Metodi statistici per la Neuropsicologia Forense

A.A. 2023/2024

Giorgio Arcara

IRCCS San Camillo, Venezia
Università degli Studi di Padova

4. Validità e Affidabilità

Validità ed Affidabilità

Validità e **Affidabilità** sono due qualità fondamentali dei test
(non solo didatticamente)

Nelle prossime slides farà una panoramica generale senza scendere
nei dettagli delle formule che vedremo in un passo successivo

Validità ed Affidabilità

Usare un test senza conoscerne validità e affidabilità può portare a **grossi errori**:

Per esempio usare un test con bassa affidabilità non valido per un'interpretazione e potrebbe portare a misurare l'altezza di una persona con una bilancia che ha un margine di errore di 10 kg.

Validità ed Affidabilità

La **validità** è la qualità di uno strumento di misurare effettivamente il costrutto che vuole misurare.

L'**affidabilità** indica la precisione con cui un test misura un certo costrutto.

Esistono diverse classificazioni di validità e affidabilità. Quelle che segue è solo una delle possibili (Urbina, 2004, Essentials of Psychological Testing).

Come vorremmo fossero validità e affidabilità (ma non sono)

Validità: il mio test misura il mio costrutto in maniera nota e quantificabile.
Es.
- l'80% del punteggio osservato è riconducibile al costrutto di interesse
- il punteggio riconducibile al costrutto di interesse mentre +/-5% ad altri costrutti.

Affidabilità: il mio test misura il mio costrutto con precisione quantificabile.
Es.
- il mio punteggio finale può essere sbagliato di +/-3
- il punteggio finale puà essere sbagliato di +/-5%

Purtroppo Validità e Affidabilità ci danno informazioni meno precise

Come vorremmo fossero validità e affidabilità (ma non sono)

Anche se ci sono dei criteri talvolta di valore minimo (di affidabilità o validità) per considerare un test adeguato, spesso questi valori ci aiutano a distinguere test tra di loro, per scegliere il migliore (o il meno peggiore)

Peggiore Validità e/o Affidabilità

Migliore Validità e/o Affidabilità

Peggio

Meglio

La Validità

La validità è la qualità di un test di misurare ciò che effettivamente vuole misurare. Il termine validità è sempre riferito ad un'**utilizzo** che si fa dei punteggi.

Non ha senso dire che **un test è valido**

La Validità

Un test può avere validità per una certa utilizzo o per più utilizzi.

La validità non è una qualità tutto-o-nulla, ma una qualità lungo un continuum

Ad esempio ci sono evidenze che il ”Free and Cued Selective Reminding Test-it” (Clerici et al., 2017) è un test con buona validità per misurare le capacità di memoria.

La Validità

- **Validità di contenuto**
- **Validità convergente/divergente (validità di costrutto)**
- **Validità di criterio**
- **Validità di facciata**
- **Validità ecologica**

Esistono diverse classificazioni di validità Questa che sto utilizzando è principalmente basata su Urbina, 2004, Essentials of Psychological Testing.

Tipi di Validità

Validità di contenuto : la proprietà degli item di essere sufficienti ed adeguati per valutare il costrutto di interesse. Può essere valutata qualitativamente o quantitativamente (Lawshe, 1975).

Spesso non è riportata nei test o c'è solo una valutazione qualitativa. Due test che riportano validità di contenuto quantitativamente sono Abaco (Sacco et al., 2008) e APACS (Arcara & Bambini, 2016), ma non usano statistiche inferenziali.

Tipi di Validità

Esempio di test con scarsa validità di contenuto:

Un test che vuole misurare l'abilità di lettura nella vita quotidiana e utilizza solamente parole in isolamento.

Esempio di test con buona validità di contenuto:

Un test che vuole misurare le abilità di guida. E simula le abilità di guida tramite un sorta di videogioco in numerose condizioni.

Tipi di Validità

Validità convergente-divergente (o validità di costrutto): indaga quanto il test misura effettivamente il costrutto che intende misurare valutando corenza interna degli item, oppure correlazione con altri test.
Si parla anche di validità "divergente" perché valuta anche la qualità di *non* correlare con test con cui non dovrebbe correlare (ad esempio un test specifico per memoria visuospaziale, dovrebbe non correlare troppo con memoria verbale).

