PubH 7470: STATISTICS

FOR TRANSLATIONAL & CLINICAL RESEARCH

DIAGNOSTICS:

SOME FUNDAMENTAL ISSUES

THE DEVELOPMENTAL STAGE

- In the Developmental Stage, the basic question is: Does the idea work? It's the investigator's burden to prove to public or regulatory agencies.
- Approach: Trying the test's idea on a "pilot population" where one compares the test results versus truth; we have data.

KEY PARAMETERS

- **Two parameters:**
  Sensitivity, $S^+ = Pr(T=+|D+)$
  Specificity, $S^- = Pr(T=-|D=-)$
- **Sensitivity is the probability to correctly identify a diseased individual and Specificity the probability of correctly identify a healthy individual**
- **At the present time, mammography is about 96.6% specific and 64.7% sensitive.**

Since all cancers, for example, are "rare" (low prevalence and incidence), specificity - **not** sensitivity - has its dominating effect on the "**response/test rate**", $\pi_t = \Pr(T=+)$.

$$\pi_t = \Pr(T=+) = \Pr(T=+, D=+) + \Pr(T=+, D=-)$$

$$\pi_t = \Pr(T=+ \mid D=+)\Pr(D=+) + \Pr(T=+ \mid D=-)\Pr(D=-)$$

$$\pi_t = S^+\pi + (1-S^-)(1-\pi)$$

Why Specificity more dominating?

Reason: larger coefficient, $(1-\pi)$ versus π.

EXAMPLE #1

Scenario	Sensitivity	Specificity	Prevalence	Response Rate
1A	0.9	0.9	0.002	0.1016
1B	0.9	0.9	0.005	
1C	0.9	0.9	0.200	
2A	0.9	0.8	0.002	0.2014
2B	0.9	0.8	0.005	
3A	0.8	0.9	0.002	0.1014

Compare: 1A versus 2A, then 1A versus 3A

THE APPLICATIONAL STAGE

- In the Applicational Stage, the basic question is: Does it work for "me"? It's the user's concern.
- Problem: One can't resolve the concern, like comparing the test result versus the truth, i.e. no data (one person; truth is known).

KEY PARAMETERS

- **Two parameters:**
  Positive Predictive Value, $P^+ = Pr(D=+|T=+)$
  Negative Predictive Value, $P^- = Pr(D=-|T=-)$
- **Positive predictive value is the probability having an accurate positive result and negative predictive value is the probability having an accurate negative result; (Perhaps, users are more often concerned about P^+ than P^-).**

Issue #1:
ESTIMATION OF PARAMETERS

- **Unlike sensitivity & specificity, predictive values P^+ and P^- cannot be estimated directly because there are no data.**
- **However, they can be "estimated" indirectly using the Bayes' theorem or Bayes' rule.**
- **Actually, we can't estimate them; we can only approximate them.**

BAYES' RULE

$$\Pr(B \mid A) = \Pr(B \text{ and } A)/\Pr(A)$$

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \text{ and } B) + \Pr(A \text{ and Not } B)}$$

$$\Pr(B \mid A) = \frac{\Pr(A \mid B)\Pr(B)}{\Pr(A \mid B)\Pr(B) + \Pr(A \mid \text{Not } B)\Pr(\text{Not } B)}$$

APPLICATION

$$\Pr(\text{B} \mid \text{A}) = \frac{\Pr(\text{A} \mid \text{B})\Pr(\text{B})}{\Pr(\text{A} \mid \text{B})\Pr(\text{B}) + \Pr(\text{A} \mid \text{Not B})\Pr(\text{Not B})}$$

Let A = (T=+) and B = (D=+), we have:

$$\Pr(D = + \mid T = +) = \frac{\Pr(T = + \mid D = +)\Pr(D = +)}{\Pr(T = + \mid D = +)\Pr(D = +) + \Pr(T = + \mid D = -)\Pr(D = -)}$$

Note: "not B" = (D=-)

RESULTS

Both predictive values are functions of disease prevalence, $\pi = \Pr(D = +)$:

$$P^+ = \frac{S^+\pi}{S^+\pi + (1 - S^-)(1 - \pi)}$$

$$P^- = \frac{S^-(1 - \pi)}{S^-(1 - \pi) + (1 - S^+)\pi}$$

EXAMPLE #2:
1A vs. 1B, 1A vs. 2A, & 1A vs. 3A

Since all cancers are "rare" (low prevalence and incidence), specificity has its dominating effect on the positive predictive value but not much on the negative predictive value.

