA SOTTOGRUPPI DI PARTECIPANTI, LETTORI O CENTRI (SE EFFETTUATE)

Esempio
La sensibilità e la specificità (…) per l'angiografia-risonanza magnetica (…) per la diagnosi di una stenosi emodinamicamente significativa di una arteria renale principale erano del 90%. Una volta esclusi dalla diagnosi i pazienti con displasia fibromuscolare, la sensibilità della angiografia-risonanza magnetica è aumentata al 97%, con un valore predittivo negativo del 98%[50].

Poichè la variabilità è la regola piuttosto che l'eccezione, i ricercatori devono esplorare le possibili cause di eterogeneità nei risultati, entro i limiti delle dimensioni del campione disponibile. La pratica migliore e quella di pianificare l'analisi in sottogruppi prima dell'inizio dello studio[84].
Nell'esempio, gli autori riportano stime separate per i pazienti con displasia fibromuscolare. Non hanno specificato se avevano progettato questa suddivisione in sottogruppi prima della raccolta dei dati.

PUNTO 24. INDICARE LE STIME DELLA RIPRODUCIBILITÀ DELL'ESAME (SE EFFETTUATE)

Esempio
La variabilità degli osservatori nel graduare le lesioni stenotiche dell'arteria renale (gradi da 1 a 4) con l'angiografia tradizionale-MRA era identica, con un valore k di 0.77 ed intervalli di confidenza al 90% tra 0.67 e 0.86. Per l'individuazione di lesioni significative dal punto vista emodinamico, la variabilità tra osservatori era 0.87 (0.78-0.95) per l'MRA e 0.88 (0.79-0.97) per l'angiografia convenzionale[50].

Raccomandiamo che gli autori riportino tutte le misure di riproducibilità dell'esame che hanno eseguito durante lo studio (vedi punto 13). Per i metodi analitici quantitativi, riportare il coefficiente di variazione (CV) alle concentrazioni rilevanti per lo studio, definire tali concentrazioni ed il numero di determinazione (per CV intra-dosaggio, se rilevante) o il numero di giorni di analisi (per CV tra giorni, totale) o entrambi.

PUNTO 25. DISCUTERE L'APPLICABILITÀ CLINICA DEI RISULTATI DELLO STUDIO

Esempio
Anche se sono stati pubblicati molti articoli sul peptide natriuretico cerebrale in gruppi selezionati di pazienti, questi sono i primi dati sulle caratteristiche di prestazioni di un dosaggio di NT-proBNP in una ampia serie generalizzabile di adulti selezionati in modo casuale con una diagnosi confermata di insufficienza cardiaca e con una popolazione di confronto selezionata in modo causale dalla stessa popolazione. (…) Questi dati suggeriscono che nella pratica clinica il dosaggio ha tre impieghi: screening di pazienti precedentemente classificati come affetti da insufficienza cardiaca (in 79 dei 103 pazienti classificati in questo modo l'insufficienza car-

diaca è stata esclusa con la determinazione dell'NT-proBNP); triage per ecocardiogramma dei pazienti che si presentano con sintomi suggestivi di insufficienza cardiaca (difficoltà respiratorie, letargia) screening di pazienti a rischio elevato di insufficienza cardiaca. Riteniamo che il dosaggio dia buoni risultati in questi ambiti ma la prima indicazione non era stata verificata in modo formale in questo studio e la terza indicazione è stata valutata solo in 134 pazienti [85].

A causa della variabilità delle caratteristiche dell'esame dovute a differenze nel disegno, nei pazienti e nelle procedure, i risultati di un particolare studio possono non essere applicabili al problema decisionale di interesse per i lettori[13].

Oltre a discutere i potenziali limiti metodologici dello studio ed una interpretazione generale dei risultati nel contesto delle prove attualmente disponibili, raccomandiamo che gli autori sottolineino le differenza tra il contesto dello studio e le altre sedi e gli altri gruppi di pazienti in cui l'esame sarà probabilmente impiegato.

Commenti

Sappiamo che gli studi di accuratezza diagnostica non costituiscono l'unico tipo di studi per valutare esami diagnostici. E' impiegato un ampio spettro di altri disegni, compresi i trial clinici randomizzati[2].

