NORTH DENMARK REGION

Regression models for accuracy estimation

Niels Henrik Bruun,
Research Data and Biostatistics,
Aalborg University Hospital

NORTH DENMARK REGION

Acknowledgement

In collaboration with:

Professor Helle Damgaard Zacho,
Department of Nuclear Medicine and Clinical Cancer Research Center,
Aalborg University Hospital

Dr. Farid Gossili,
Senior Registrar and Research Fellow in Nuclear Medicine,
Aalborg University Hospital

NORTH DENMARK REGION

Case origin:

- 50 men diagnosed with prostate cancer.
- After diagnosis, all men underwent prostatectomy.
- Histology (pathology) served as the reference standard.
- All underwent two different types of PET scans and a MR scan.
- Presence or absence of cancer was marked for six anatomical segments in the prostate.

- **Report and compare the diagnostic accuracy for the 3 modalities**

NORTH DENMARK REGION

Accuracy Measures

- Measures (Bland (2015))
  - Sensitivity (TPR), Specificity (1 - FPR).
  - AUC, LR+, LR- (Pos and neg likelihood ratios).
  - PPV, NPV (Pos and neg predicted values), Accuracy.

		Index-test		
		Positive	Negative	Rates
Reference	Positive (P)	True Positive(TP)	False Negative (FN)	TPR = Sensitivity = TP / P
	Negative (N)	False Positive (FP)	True Negative (TN)	FPR = 1 - Specificity = FP / N

The confusion matrix for a binary test

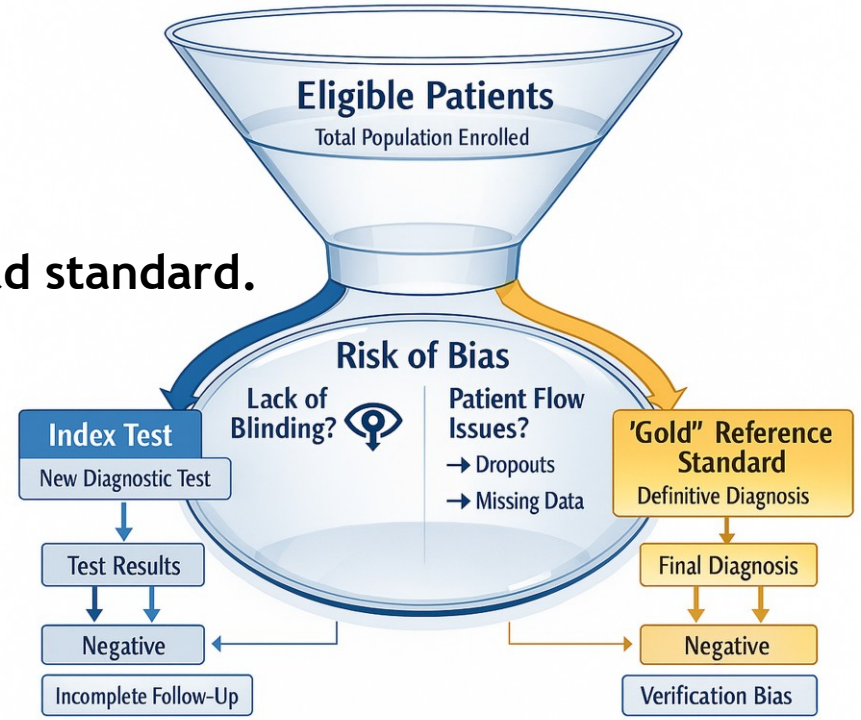
Estimating accuracy measures from TPR and FPR

- Prevalence (prv)
  - The proportion of actual positives in the sample.
  - Sensitivity and specificity can be used across different prevalences.
- Accuracy measures based solely on TPR and FPR.
  - $Sensitivity = TPR$
  - $Specificity = 1 - FPR$
  - $LRp = \frac{TPR}{FPR}$
  - $LRm = \frac{1-TPR}{1-FPR}$
  - $AUC_{bin} = \frac{1+TPR-FPR}{2}$
- Accuracy measure formulas also requiring prv
  - $Accuracy = prv \cdot TPR + (1 - prv) \cdot (1 - FPR)$
  - $PPV = \frac{prv \cdot TPR}{prv \cdot TPR + (1-prv) \cdot FPR}$
  - $NPV = \frac{(1-prv) \cdot (1-FPR)}{(1-prv) \cdot (1-FPR) + prv \cdot (1-TPR)}$