Scenario	Sensitivity	Specificity	Prevalence	P+	P-
1A	0.9	0.9	0.002	0.0178	0.9998
1B	0.9	0.9	0.005	0.0433	0.9994
2A	0.9	0.8	0.002	0.0089	0.9997
2B	0.9	0.8	0.005		
3A	0.8	0.9	0.002	0.0158	0.9996
3B	0.8	0.9	0.005		

Issue #2: Again, Should We Conduct "RANDOM TESTING" For Diseases, Such As AIDS?

EXAMPLES: AIDS SCREENING

Example A: $S^+=.977$, $S^-=.926$, and $\pi=.003$:

$$P^+ = \frac{(.977)(.003)}{(.977)(.003)+(.074)(.997)} = .038 \text{ or } 3.8\%$$

Example B: $S^+=.977$, $S^-=.926$, and $\pi=.20$:

$$P^+ = \frac{(.977)(.20)}{(.977)(.20)+(.074)(.80)} = .767 \text{ or } 76.7\%$$

Note: Current Estimate for USA's AIDS: .3% as above and S^+ and S^- are for ELISA in Weiss, 1985.

IMPLICATION

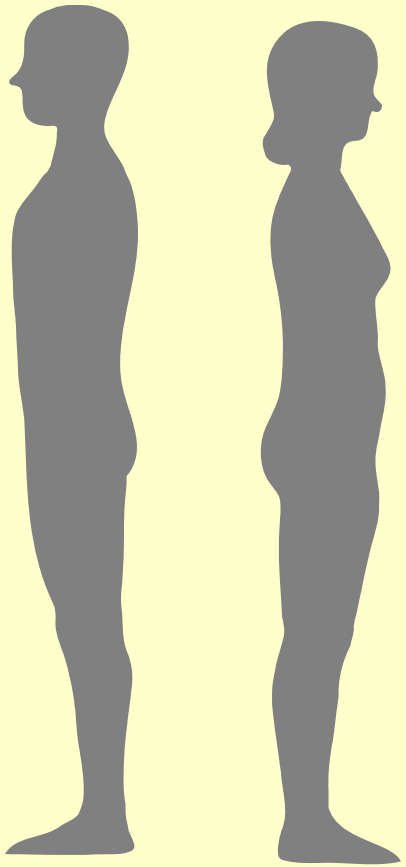
- Predictive values of a screening test depend not only on sensitivity and specificity but on disease prevalence too.
- The higher the prevalence, the higher the positive predictive value; "random screening" or "random testing" might not do much good – many false positives.
- The higher the prevalence, the lower the negative predictive value (but the effect is much weaker for P^-)

MORE ABOUTBREAST CANCER

- Breast Cancer is an uncontrolled proliferation of cells (when normal process goes wrong, new cells form unnecessarily and old cells do not die when they should); extra cells form tumors, some are malignant.
- It's a very diverse disease of many varying histological subtypes; different subtypes make it more difficult to treat and to screen.
- The lifetime risk for American women is 1 in 8 – up from 1 in 20 in 1960; In 2009, there were over 200,000 new cases – majority are invasive.

2009 Estimated US Cancer Cases

Men 766,130	
Prostate	25%
Lung & bronchus	15%
Colon & rectum	10%
Urinary bladder	7%
Melanoma of skin	5%
Non-Hodgkin lymphoma	5%
Kidney & renal pelvis	5%
Leukemia	3%
Oral cavity	3%
Pancreas	3%
All Other Sites	19%



Women 713,220	
27%	Breast
14%	Lung & bronchus
10%	Colon & rectum
6%	Uterine corpus
4%	Non-Hodgkin lymphoma
4%	Melanoma of skin
4%	Thyroid
3%	Kidney & renal pelvis
3%	Ovary
3%	Pancreas
22%	All Other Sites