A livello di analisi di dati.
Per la coerenza degli items di un test tra loro si usa l'analisi fattoriale o l'alpha di Cronbach (per quest'ultima analisi vedi sezione affidabilità).

Per la relazione dei punteggi totali con quelli di altri test si usano spesso correlazioni.

Non esistono delle regole su che valori (nell'analisi fattoriale, o nelle correlazioni) siano accettabili. Dipende dal costrutto misurato e dalle relazioni attese.

Tipi di Validità

Un test con buona validità di costrutto correla con test che misurano lo stesso costrutto o correla con test che misurano altri costrutti in maniera coerente con le aspettative.

Ad esempio, un test di memoria di lavoro con buona validità di costrutto, correla con altri test di memoria di lavoro. Un test di comprensione di linguaggio figurato (costrutto più ampio), dovrebbe correlare con altri test che misurano lo stesso costrutto, ma in maniera moderata, potrebbe correlare anche con attenzione o altre funzioni cognitive legate.

Tipi di Validità

La **validità di costrutto** è da un lato forse la validità più importante, ma dall'altro quella più difficile da valutare se adeguata quando si indaga in relazione ad altri test che misurano lo stesso costrutto. Questo perché non esistono indicazioni di "correlazione minima" che dovrebbe avere un test con un altro per avere supporto alla sua validità di costrutto.

Ad esempio: se sviluppo un nuovo test di memoria, non è detto che ci sia un valore minimo di correlazione con un altro test di memoria che sia considerato come sufficiente.

Per l'analisi fattoriale o il Cronbach's alpha esistono invece principi più chiari: l'analisi deve mostrare che gli item si comportano in maniera statisticamente adeguata relativamente al costrutto (questo sarà più chiaro nella sezione di approfondimento statistico di queste analisi).

Tipi di Validità

Validità di criterio: La proprietà di unt test di fornire risultati legati ad un criterio esterno. Quest criterio è spesso l'appartenenza ad una patologia, oppure un aspetto prognostico (lo sviluppare una patologia in futuro, il migliorare dopo un trattamento, etc.)

La validità di criterio è spesso espressa da valori di sensibilità/specificità.
In generale sono fornite percentuali che esprimono l'accuratezza del test nel predire il criterio.

Esistono diversi standard riportati per considerare accettabile la validità di criterio, ma dovrebbe essere alta (sens/spec superiore all'80%)

Tipi di Validità

Validità di facciata: si riferisce alla qualità di un test di esssere chiaramente riconducibile al costrutto che vuole misurare. Es. un test di memoria che richiede di ricordare gli items.

Data la natura prettamente qualitativa non esistono analisi per verificarla.

Tipi di Validità

Validità ecologica: si riferisce alla qualità di un test di riflettere effettivamente delle abilità che hanno ripercussioni nella vita quotidiana.
(Ad esempio, che la performance deficitaria ad un test di memoria rifletta difficoltà di memoria nella vita quotidiana).

Può essere valutata tramite diverse analisi, es. Correlazioni con altri questionari o scale riferite alla vita quotidiana.

Non esistono valori soglia condivisi che un test deve avere.

Nota che la validità ecologica è raramente disponibile vista la difficoltà nell'essere verificata.

Tipi di Validità

Nota che la **validità ecologica** è spesso trascurata ma particolarmente importante perché spesso nelle valutazioni sono fatte inferenze implicite sull'impatto del deficit sulla vita quotidiana.

Identificare un deficit di memoria è infatti relativamente importante: quello che è rilevante è spesso capire che impatto ha questo deficit nella vita del paziente

Questo è particolarmente rilevante nel caso di valutazioni del *danno* subito dal paziente. Avere un deficit cognitivo che sia particolarmente invalidante è ovviamente diverso da avere un deficit che però non ha impatto sulla vita quotidiana.

Un test con dati su validità ecologica ci aiuterebbe a fare queste interepretazioni perché ci sarebbe supporto scientifico sulla relazione tra performance al test e comportamento di vita quotidiana.

Alcuni aspetti importanti sulla Validità

- Per conoscere la validità di un test occorre documentarsi (manuale, articolo scientifico).
- Per comprendere gli effetti della validità occorre conoscere degli aspetti (veramente base) di statistica.
- In italia e nei test neuropsicologici c'è spesso poca attenzione alla validità. Pochi test riportano dati su validità. Il termine *validato* viene in certi casi usato per indicare che sono stati raccolti i dati normativi, un aspetto completamente diverso.

