Do physicians Really understand diagnostic tests

Edward A. Panacek, MD, MPH
Department of Emergency Medicine
USA Medical Center, Mobile, AL

NDAFP, Big Sky, Montana, 2016
(Syllabus)

This session will cover 5 issues:

1. Critical "thresholds" in testing and decision making
2. The REAL purpose of diagnostic tests
3. Limitation of Sensitivity and specificity
4. Why PPV and NPV predictive values are limited
5. How Likelihood ratios, computers and better educated clinicians are the future.

Question

Why do clinicians, sometimes of equivalent training and experience, vary so much in their test ordering and clinical decision making patterns?

Answer:

A major reason is different clinical decision "thresholds"

This phenomenon has been noted for some time

- Green SM, Rothrock SG. Ann Emerg Med. 1999 Feb;33(2):211-4.
  Evaluation styles for well-appearing febrile children: "risk-minimizers" versus "delay-minimizers"?
- Graham T. Can Fam Physician. 2000 Jan;46:29-30, 33-5.
  Are you a "risk-minimizer" or a "test-minimizer"?

Critical testing thresholds

- Vary by clinician
  - Test avoiders
    - Tend to have higher RVUs, faster pt care dispositions
  - Risk Avoiders
    - Tend to have less QA issues
- Vary with recent experience
- Varies by patient type
  - CYA with some patients more than others
- Varies with the complaint (& the Diff. Dx)
  - Life threatening vs. benign diseases

"Risk avoiders" versus "test avoiders" are prone to different criticisms

Speeding tickets

Versus

Parking tickets

What is the purpose of diagnostic testing?

Most physicians would say:

"To definitively rule in some diagnosis, or to rule it out"

Let's look at diagnostic tests more closely

It may surprise you

Question:

Which test performance parameters are best for clinicians to use?

Sensitivity & specificity ?

Positive and negative predictive values ?

Relative risks or Odds Ratios

Something else?

Answer:

It depends.

But….

….it may not be what you think,

…or are currently using

Let's start with a review

Refresh our memories about some basic concepts

Diagnostic test performance measures

- **Dichotomous data**
  - **Sensitivity and specificity**
  - **Positive and negative predictive values**
  - **Positive and negative Likelihood Ratios**
- **Continuous data**
  - **Use test cut-off points to dichotomize**
  - **Interval Likelihood Ratios**
  - **ROC curves**

Standard R x C table organization for performance of dichotomous tests

	Disease (+)	Disease (-)
Test result (+)	a	b
Test result (-)	c	d

True & false positives & negatives

Each cell can also be classified as a true or false positive or negative test result

	Disease (+)	Disease (-)
Test (+)	TP	FP
Test (-)	FN	TN

Sensitivity and specificity

- **The most commonly used test performance parameters**
- **Considered to be "fixed" test properties**
- **Poorly understood and often misapplied**

Sensitivity

Definition: of subjects who truly have the disease, the proportion who test positive for the disease

- = TP / [TP+FN]

= a / (a+c)

	Disease (+)	Disease (-)
Test (+)	a	b
Test (-)	c	d

Specificity

Definition: of those who truly do not have the disease, the proportion that test negative

- = TN / [TN+FP]

= d / (b+d)

	Disease (+)	Disease (-)
Test (+)	a	b
Test (-)	c	d

Three problems with sensitivity and specificity

- Despite what is commonly believed, they are not well oriented to clinical decision making
- Misunderstood and misapplied by most physicians
- Not absolutely "fixed" test properties

Problems of spectrum bias in evaluating the accuracy of dx tests

Ransohoff DF, Feinstein AR. NEJM.1978;299:926.

- Sensitivity and specificity considered to be "fixed" test properties
- However, diagnostic tests often perform differently in actual practice than when first studied
- Sensitivity and specificity shown to change
- Problem is sample of patients in which first studied
  - Often pts with classic or severe disease
  - → bias (spectrum bias), ↑test performance

Most physicians cannot correctly describe what kind of test *(highly sensitive or highly specific)* they want to use to rule-in, versus rule out, a diagnosis

It can be confusing

Properly applying sensitivity and specificity (SPIN & SNOUT)

Which to use to rule-in vs out disease?

- *SNOUT* = sensitivity to rule *OUT* disease
  - Goal is to have no false negatives
  - = everyone with dz tests positive
- *SPIN* = specificity to rule *IN* disease
  - Goal of no false positives
  - = everyone without disease tests negative

SPIN versus SNOUT examples

- Imagine you want to rule out cancer via a screening program
  - You want a test that is highly sensitive
  - You don't want any false negatives
  - You care less about false positives, because they can undergo further testing
- Imagine you want to be certain about the Dx of brain CA, to rule it in, before surgery, etc.
  - You don't want any false positives, so want high specificity

Example:
Question: Is a D-dimer more useful to rule-in or rule out disease?