The setup

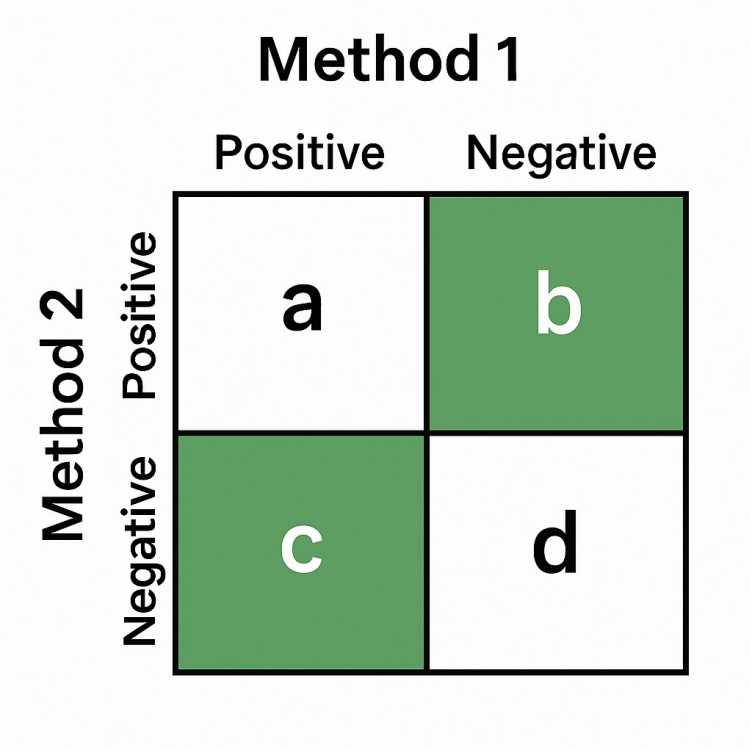
- We use a binary reference and a binary test
- TPR estimator is conditioned on positives
  - does not depend on prevalence.
- FPR estimator is conditioned on negatives
  - does not depend on prevalence.
- The TPR and FPR estimators are uncorrelated
  - The estimators are based on disjoint subsets

- **TPR and FPR describe how well the test matches the gold standard.**
  - The index test is never as good as the gold standard



Standard Methods

- Statistical Methods
  - McNemar's test.
  - Risk differences (RD).
  - Confidence intervals (CI).
- Formulas
  - $RD = \frac{b-c}{n}$
  - Simplified Wald CI: $RD_{ci} = RD \pm z_{\alpha/2} \cdot \sqrt{b+c}/n$



Sample design and representativeness

- Estimating TPR and FPR in one population
  - Sample prevalence equals the population proportion
  - Power calculation, Buderer (1996)
  - Require large samples

- Or ignore representativeness Rothman et al. (2013).
  - Estimating TPR in a subpopulation of mainly positives
  - Estimating FPR in a subpopulation of mainly negatives
  - Sample prevalence ≠ population prevalence
  - This design requires instruments to behave the same way across subpopulations.
  - TPR seems overestimated, Whiting et al. (2013)
  - FPR seems underestimated, Whiting et al. (2013)
  - Case–control designs tend to inflate both TPR and FPR.
  - Extreme cases and healthy controls exaggerate diagnostic contrast.

On Diagnostic Accuracy

- Pepe (2003)
- Whiting et al. (2004) → Identified bias sources.
- Whiting (2011) QUADAS-2
  - Quality Assessment Tools
  - Evaluates *risk of bias* and *applicability* across 4 domains.
    - Patient selection (bias?)
    - Index test (blinding?)
    - Reference standard (blinding?)
    - Flow and timing (time between tests, dropout)
- Whiting et al. (2013) → Updated evidence using QUADAS-2.
- STARD 2015 (Cohen et al. (2016)):
  - Reporting Guidelines
  - 30-item checklist for transparency & completeness

NORTH DENMARK REGION

Simulated data (back to the case)

- 800 men for one segment
- 3 measurement types f, ga, mr (wide format)
- Pathology (pa) as gold standard

```
. use simdata, clear
. list in 1/5, noobs
  +----------------------+
  | pa     f    ga    mr |
  |----------------------|
  | no    no    no    no |
  | no    no   yes    no |
  | no   yes    no    no |
  | no   yes   yes    no |
  | no   yes   yes   yes |
  +----------------------+
```