#1 in Incidence

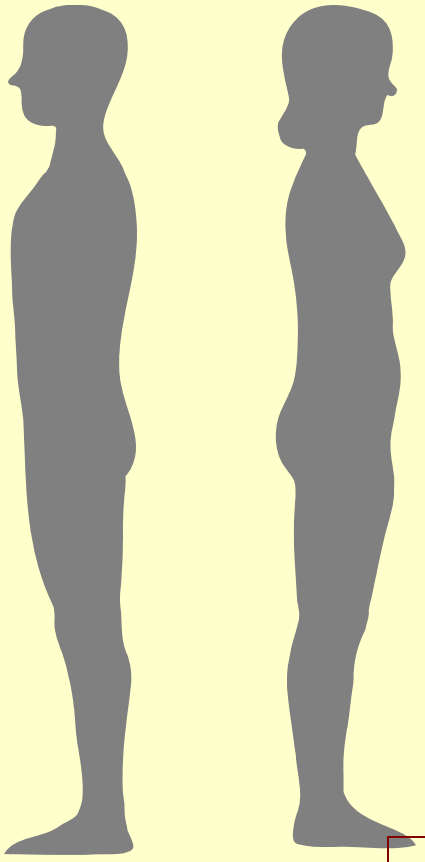
Lifetime Probability of Developing Cancer, U.S. Women, 2003-2005

Site	Risk
All sites†	1 in 3
Breast	**1 in 8**
Lung & bronchus	1 in 16
Colon & rectum	1 in 20
Uterine corpus	1 in 40
Non-Hodgkin lymphoma	1 in 53
Urinary bladder‡	1 in 84
Melanoma§	1 in 58
Ovary	1 in 72
Pancreas	1 in 75
Uterine cervix	1 in 145