Validità ed Affidabilità

Cognitive screening test	N	Validity			
		CV	CrV	DV	FS
General					
In-person					
MMSE	6	□		□	
MoCA	6	■		□	■
ACE-R	3	■			
CDT	3	■			■
MODA	2	■			
ACE-III	1				
BNS	1	■	■		
HAMTS	1				
GPCOG	1	■			
Mini-Cog	1				
Qmci-i	1	■			
SCEB	1	■			
TICS	1		■		
TYM-I	1	■			
Remote					
Itel-MMSE	2	■	■		
I-TICS	1	■	■		
VMMSE	1				
Domain-specific					
FAB	3	■		■	■
SAND	2	■		■	
AQT	1				
ART	1				
I-AABT	1	■			

Aiello et al. (2021)

Validità nei test di Screening Italiani (Review Sistematica).

Alcuni aspetti importanti sulla Validità

- La validità non è qualità “immutabile” e data di un test, ma dipende da come lo si utilizza.

Se il test viene utilizzato in maniera inappropriata, allora può non diventare più valido.

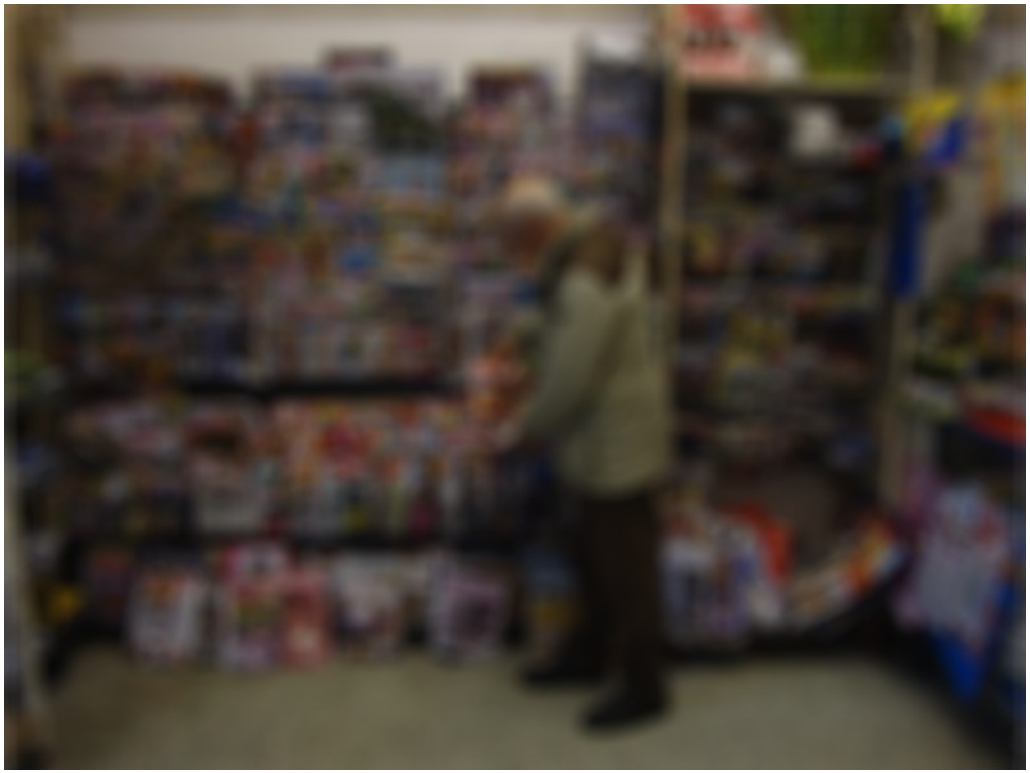
L’appropriatezza dell’utilizzo di un test è determinata dall’adesione alle procedure legate all’utilizzo del test stesso. Quanto più devio dalle procedure iniziali e quanto meno valido è il test.

Alcuni aspetti importanti sulla Validità

Esempio, il test di descrizione scene (APACS)

Alcuni aspetti importanti sulla Validità

Esempio, il test di descrizione scene (APACS)



Immaginate di stampare la foto con questa qualità di immagine

Alcuni aspetti importanti sulla Validità

Nell'esempio precedente qualcosa è cambiata nella procedura (i.e. la qualità della foto).
Non è più usato lo stimolo simile a quello usato originariamente e per questo
Il test, verosimilmente, non misura più ciò che intendeva misurare.