La metodologia per il disegno e la conduzione di studi di accuratezza diagnostica sono ancora in fase di maturazione. La nostra comprensione delle fonti di variabilità e del potenziale di accuratezza diagnostica sta crescendo. Prevediamo quindi di aggiornare periodicamente la lista di controllo STARD.

Gli esami diagnostici costituiscono una parte essenziale della medicina. Articoli completi ed informativi possono solo portare e migliori decisioni in sanità.

Bibliografia

1. Guyatt GH, Tugwell PX, Feeny DH, Haynes RB, Drummond M. A framework for clinical evaluation of diagnostic technologies. CMAJ 1986;134:587–94.
2. Sackett DL, Haynes RB. The architecture of diagnostic research. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:19–38.
3. Jaeschke R, Guyatt G, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. A. Are the results of the study valid? Evidence-Based Medicine Working Group. JAMA 1994;271:389–91.
4. Jaeschke R, Guyatt GH, Sackett DL. Users' guides to the medical literature. III. How to use an article about a diagnostic test. B. What are the results and will they help me in caring for my patients? The Evidence-Based Medicine Working Group. JAMA 1994;271:703–7.
5. Mower WR. Evaluating bias and variability in diagnostic test reports. Ann Emerg Med 1999;33:85–91.
6. Begg CB. Biases in the assessment of diagnostic tests. Stat Med 1987;6:411–23.
7. Sheps SB, Schechter MT. The assessment of diagnostic tests. A survey of current medical research. JAMA 1984;252:2418–22.
8. Reid MC, Lachs MS, Feinstein AR. Use of methodological standards in diagnostic test research. Getting better but still not good. JAMA 1995;274:645–51.
9. Lijmer JG, Mol BW, Heisterkamp S, et al. Empirical evidence of design-related bias in studies of diagnostic tests. JAMA 1999; 282:1061–6.
10. Knottnerus JA, Muris JW. Assessment of the accuracy of diagnostic tests: the cross-sectional study. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:39–59.
11. Griner PF, Mayewski RJ, Mushlin AI, Greenland P. Selection and interpretation of diagnostic tests and procedures. Principles and applications. Ann Intern Med 1981;94:557–92.
12. Habbema JDF, Eijkemans R, Krijnen P, Knottnerus JA. Analysis of data on the accuracy of diagnostic tests. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:117–44.
13. Irwig LM, Bossuyt PM, Glasziou PP, Gatsonis C, Lijmer JG. Designing studies to ensure that estimates of test accuracy will travel. In: Knottnerus JA, ed. The evidence base of clinical diagnosis. London: BMJ Publishing Group, 2002:95–116.
14. Yee J, Akerkar GA, Hung RK, Steinauer-Gebauer AM, Wall SD, McQuaid KR. Colorectal neoplasia: performance characteristics of CT colonography for detection in 300 patients. Radiology 2001; 219:685–92.
15. Haynes RB, Wilczynski N, McKibbon KA, Walker CJ, Sinclair JC. Developing optimal search strategies for detecting clinically sound studies in MEDLINE. J Am Med Inform Assoc 1994;1:447–58.
16. Dickersin K, Scherer R, Lefebvre C. Identifying relevant studies for systematic reviews. BMJ 1994;309:1286–91.
17. McKibbon KA, Walker-Dilks CJ. Beyond ACP Journal Club: how to harness MEDLINE for diagnostic problems. ACP J Club 1994; 121(Suppl 2):A10–2.
18. Deville WL, Bezemer PD, Bouter LM. Publications on diagnostic test evaluation in family medicine journals: an optimal search strategy. J Clin Epidemiol 2000;53:65–9.
19. Wilczynski NL, McKibbon KA, Haynes RB. Enhancing retrieval of best evidence for health care from bibliographic databases: calibration of the hand search of the literature. Medinfo 2001;10:390–3.
20. Kim WY, Danias PG, Stuber M, Flamm SD, Plein S, Nagel E, et al. Coronary magnetic resonance angiography for the detection of coronary stenoses. N Engl J Med 2001;345:1863–9.
21. World Medical Association Declaration of Helsinki. Recommendations guiding physicians in biomedical research involving human subjects. JAMA 1997;277:925–6.
22. Newhall WJ, Johnson RE, DeLisle S, Fine D, Hadgu A, Matsuda B, et al. Head-to-head evaluation of five chlamydia tests relative to a quality-assured culture standard. J Clin Microbiol 1999;37: 681–5.
23. Philbrick JT, Horwitz RI, Feinstein AR. Methodologic