#1 in Lifetime Risk

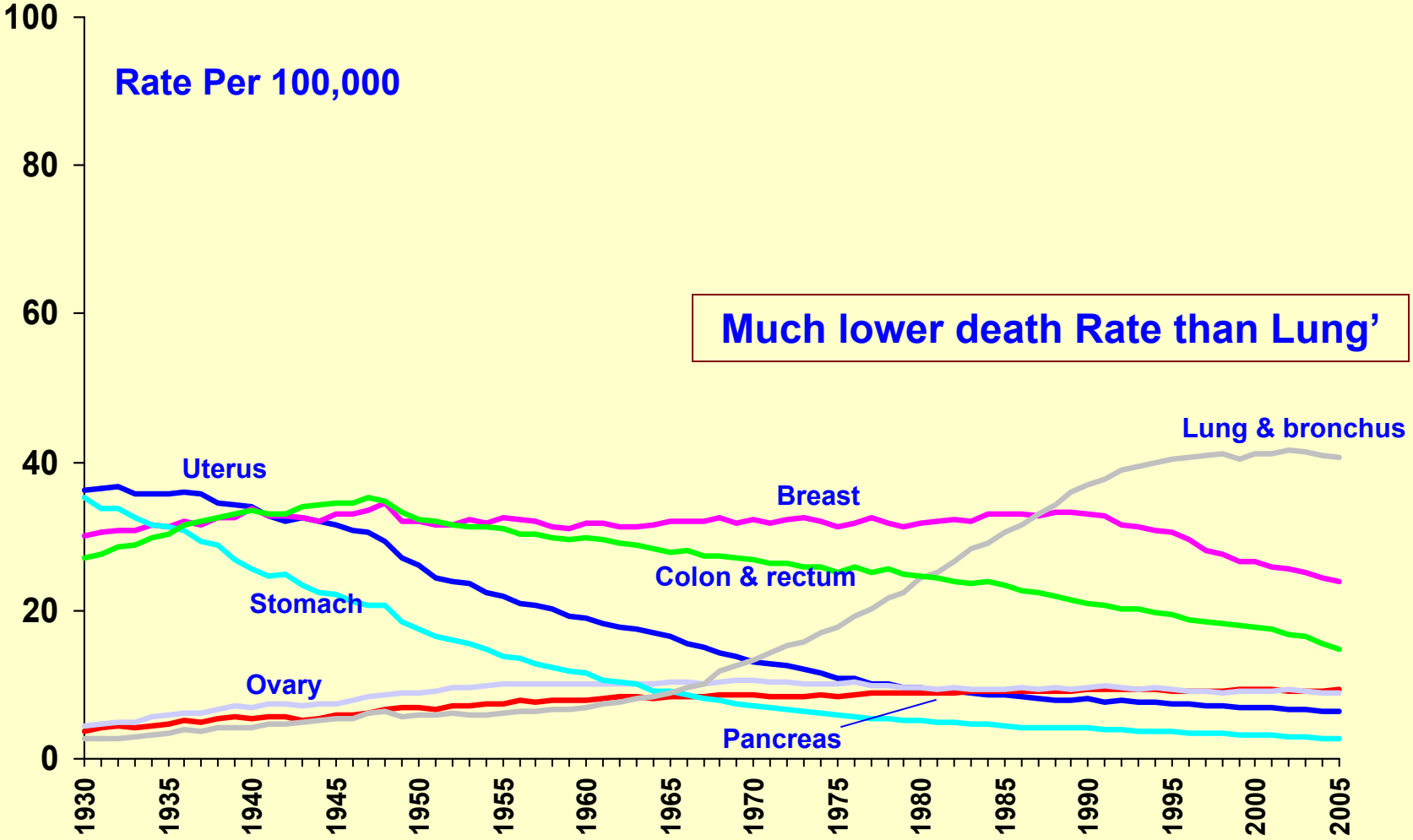
2009 Estimated US Cancer Deaths

Men		Women	
292,540		269,800	
Lung & bronchus	30%	26%	Lung & bronchus
Prostate	9%	**15%**	**Breast**
Colon & rectum	9%	9%	Colon & rectum
Pancreas	6%	6%	Pancreas
Leukemia	4%	5%	Ovary
Liver & intrahepatic bile duct	4%	4%	Non-Hodgkin lymphoma
Esophagus	4%	3%	Leukemia
Urinary bladder	3%	3%	Uterine corpus
Non-Hodgkin lymphoma	3%	2%	Liver & intrahepatic bile duct
Kidney & renal pelvis	3%	2%	Brain/ONS
All other sites	25%	25%	All other sites



#2 "Cancer Killer"

Cancer Death Rates* Among Women, US,1930-2005



*Age-adjusted to the 2000 US standard population.
Source: US Mortality Data 1930-2005, National Center for Health Statistics

Trends in Five-year Relative Survival (%)* Rates, US, 1975-2004

Site	1975-1977	1984-1986	1996-2004
• All sites	50	54	66
• **Breast (female)**	**75**	**79**	**89***
• Colon	52	59	65
• Leukemia	35	42	51
• Lung and bronchus	13	13	16
• Melanoma	82	87	92
• Non-Hodgkin lymphoma	48	53	65
• Ovary	37	40	46
• Pancreas	3	3	5
• Prostate	69	76	99
• Rectum	49	57	67
• Urinary bladder	74	78	81

Not as bad as you think

***It's 91% in 2009**

SCREENING FOR BC

- Genetic predisposition, genes BRCA1 and BRCA2, accounts for only 5% to10% of all breast cancer cases.
- No obvious risk factors other than family history and age (& gender); by the age of 50 years, more than 50% of the BRCA1 or BRCA2 mutation carriers have already developed the disease.
- Existing screening methods are: Self Breast Exam, Ultrasound, Magnetic Resonance Imaging (MRI), and Mammography.

BREAST SELF-EXAMINATION

A large randomized trial (n = 266,064) in Shanghai (1989-1991) lead to the following conclusion: "Women who choose to practice breast self-examination should be informed that its efficacy is unproven and that it may increase their chances of having a benign breast biopsy"; after 10 years of follow-up, breast cancer mortality rates in 2 groups were identical (JNCI 94: 1445-1457, 2002).

In short, it might do more harm than good.

ULTRASONOGRAPHY (US)

- The term "ultrasound" applies to all acoustic energy with a frequency above human hearing.
- Medical Ultrasonography is an ultrasound-based diagnostic imaging technique used to visualize muscles and internal organs, their size, structures and possible pathologies.
- More popular in OBGYN for prenatal care but not so popular for Breast Cancer Screening in general; often only used for BC during pregnancy to avoid radiation (of mammography).
- It is about as sensitive but a little less specific than mammography; specificity ranges 80-93%; it picks up a few more benign tumors.

MAGNETIC RESONANCE IMAGING

- Magnetic resonance imaging (MRI) is a non-invasive method used to render images of the inside of an object.
- It uses radio waves and a strong magnetic field to provide remarkably clear and detailed pictures of internal organs and tissues.
- It requires specialized equipment to evaluate body structures that may not be as visible with other imaging methods; e.g. you can see not only the organs but even blood vessels too.