It's sensitivity is much better that it's specificity, so

SNOUT

Sensitivity and specificity problem: Not well suited to clinical care decisions

- **Orientation of these parameters is the opposite of clinical decision making !**
- **The definition of sensitivity :** Given that a pt has a disease, what is the probability that the test will be positive
- **The definition of specificity :** Given that a pt does not have a disease, what is the probability that the test will be negative
- **Clinicians have the opposite questions !**

Using sensitivity and specificity in actual practice

- **Most clinicians, intuitively, can accurately apply sensitivity and specificity information in clinical practice only when they approach 100% for a given test**
- **At lower %'s, most clinicians have a poor understanding of how diagnostic tests change the probability of disease or how they should apply the results in an individual patient**

Let us examine another test performance parameter

Predictive values

Positive

Negative

A Clinical Question to Consider

- **"NASTY" is a very bad disease, 99% fatal within 3 years.**
  - **There is one treatment available, but it is quite toxic and it alone causes a 10% mortality rate in 6 months**
- **There is one test for NASTY and it has a sensitivity of .99 and a specificity of .99**
- **You have just tested positive for the disease**
- **What do you do?**
- **Do you take the treatment or your chances?**

Predictive values

- **Seem to tell us exactly what we want to know clinically in such situations**
- **Positive predictive value (PPV)**
  - **proportion of subjects who test positive who truly do have the disease**
- **Negative predictive value (NPV)**
  - **proportion of subjects who test negative who truly do not have the disease**

2 x 2 table Dx test parameters

	Dz (+)	Dz (-)
Test (+)	a	b
Test (-)	c	d

- **Sensitivity = a/ (a+c)**
- **Specificity = d/ (b+d)**
- **PPV = a/ (a+b)**
- **NPV = d/ (c+d)**
- **True positives = a**
- **True negatives = d**
- **False positives = b**
- **False negatives = c**

Effects of prevalence on test parameter performance properties

- Sensitivity and specificity are not generally affected by prevalence
  - they are considered relatively "fixed" properties of test performance
- PPV and NPV are highly unstable parameters
  - directly affected by changes in prevalence in the population studied

Example of the effect of prevalence on PPV and NPV

- You have tested positive for "NASTY" disease which has a 100% 3 year mortality
- The only treatment is very toxic and carries its own 10% mortality in 6 months
- The test for NASTY is 99% sensitive and 99% specific
- However, NASTY is a rare disease that has a prevalence in the population of only 1/100,000

PPV of NASTY test with prevalence of 1/100,000*

	Disease (+)	Disease (-)	marginals
Test (+)	10	9999	10,009
Test (-)	0	989,991	989,991
marginals	10	999,990	1,000,000

PPV = 10/10,009

.001

= 0.1%

NPV = 989,990

.999

* Test = 99% sensitive and 99% specific

PPV of NASTY test with a disease prevalence of 1/100 (1%)*

	Disease (+)	Disease (-)	marginals
Test (+)	9900	9900	19,800
Test (-)	100	980,100	980,200
marginals	10,000	990,000	1,000,000

PPV = 9900/19800

= .5

NPV

=980100/980,200

= .9998

* Test = 99% sensitive and 99% specific

NASTY test with prevalence of 1/10 (10%)

	Disease (+)	Disease (-)	marginals
Test (+)	99	9	108
Test (-)	1	891	892
marginals	100	900	1,000

PPV = 99/108

= .9

NPV

= 891/892

= .99

Question:

Are predictive values the test performance parameter that most help us make better clinical decisions at the bedside?

Answer:

No, unfortunately not.

They are too unstable

They are also poorly understood by clinicians.

They do demonstrate why it is inappropriate to apply diagnostic tests in very low risk populations.

Most of the positive results will be false positives.

Likelihood Ratios !

- Combine the conventional fixed test properties of sensitivity and specificity into a summary index measure of the test
- Are the most clinically useful parameter of diagnostic test performance
  - Can be applied individually to any given patient
- Problem is they are not well understood
  - Can be cumbersome to use at the bedside

Likelihood ratios in clinical use

- LRs can be used for any test that has known sensitivity and specificity
- They start with the probability of the disease/diagnosis in a given pt and calculate the new probability after the test results are known.