problems of exercise testing for coronary artery disease: groups, analysis and bias. Am J Cardiol 1980;46:807–12.
24. Detrano R, Gianrossi R, Froelicher V. The diagnostic accuracy of the exercise electrocardiogram: a meta-analysis of 22 years of research. Prog Cardiovasc Dis 1989;32:173–206.
25. Stein PD, Gottschalk A, Henry JW, Shivkumar K. Stratification of patients according to prior cardiopulmonary disease and probability assessment based on the number of mismatched segmental equivalent perfusion defects. Approaches to strengthen the diagnostic value of ventilation/perfusion lung scans in acute pulmonary embolism. Chest 1993;104:1461–7.
26. Knottnerus JA, Knipschild PG, Sturmans F. Symptoms and selection bias: the influence of selection towards specialist care on the relationship between symptoms and diagnoses. Theor Med 1989; 10:67–81.
27. Knottnerus JA, Leffers P. The influence of referral patterns on the characteristics of diagnostic tests. J Clin Epidemiol 1992;45: 1143–54.
28. Melbye H, Straume B. The spectrum of patients strongly influences the usefulness of diagnostic tests for pneumonia. Scand J Prim Health Care 1993;11:241–6.
29. van der Schouw YT, Van Dijk R, Verbeek AL. Problems in selecting the adequate patient population from existing data files for assessment studies of new diagnostic tests. J Clin Epidemiol 1995;48:417–22.
30. Kline JA, Israel EG, Michelson EA, O'Neil BJ, Plewa MC, Portelli DC. Diagnostic accuracy of a bedside D-dimer assay and alveolar dead-space measurement for rapid exclusion of pulmonary embolism: a multicenter study. JAMA 2001;285:761–8.
31. Vande Berg BC, Lecouvet FE, Poilvache P, Dubuc JE, Bedat B, Maldague B, et al. Dual-detector spiral CT arthrography of the knee: accuracy for detection of meniscal abnormalities and unstable meniscal tears. Radiology 2000;216:851–7.
32. Sorensen HT, Sabroe S, Olsen J. A framework for evaluation of secondary data sources for epidemiological research. Int J Epidemiol 1996;25:435–42.
33. Mayeux R, Saunders AM, Shea S, Mirra S, Evans D, Roses AD, et al. Utility of the apolipoprotein E genotype in the diagnosis of Alzheimer's disease. Alzheimer's Disease Centers Consortium on Apolipoprotein E and Alzheimer's Disease. N Engl J Med 1998; 338:506–11.
34. Mol BW, Lijmer JG, van der Meulen J, Pajkrt E, Bilardo CM, Bossuyt PM. Effect of study design on the association between nuchal translucency measurement and Down syndrome. Obstet Gynecol 1999;94:864–9.
35. Corley DE, Kirtland SH, Winterbauer RH, Hammar SP, Dail DH, Bauermeister DE, et al. Reproducibility of the histologic diagnosis of pneumonia among a panel of four pathologists: analysis of a gold standard. Chest 1997;112:458–65.
36. Magklara A, Scorilas A, Catalona WJ, Diamandis EP. The combination of human glandular kallikrein and free prostate-specific antigen (PSA) enhances discrimination between prostate cancer and benign prostatic hyperplasia in patients with moderately increased total PSA. Clin Chem 1999;45:1960–6.