ADVANTAGES OF MR IMAGING

- Use of MRI first reported in 1985.
- MRI not associated with ionizing radiation; no known long-term side effects.
- MRI is not impaired by dense parenchyma; sensitivity improves;
- MRI could measure not only physiological but functional properties of tissues as well.
- However, for now, breast MR imaging is not used routinely in a screening setting. Why? High cost is the major inhibitive reason; the machine costs 2-4 million dollars and each episode $1500-$2,000

MAMMOGRAPHY

- Mammography is the process of using low-dose X-rays to examine the human breast; It uses doses of ionizing radiation to create image
- It is used to detect and diagnose breast disease or tumor, both in women with or without breast complaints or symptoms - i.e. more routine.
- Modern mammography has only existed since 1969, when the first x-ray machines used just for breast imaging became available. Technology has advanced, so that today's mammogram is very different even from those of the mid-1980s.

THE ISSUE

- The need is not the issue; it decreases BC mortality by 32% (Tabar, 2000; from "the Swedish two-county trial").
- The test "characteristics" may not be the major issue; sensitivity is low (Kuhl, 2000) but the specificity ranges from 93%-99.7% in high-risk women (Warner, 2001).
- But is forty or fifty "old enough"? (to be at "higher risk" for efficient screening)

SCREENING GUIDLINE?

- There are guidelines, by federal panels and/or ACS, but are there any justification? Why 40? Why 50? Or, why not starting at 35?
- Here are some post-hoc overall data by ACS: about 10% or less* are "recalled" for more tests (because the first mammogram is "positive"); 8%-10% of those need biopsy – because mammogram is positive again, and 20% of those with biopsy have cancer. That puts the positive predictive value (of first test) at most 1.6%-2%.

*It's 3%-4% for women age 40

HOW GOOD IS GOOD?

Some investigators imply that a "good test" must yield $P^+ \geq 50\%$; by either improving its characteristics (S^+ and S^-) or by selecting the population in which the test is used so that the background prevalence is higher.

But if you cannot improve (S^+ and S^-); When Does It Make Sense to Screen?

Issue #3:

When Does It Make Sense to Screen?

From:

$$P^+ = \frac{S^+\pi}{S^+\pi + (1-S^-)(1-\pi)}$$

We can "solve" for π:

$$\pi = \frac{(1-S^-)P^+}{S^+ + (1-S^-)P^+ - S^+P^+}$$

Then set a "desirable level" for P+ to obtain "screenable prevalence"

SCREENABLE PREVALENCE

For example, setting P^{+}=.80 or 80%

Table 2.1		Sensitivity, S+				
		0.5	0.9	0.95	0.98	0.99
Specificity, S-	0.5					
	0.9		0.308		0.29	
	0.95			0.174		
	0.98		0.082		0.075	
	0.99					

That is, if $S^{+}=S^{-}$=.98, we "attain" a positive predictive value of 80% if prevalence $\pi \geq .075$; much higher if the test is not that good.

Specificity has more influence on Screenable Prevalence:

$\pi \geq .082$ when (s^{-}=.98,s^{+}=.90) but $\pi \geq .29$ when (s^{-}=.90,s^{+}=.98)

THE PAP TEST: NOT THAT BAD!

- The "Pap" test or Pap Smear test, or cytology test, is an important part of women's health care. The smeared cells or cell suspension is placed on a glass slide, stained with a special dye (Pap stain), and viewed under a microscope. It is used to detect cervical cancer as well as some vaginal or uterine infections.
- As for cervical cancer, it is still not very sensitive, especially cases in early stage. However, because it is highly specific (could be about 99%), its positive predictive value is high making it suitable for "case identification".

SOME RESULTS OF MEMMOGRAPHY

(Currently) S^-=.966, S^+=.647 & $\pi = \dfrac{(1-S^-)P^+}{(1-S^-)P^+ + S^+(1-P^+)}$

Predictive Value, P	Screenable Prevalence
1%	53 per 100,000
2%	**107 per 100,000**
5%	**276 per 100,000**
10%	581 per 100,000

Prevalences from SEER:

Age Group	Rate
35-39	59 per 100,000
40-44	**119 per 100,000**
45-49	194 per 100,000
50-54	**254 per 100,000**
55-59	313 per 100,000

COMPETING STRATEGIES
FOR BREAST CANCER SCREENING

Starting at age 40: Incidence Rate is about 119 per 100,000

Positive Predictive Value is 2%

Negative Predictive Value is 99.96%

Starting at age 50: Incidence Rate is about 254 per 100,000

Positive Predictive Value is 5%

Negative Predictive Value is 99.91%

Would it be justified to reduce from 50 to 40?