The challenge with applying LRs at the bedside: Historically too complicated

Calculating LRs at the bedside was a complicated, multi-step process:

- Establish pre-test probabilities
- Convert to pre-test odds
- Calculate the LR value
- Multiply by the appropriate LR
- Get post-test odds
- Convert to post-test probability of disease

LR shortcuts

- Smart phone calculators
- Web site calculators
  - McMaster site
  - EBM sites
- Fagan Nomogram
- Jumping probability categories
  - rather than using exact %
- Future: Clinical decision analysis computer programs

To properly review LRs would take an entire lecture

I only have time to briefly mention them…. And give an example

Clinical decision problem example

- The resident is seeing a young adult woman with pleuritic CP, but no DVT risks or findings
- The ECG, CXR and exam are all benign
- The HO wants to send a D-dimer to further r/o a PE
- The attending says no, that test does not have a high enough sensitivity for PE and he could still be missing 10-20% of the PE cases (i.e. would be sending up to 20% of the PE cases home)
  - *The attending says "do a spiral CT or do nothing"*

The pleuritic CP patient problem

- The HO says that he feels he needs to do some study and would at least feel better if he knew the D-dimer was negative
- The attending says that would be worthless and that a more definitive test with higher sensitivity is needed…
  "get a spiral CT scan or do nothing"
- They come to you to for your opinion of what to do…….

…What is the argument for further testing?
What is the argument for no further testing?
….So, what do you recommend, and why?

Case of the adult with pleuritic chest pain

- The HO already had sent the D-dimer (Simpli-Red)
- The result is negative
- You all sit down and discuss where you are now in terms of needing to rule out a PE in this pt
- First, you agree the test has a sensitivity of at least .90 and spec of .60 for disease
- There is some disagreement on the patient's pre-test probability of disease.
  - Resident says it is low probability, "maybe 5%"
  - Attending says maybe intermediate, as high as 20%
- That was the pre-test probability, now what is the post-test probability?

Post test probability

- HO believes the pre-test Pr = 5% (.05)
- LR(-) for a negative D-dimer =
  (1-sens)/spec= (1-.9)/.6 =.1/.6 = .16
  - *<Fagan nomogram or calculator used>*
- post-test Pr = 0.8%
- So, after the D-dimer, the (HO) patients probability (risk) of PE is <u>now less than 1%</u>
  - i.e. the pt dropped from low, to a very low probability category

Post test probability: Attending

- However the attending believes the pre-test Pr = 20% (.20)
- LR(-) for a negative D-dimer =
  (1-sens)/spec =.10/.6=.16
  - *<Fagan nomogram or calculator used>*
- post-test Pr = .04 = 4%
- So, even with the much higher pre-test Pr, the patients probability (risk) of PE is now only 4%
  - i.e. it dropped the pt from a moderate risk to a low risk category

Using LRs in combination

- BTW, the HO also sent an ABG on the pt thinking that might be useful.
  - The attending says *"what a worthless test for PE!"*
  - It came back entirely normal.
- You do a lit search, and conclude that the ABG is 85% sensitive and 60% specific for PE
- Now what is the patient's post-test probability of disease?
- You start with the new Pr, after the D-dimer

Post-post test probability: HO

- HO current pre-test Pr is = 1% (.01)
- LR(-) for a negative ABG = (1-sens)/spec =.15/.6=.25
  - <use Fagan nomogram>
- post-test Pr = .0025 = 0.25%
- *So, after the normal (negative) ABG, the patients post-test probability (risk) of PE is now about 0.25% (1 in 400)*

Post-post test probability: Attending

- Attending current pre-test Pr = 4% (.04)
- LR(-) for a negative D-dimer = (1-sens)/spec =.15/.6 = .25
  - *<use the Fagan nomogram>*
- post-test Pr = .01 = 1%
- *So, even with the higher pre-test Pr, the patients probability (risk) of PE is now at or under 1%*

This is what diagnostic tests <u>really</u> do. They keep revising the probability of disease.

Physicians usually order tests until there is enough data to reach one of their critical thresholds.

This is *Bayesian* thinking and physicians are doing it constantly, whether they realize it or not.

LRs formally quantitate and make conscious…….a (Baysian) process that goes on unconsciously in most physicians brains

Some physicians are more aware of this than others

Dx testing in perspective

- Dx tests =Tools we use to help evaluate patients
  - Usually are not absolutely diagnostic
  - Physicians vary in critical decision thresholds
- Best applied when we ***<u>really</u>*** understand them
  - including their limitations
- Physician evaluations are "Bayesian", incorporating all data and revising probabilities
  - Diagnostic test use should be Bayesian also
- Ordering more tests is not always helpful or safe
- Even great tests can do more harm than good if used inappropriately in very low risk populations

There is a great deal of unnecessary diagnostic testing.

So what…. ?

Let me leave you with a story…

The hazards of performing tests in very low risk individuals: *A true story !*

- Healthy young adult (emergency physician)
- @ Hospital charity event dinner
- Won a free whole body screening CT scan
- That study found a pulmonary nodule
- Probably benign, but she favored a biopsy
  - Guided biopsy performed
- Sat up after study, → air embolism → CNS
- Major CVA, brain swelling, brain death

The end