37. Smith H, Pickering RM, Struthers A, Simpson I, Mant D. Biochemical diagnosis of ventricular dysfunction in elderly patients in general practice: observational study. BMJ 2000;320:906–8.
38. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130:515–24.
39. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15: 361–87.
40. Hartz AJ, Guse C, Kajdacsy-Balla A. Identification of heavy drinkers using a combination of laboratory tests. J Clin Epidemiol 1997; 50:1357–68.
41. Elmore JG, Feinstein AR. A bibliography of publications on observer variability (final installment). J Clin Epidemiol 1992;45: 567–80.
42. Brealey S, Scally AJ, Thomas NB. Review article: methodological standards in radiographer plain film reading performance studies. Br J Radiol 2002;75:107–13.
43. Elmore JG, Wells CK, Lee CH, Howard DH, Feinstein AR. Variability in radiologists' interpretations of mammograms. N Engl J Med 1994;331:1493–9.
44. Brealey S, Scally AJ. Bias in plain film reading performance studies. Br J Radiol 2001;74:307–16.
45. Ronco G, Montanari G, Aimone V, Parisio F, Segnan N, Valle A, et al. Estimating the sensitivity of cervical cytology: errors of interpretation and test limitations. Cytopathology 1996;7:151–8.
46. Cohen MB, Rodgers RP, Hales MS, Gonzales JM, Ljung BM, Beckstead JH, et al. Influence of training and experience in fine-needle aspiration biopsy of breast. Receiver operating characteristics curve analysis. Arch Pathol Lab Med 1987;111:518–20.
47. Egglin TK, Feinstein AR. Context bias. A problem in diagnostic radiology. JAMA 1996;276:1752–5.
48. Doubilet P, Herman PG. Interpretation of radiographs: effect of clinical history. AJR Am J Roentgenol 1981;137:1055–8.
49. Berbaum KS, Franken EA Jr, Dorfman DD, Barloon T, Ell SR, Lu CH, et al. Tentative diagnoses facilitate the detection of diverse lesions in chest radiographs. Invest Radiol 1986;21:532–9.
50. Leung DA, Hoffmann U, Pfammatter T, Hany TF, Rainoni L, Hilfiker P, et al. Magnetic resonance angiography versus duplex sonography for diagnosing renovascular disease. Hypertension 1999;33: 726–31.
51. Lang TA, Secic M. Generalizing from a sample to a population: reporting estimates and confidence intervals. In: Lang TA, Secic M, eds. How to report statistics in medicine: annotated guidelines for authors, editors, and reviewers. Philadelphia: American College of Physicians, 1997:55–63.
52. Quinn MF. Relation of observer agreement to accuracy according to a two-receiver signal detection model of diagnosis. Med Decis Making 1989;9:196–206.
53. DesJardin JA, Falagas ME, Ruthazer R, Griffith J, Wawrose D, Schenkein D, et al. Clinical utility of blood cultures drawn from indwelling central venous catheters in hospitalized patients with cancer. Ann Intern Med 1999;131:641–7.
54. Wells PS, Brill-Edwards P, Stevens P, Panju A, Patel A, Douketis J, et al. A novel and rapid whole-blood