Il test con stimoli così , da un test per valutare le abilità di comunicare e descrivere scene semplici, diventa un test di acuità visiva.

Esempio 1

Per che cosa è valido il MoCA?

Spesso la tendenza è di giudicare la validità di un test solo sull'intuizione
o sul nome del test.

Nel paper originale Nasreddine et al., 2005.
Il MoCA era valido per discriminare tra MCI, AD e, controlli.

Anche il MMSE (Folstein et al., 1975), nasceva per distinguere Demenze da Depressione.

Esempio 2

Per che cosa è valida la FAB?

Finally, reliabilities were computed by means of Pearson's correlation coefficient. On a subset of 56 subjects, FAB performance was rated by two independent examiners; the inter-rater reliability was r=0.96 (df=54, p<0.001); on a different subset of 45 subjects, the FAB was repeated after 2–4 weeks; the test–retest reliability was r=0.85 (df=43, p<0.001).

The MMSE raw and adjusted mean scores were 29.0±1.3 (range 23–30) and 29.3±1.2 (range 24–30), respectively; interestingly, correlations of the FAB raw scores with MMSE raw and adjusted scores were 0.41 (p<0.001) and 0.09 (p=ns), respectively; correlations of the FAB adjusted scores were in the same direction (r=0.23, p<0.001 and r=0.10, p=ns, respectively).

(Da Appollonio et al. 2005)

Nell'articolo originale in realtà non ci sono effettive prove di validità Questo non significa che il FAB-it non abbia prove di validità per misurare funzioni esecutive(le versioni di altre lingue del FAB le hanno), ma che di fatto non ci sono prove per quella italiana.

Esempio 3

Validità di un test

Il test AbAco per le valutazioni di capacità comunicative è un esempio di eccellente test riguardo validità.

Esempio

Validità del test Abaco

Sacco 2008

5.3 Validation of the battery

Some questions must be answered when using a new clinical tool. The first main question concerns the reliability of the instrument. In particular:

(a) Is each scale/subscale of the Battery composed of congruent items, i.e. items that are intrinsically related to one another? This question concerns *internal consistency* and was answered by calculating the cohesion within each subscale (Cronbach alpha).

(b) Is the scoring system sufficiently clear and objective to be used by any trained examiner/rater? This question concerns *inter-rater reliability* and a measure of agreement between the ratings given by two independent judges (Cohen's kappa) was computed to satisfy such a goal.

The second main question concerns the validity of the instrument. In particular:

(c) At item level, do the test questions match the test objectives, i.e. does their content precisely address the subject area they are intended to assess? (b) Are the items appropriate for the age group of the subjects the instrument is intended for? In the case of our battery, are the contents of the items suitable for both developing children and adult subjects? These two questions concern *content validity*, and were answered through item evaluation by independent pragmatic experts.

(d) At a more general level, is the instrument actually measuring what it is assumed to measure, i.e. are the five scales of the battery referable to pragmatic abilities? This question concerns the *construct validity*, and was dealt with by computing a factor analysis.

L'Affidabilità

L'affidabilità (o **Attendibilità**) è la qualità di un test di fornire punteggi consistenti in diverse misurazioni, e può essere quindi intesa come la precisione di un test.

L'Affidabilità

- **Affidabilità inter-rater**
- **Affidabilità test-retest**
- **Consistenza Interna**

Esistono diverse classificazioni di affidabilità. Questa che sto utilizzando è principalmente basata su Urbina, 2004, Essentials of Psychological Testing.

Tipi di Affidabilità

Affidabilità inter-rater: è una misura della consistenza con cui diversi esaminatori valutano la stessa prestazione dello stesso paziente. É legata alla chiarezza istruzioni su come attribuire i punteggi e alla complessità dei comportamenti osservati nel test.:

In genere è espressa da un coefficiente, *l'intraclass correlation*, che varia tra 0 e 1, dove 0 indica completa inconsistenza tra i punteggi e 1 indica assoluta consistenza tra i punteggi di due o più esaminatori.

valori maggiori a 0.60 sono desiderabili

Tipi di Affidabilità

Affidabilità test-retest: rappresenta la correlazione di due misure con lo stesso test effettuato dallo stesso esaminatore e sullo stesso individuo dopo un intervallo di tempo *in cui si assume che non sia avvenuto nessun cambiamento*.