CAN IT BE IMPROVED?

- Unfortunately, very often, neither maneuvers - by either improving its characteristics or by selecting the population with higher prevalence may be possible to yield $P^{+} \geq 50\%$; That's may be reasonable but too much to ask, even tests useful clinically may not pass!
- For AIDS, maybe one should only screen "high-risk" sub-populations, like drug IV abusers or prisoners; but what's breast cancer, what should we do? We know that early detection is proven to save lives.

WHAT ABOUT A RE-TEST?

- If starting at age 40, and if "recalled", the chance to have cancer would be about 2%. Another recall for biopsy would raise the predictive value to 28% (which is similar to ACS' data of about 20% - perhaps including younger users).
- If starting at age 50, and if "recalled", the chance to have cancer would be about 5%. Another recall for biopsy would raise the predictive value to 50%-51%; that qualifies it as a "good" procedure as stipulated by some investigators.

SCREENING EFFICACIES

Starting at age 40: Incidence Rate is about 119 per 100,000

Positive Predictive Value is 2%

Two recalls raise predictive value to 28%

Starting at age 50: Incidence Rate is about 254 per 100,000

Positive Predictive Value is 5%

Two recalls raise predictive value to 51%

Even at age 35-39, if your memmogram is positive, there is still a 1% chance that you have breast cancer. Is 1% a worthy chance?

For some, when it comes to saving life, no chance is a slim chance; there are other costs but no cost is as pricey as life.

But, remember that false positives are not without consequences.

Should Women Start Mammograms at Age 40 or 50?

- For those with "reason" to test, i.e. women with family history (mother or sisters with BC), decision is easier – and should be recommended (by age 50 it might be too late, more than 50% of the BRCA1 or BRCA2 mutation carriers have already developed the disease).
- For others, it may boil down to this not-very-simple question: are you prepared for unwanted consequences? At age 40, 98% of positive mammograms are false positives and, after another recall, 72% of biopsies are negative

- **Besides false positives, it's the possibility of … over-diagnosis.**

- **Many benign tumors need not be removed.**
- **Some breast cancers may lie dormant for years, even never develop.**
- **Some breast cancers grow slowly, do not spread, even shrink/go away on their own.**

Issue #4: To form an INDEX measuring "**Diagnostic Competence**"

- Other things (cost, ease of application, etc…) being equal, a test with larger values of both sensitivity and specificity is obviously better.
- If not that clear cut, one has to consider the relative costs associated with 2 forms of error.
- If the 2 types of error are equally important, it may be desirable to have a single index to measure the "diagnostic competence" of the test.
- Candidates might include "Overall Agreement" and "Kappa Statistic".

OVERALL AGREEMENT

- The simplest measure would be the "overall agreement", Pr (T=D).
- However, unlike sensitivity and specificity, the **overall agreement is influenced by the disease prevalence**.

$$\Pr(T = D) = \Pr(T = D = +) + \Pr(T = D = -)$$

$$\Pr(T = D) = \Pr(T = + \mid D = +)\Pr(D = +) + \Pr(T = - \mid D = -)\Pr(D = -)$$

$$\Pr(T = D) = \pi S^{+} + (1 - \pi) S^{-}$$

KAPPA STATISTIC

- Another alternative is “Kappa Statistic”
- Kappa is a popular statistic often used to measure agreement between observers.
- It adjusts overall agreement for “chance agreement”.
- But, similar to the case of overall agreement, **Kappa is still influenced by the disease prevalence.**

$$\kappa = \frac{\{\text{Overall agreement}\} - \{\text{Chance agreement}\}}{1 - \{\text{Chance agreement}\}}$$

$$\kappa = \frac{[\Pr(T = D = +) + \Pr(T = D = -)] - [\Pr(T = +)\Pr(D = +) + \Pr(T = -)\Pr(D = -)]}{1 - [\Pr(T = +)\Pr(D = +) + \Pr(T = -)\Pr(D = -)]}$$