assay for D-dimer in patients with clinically suspected deep vein thrombosis. Circulation 1995; 91:2184–7.
55. Garcia Pena BM, Mandl KD, Kraus SJ, Fischer AC, Fleisher GR, Lund DP, et al. Ultrasonography and limited computed tomography in the diagnosis and management of appendicitis in children. JAMA 1999;282:1041–6.
56. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. N Engl J Med 1978; 299:926–30.
57. Choi BC. Sensitivity and specificity of a single diagnostic test in the presence of work-up bias. J Clin Epidemiol 1992;45:581–6.
58. Zhou XH. Effect of verification bias on positive and negative predictive values. Stat Med 1994;13:1737–45.
59. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983; 39:207–15.
60. Greenes RA, Begg CB. Assessment of diagnostic technologies. Methodology for unbiased estimation from samples of selectively verified patients. Invest Radiol 1985;20:751–6.
61. Cecil MP, Kosinski AS, Jones MT, Taylor A, Alazraki NP, Pettigrew RI, et al. The importance of work-up (verification) bias correction in assessing the accuracy of SPECT thallium-201 testing for the diagnosis of coronary artery disease. J Clin Epidemiol 1996;49: 735–42.
62. Diamond GA, Rozanski A, Forrester JS, Morris D, Pollock BH, Staniloff HM, et al. A model for assessing the sensitivity and specificity of tests subject to selection bias. Application to exercise radionuclide ventriculography for diagnosis of coronary artery disease. J Chronic Dis 1986;39:343–55.
63. Diamond GA. Off Bayes: effect of verification bias on posterior probabilities calculated using Bayes' theorem. Med Decis Making 1992;12:22–31.
64. Egger M, Juni P, Bartlett C. Value of flow diagrams in reports of randomized controlled trials. JAMA 2001;285:1996–9.
65. Fletcher RH. Carcinoembryonic antigen. Ann Intern Med 1986; 104:66–73.
66. Lachs MS, Nachamkin I, Edelstein PH, Goldman J, Feinstein AR, Schwartz JS. Spectrum bias in the evaluation of diagnostic tests: lessons from the rapid dipstick test for urinary tract infection. Ann Intern Med 1992;117:135–40.
67. O'Connor PW, Tansay CM, Detsky AS, Mushlin AI, Kucharczyk W. The effect of spectrum bias on the utility of magnetic resonance imaging and evoked potentials in the diagnosis of suspected multiple sclerosis. Neurology 1996;47:140–4.
68. Moons KG, van Es GA, Deckers JW, Habbema JD, Grobbee DE. Limitations of sensitivity, specificity, likelihood ratio, and Bayes' theorem in assessing diagnostic probabilities: a clinical example. Epidemiology 1997;8:12–7.
69. Hlatky MA, Pryor DB, Harrell FE Jr, Califf RM, Mark DB, Rosati RA. Factors affecting sensitivity and specificity of exercise electrocardiography. Multivariable analysis. Am J Med 1984;77:64–71.
70. Harris JM Jr. The hazards of bedside Bayes. JAMA 1981;246: 2602–5.
71. Philbrick JT, Horwitz RI, Feinstein AR, Langou RA, Chandler JP. The limited spectrum of patients studied in exercise test research. Analyzing the tip of the iceberg. JAMA 1982;248:2467–70.
72. Giard RW, Hermans J. Use and accuracy of fine-needle aspiration cytology in histologically proven thyroid carcinoma: an audit using a national nathology database. Cancer 2000;90:330–4.
73. Bizzaro N, Mazzanti G, Tonutti E, Villalta D, Tozzoli R. Diagnostic accuracy of the anti-citrulline antibody assay for rheumatoid arthritis. Clin Chem 2001;47:1089–93.
74. Bonnamy L, Marret H, Perrotin F, Body G, Berger C, Lansac J. Sonohysterography: a prospective survey of results and complications in 81 patients. Eur J Obstet Gynecol Reprod Biol 2002;102: 42–7.
75. Ioannidis JP, Lau J. Completeness of safety reporting in randomised trials: an evaluation of 7 medical areas. JAMA 2001;285: 437–43.
76. Towler BP, Irwig L, Glasziou P, Weller D, Kewenter J. Screening for colorectal cancer using the faecal occult blood test, hemoccult. Cochrane Database Syst Rev 2000:CD001216.
77. Anton RF, Dominick C, Bigelow M, Westby C. Comparison of Bio-Rad %CDT TIA and CDTect as laboratory markers of heavy alcohol use and their relationships with gamma –glutamyl-transferase. Clin Chem 2001;47:1769–75.
78. Harper R, Reeves B. Reporting of precision of estimates for diagnostic accuracy: a review. BMJ 1999;318:1322–3.
79. Begg CB, Greenes RA, Iglewicz B. The influence of uninterpretability on the assessment of diagnostic tests. J Chronic Dis 1986; 39:575–84.
80. Simel DL, Feussner JR, DeLong ER, Matchar DB. Intermediate, indeterminate, and uninterpretable diagnostic test results. Med Decis Making 1987;7:107–14.
81. Pisano ED, Fajardo LL, Tsimikas J, Sneige N, Frable WJ, Gatsonis CA, et al. Rate of insufficient samples for fine-needle aspiration for nonpalpable breast lesions in a multicenter clinical trial: The Radiologic Diagnostic Oncology Group 5 Study. The RDOG5 investigators. Cancer 1998;82:679–88.
82. Giard RW, Hermans J. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer 1992;69:2104–10.
83. Value of the ventilation/perfusion scan in acute pulmonary embolism. Results of the prospective investigation of pulmonary embolism diagnosis (PIOPED). The PIOPED Investigators. JAMA 1990; 263:2753–9.
84. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses. Ann Intern Med 1992; 116:78–84.
85. Hobbs FD, Davis RC, Roalfe AK, Hare R, Davies MK, Kenkre JE. Reliability of N-terminal pro-brain natriuretic peptide assay in diagnosis of heart failure: cohort study in representative and high risk community populations. BMJ 2002;324:1498–500.
86. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Ann Intern Med 2003;138:40–4.
87. Bossuyt PM, Reitsma JB, Bruns DE, Gatsonis CA, Glasziou PP, Irwig LM, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD Initiative. Clin Chem 2003;49:1–6.