In genere è espressa da un coefficiente di correlazione che varia tra -1 e 1,

valori maggiori a 0.70 sono desiderabili

Tipi di Affidabilità

Affidabilità test-retest: rappresenta la correlazione di due misure con lo stesso test effettuato dallo stesso esaminatore e sullo stesso individuo dopo un intervallo di tempo *in cui si assume che non sia avvenuto nessun cambiamento*.

In genere è espressa da un coefficiente di correlazione che varia tra -1 e 1,

valori maggiori a 0.70 sono desiderabili

Tipi di Affidabilità

Affidabilità test-retest: nota che l'affidabilità test-retest è sempre valutate con uno specifico intervallo (es. 1 mese). Per tale ragioni esistono infinite possibili affidabilità test-retest, dal momento che questo valore potrebbe variare a seconda dell'intervallo.

Generalmente. Più corto è l'intervallo, più è probabile sia alta l'affidabilità test-retest.

Tipi di Affidabilità

L' **Affidabilità test-retest** NON indica la stabilità dei punteggi nel tempo (non necessariamente).

Punteggi potrebbero avere una affidabilità test-retest (correlazione) di 0.98, ma essere poco stabili perché soggetti a variazioni *sistematiche*.

L'affidabilità test retest non tiene conto dell'**effetto pratica** (trattata in slides successive), visto che assume (anche matematicamente) che non ci siano cambiamenti nel tempo. Questo aspetto sarà chiaro, studiando la formula con cui è calcolato il test-retest, sia tramite simulazioni.

Tipi di Affidabilità

Consistenza interna: rappresenta la consistenza tra gli item di un test

In genere è calcolata tramite l'alpha di Cronbach, un indice che varia da 0 a 1.
Ci sono altri metodi (es. split-half reliability), che misurano la stessa cosa secondo una prospettiva diversa.

valori maggiori a 0.70 sono desiderabili, ma spesso in test neuropsicologici sono intorno a 0.60.

Considerazione su Affidabilità

Come per la validità anche l'affidabilità di una specifica misurazione può cambiare.

Se cambiano procedure (o se ci sono interferenze) l'affidabilità può essere inferiore.

Validità ed Affidabilità

Relazioni possibili tra Validità e Affidabilità

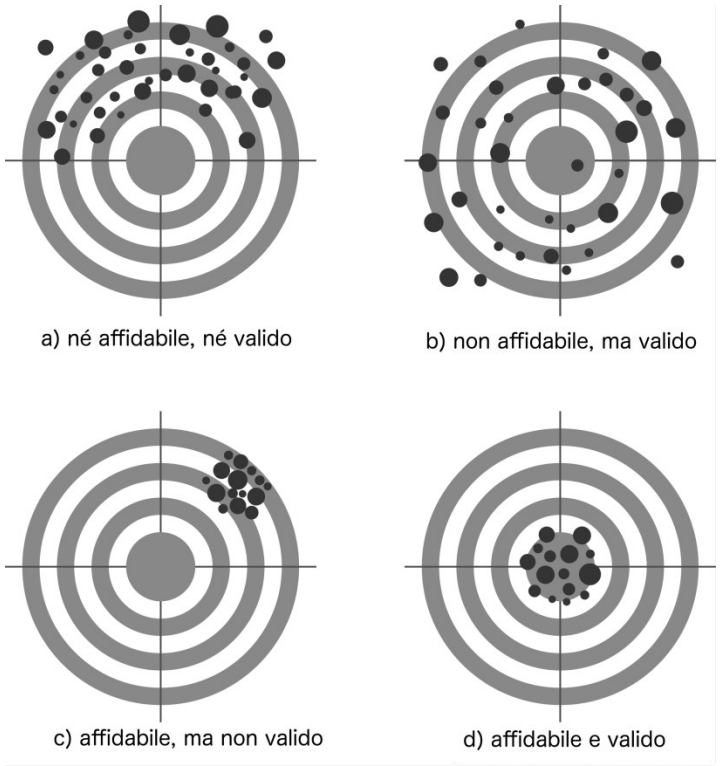


Figura adattata da *"Reliability and validity" by © Nevit Dilmen. Licensed under CC BY-SA 3.0 via Commons*.