An interesting exercise: Express Kappa as a function of sensitivity, specificity, and prevalence; then fix sensitivity and specificity to see if Kappa is a monotonic function of disease prevalence (e.g. Making a graph using S-plus)

YOUDEN'S INDEX

- That leaves one measure, Youden's Index (Cancer, 1950)
- If the 2 types of error are equally important, the Youden's Index J is defined as:

$$\mathbf{J = 1 - (\alpha + \beta) = S^{+} + S^{-} - 1}$$

The Youden's index J special with interesting characteristics: (i) it is based on a simple principle: **small sum of errors** (when neither one has priority), (ii) its value is larger when both sensitivity and specificity are high, and (iii) **It does not depend on the disease prevalence**. And **there are other reasons** as well!

Issue #5:

When does a process qualify as a test?

- To decide if a "process" is a "test", the minimum criterion it must pass is that it detects disease better than by chance alone
- That a process can only qualify as a test if **it selects diseased persons with higher probability than pure guessing: $P^{+}>\pi$.**

BASIC QUALIFICATION

$$P^+ = \frac{S^+\pi}{S^+\pi + (1-S^-)(1-\pi)}$$

$$P^+ > \pi \Leftrightarrow S^+ > 1-S^- \Leftrightarrow J > 0$$

$$\pi = \Pr(D = +)$$

$$P^+ = \Pr(D = + \mid T = +)$$

Issue #6:

THE ISSUE OF COST

For a Woman Age 40 – 44

$$100 + \frac{119}{100{,}000}[(.647)(100) + (.647)^2(5500)]$$

$$+ [1 - \frac{119}{100{,}000}][(.034)(100) + (.034)^2(5500)]$$

$$= \$112.81/\text{year}$$

Calculations are based on these cost estimates

Mammogram $80 – $120;

Needle Biopsy $5000 – $6000

SCREENING COSTS AT AGE 50

For a Woman Age 50 – 54

$$100 + \frac{254}{100{,}000}[(.647)(100) + (.647)^2(5500)]$$

$$+ [1 - \frac{254}{100{,}000}][(.034)(100) + (.034)^2(5500)]$$

$$= \$115.98/\text{year}$$

Calculations are based on these cost estimates

Mammogram $80 – $120

Needle Biopsy $5000 – $6000

ESTIMATED COSTS TO SAVE A LIFE

100,000 women age 40-44

119 with breast cancer (source: SEER)

77 identified by mammograms (sensitivity = .647)

50 identified as positive again (sensitivity = .647)

50 confirmed by biopsies (assume 100% rate): →treatment

12 died if all were not screened (assume 25% death rate)

4 would be saved by mammograms (32% rate by Tabar)

100,000 go through the process, 4 lives saved

25,000 go through the process to save 1 life (called NNS)

Age 40: NNS = 25,000 [Cost = (25,000)(112.81)=$2.82M/yr)

Age 50: NNS = 11,700 [Cost = (11,700)(115.98)=$1.36M/yr)

EXERCISES

12.1. Prove that P^+ is an increasing function of π.

12.2. Is P^- an increasing or decreasing function of π?

12.3. Express Kappa statistic (κ) as a function of S^+, S^-, and π; is it an increasing or decreasing function of π?

12.4. Prove that $P^- > (1- \pi)$ if and only if $J>0$.

12.5. Prove that, in the 2-by-2 cross-classification of D(+,-) versus T(+,-), Odds Ratio is equal to 1 if & only if $J=0$.

READINGS/REFERENCES

- Armitage P (1977). Statistical Methods in Medical Research, published by Wiley & Sons; pp. 433-438.
- Bishop YMM, Fienberg SE, and Holland PW (1988). Discrete Multivariate Analysis, published by MIT Press; pp. 393-396.
- Buck AA and Gart JJ (1966). AJE 83: 586-592.
- Carey et al. (1976). JAMA 236: 847-851
- Gart JJ and Buck AA (1966). AJE 83: 593-602
- Le CT (2003). Introductory Biostatistics, published by Wiley & Sons; pp. 5-7, 115-117.
- Moskowitz et al. (1974). NEJM 295:249-252
- Rogan WJ and Gladen B (1978). AJE: 71-75
- Stamler et al. (1976). JAMA 235: 2299-2306