Validità ed Affidabilità

L’Affidabilità non è solo qualche numero indicato nei manuali dei test, ma dei valori che hanno un impatto sulle conclusioni cliniche tratte dai test.

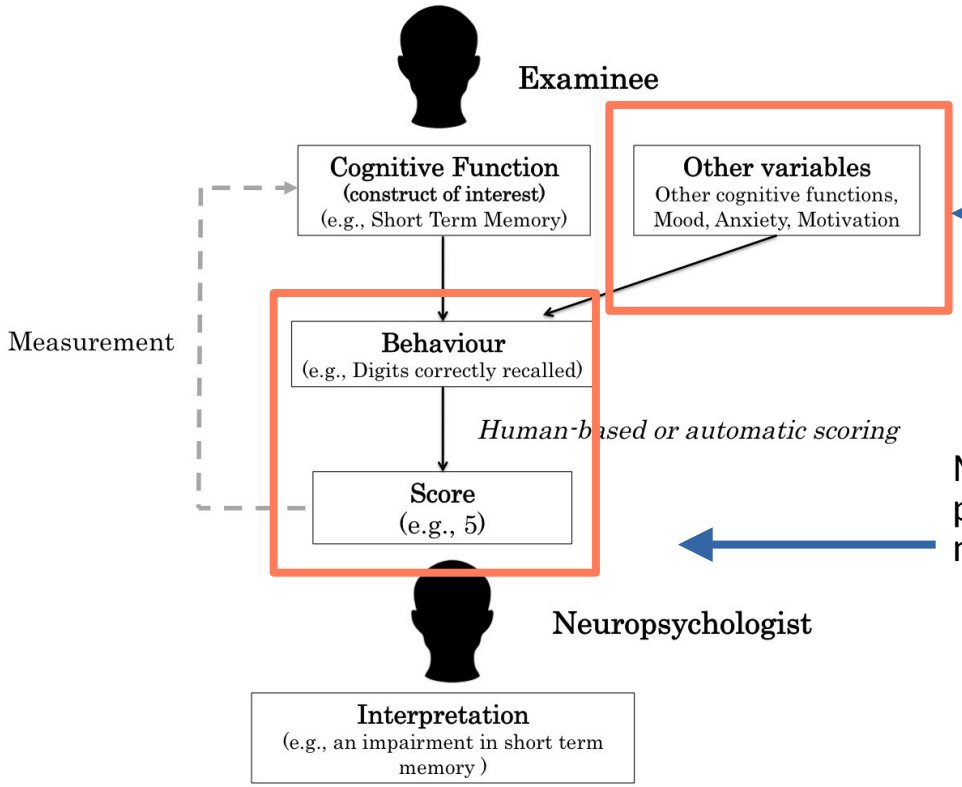
Validità ed Affidabilità

Il fatto che un test sia stato pubblicato non ci assicura che sia validato per qualche ragione e abbia affidabilità sufficiente!

(Disclaimer: Anche test pubblicati da me hanno bassa affidabilità, usare con cautela)

Es. Da Spinnler & Tognoni, 1987

- Matrici progressive di raven. Affidabilità = 0.89
- Orientamento. Affidabilità = 0.37



Minacce alla **Validità** (il test non misura quello che dovrebbe misurare)

Minacce alla **Affidabilità** (il punteggio del test oscilla molto).

NOTA: in questa slide parliamo di aspetti della qualità *del test*, non di una specifica misurazione. In slide future tratteremo il concetto di *validità* di una specifica istanza misurazione (cioè di una specifica valutazione di uno specifico individuo).

Validità ed Affidabilità

Affidabilità e validità sono due qualità importanti che ci permettono di valutare un test.

In genereale dovremmo scegliere test con elevata affidabilità e validità per le interpretazioni che ci servono

Queste due qualità non ci dicono però necessariamente, **l'utilità clinica/forense del test.**

Un test potrebbe essere eccellente per misurare un certo costrutto, ma poi, in fin dei conti, non aiutarci nella diagnosi dei disturbi del paziente o lo stabilire il piano riabilitativo, o rispondere alla nostra specifica domanda forense.