NORTH DENMARK REGION

Regressing test on reference to get TPR aQnd FPR

Estimation:

Using binreg rd

```
. binreg mr bn.pa, nocons vce(robust) rd
```

Using OLS regression

```
. regress mr bn.pa, nocons vce(robust) noheader
```

Using binreg rr

```
. binreg mr bn.pa, nocons vce(robust) rr
```

Using poisson regression

```
. poisson mr bn.pa, nocons vce(robust) irr
```

- Estimation from regression:
  - TPR = _b[1.disease])
  - FPR = _b[0.disease])
- The estimators of TPR and FPR are independent
  - The estimators are based on disjoint subsets
- Confidence intervals depend on the chosen scale (linear vs log).
- OLS with robust variance estimation is chosen for estimation
  - A probability is a mean of zeroes and ones.
  - Confidence intervals can end outside the interval [0, 1].
  - Robust variance is required.
- Inspired by Cummings (2009)

		p	[95%	CI]
TPR/sens	binreg(rd)	0.501	0.454	0.549
	OLS	0.501	0.453	0.549
	binreg(rr)	0.501	0.456	0.551
	poisson	0.501	0.456	0.551
FPR/(1-spec)	binreg(rd)	0.084	0.053	0.114
	OLS	0.084	0.053	0.114
	binreg(rr)	0.084	0.058	0.121
	poisson	0.084	0.058	0.121

NORTH DENMARK REGION

Estimating accuracy measures from TPR and FPR

Using OLS with robust variance estimation of TPR and FPR

```
. regress mr bn.pa, ///
        nocons vce(robust)
. nlcom (tpr : _b[1.pa]) ///
        (fpr : _b[0.pa]), post
```

Using nlcom and the sample prevalence for estimation

```
. su pa, mean
. local prv = r(mean)
. nlcom (sensitivity : _b[tpr]) ///
        (specificity : 1 - _b[fpr]) ///
        (AUCbin : 0.5 * (_b[tpr] + 1 - _b[fpr])) ///
        (LRp : _b[tpr] / _b[fpr]) ///
        (LRm : (1 - _b[tpr]) / (1 - _b[fpr])) ///
        (accuracy: _b[tpr] * `prv' + (1 - _b[fpr]) * (1 - `prv')) ///
         (ppv: _b[tpr] * `prv' / ///
                (_b[tpr] * `prv' + _b[fpr] * (1 - `prv'))) ///
         (npv: (1 - _b[fpr]) * (1 - `prv') / ((1 - _b[fpr]) * (1 - `prv') ///
                + (1 - _b[tpr]) * `prv')) ///
        , post
```

	p	[95%	CI]
sensitivity	0.50	0.45	0.55
specificity	0.92	0.89	0.95
AUCbin	0.71	0.68	0.74
LRp	5.99	3.71	8.28
LRm	0.54	0.49	0.60
accuracy	0.66	0.63	0.69
ppv	0.90	0.87	0.94
npv	0.54	0.51	0.56

NORTH DENMARK REGION

Adding instruments for comparisons

- Reshape into long format by ID and instrument.
- Correlation from random intercept by id

```
. use simdata, clear
. generate id = _n
. rename (f ga mr) tst=
. reshape long tst, i(id) j(msrmnt) string
. strtonum msrmnt
. list in 1/6, noobs sepby(id)
  +-------------------------+
  | id   msrmnt   tst   pa  |
  |-------------------------|
  |  1        f    no   no  |
  |  1       ga    no   no  |
  |  1       mr    no   no  |
  |-------------------------|
  |  2        f    no   no  |
  |  2       ga   yes   no  |
  |  2       mr    no   no  |
  +-------------------------+
```

NORTH DENMARK REGION

Regression estimates of TPR and FPR by instrument

Three uncorrelated samples

```
. estimates clear
. qui regress tst bn.pa#i.msrmnt, nocons vce(robust)
. nlcom ///
        (se_f : _b[1.pa#1.msrmnt]) ///
        (sp_f : 1 - _b[0.pa#1.msrmnt]) ///
        (se_ga : _b[1.pa#2.msrmnt]) ///
        (sp_ga : 1 - _b[0.pa#2.msrmnt]) ///
        (se_mr : _b[1.pa#3.msrmnt]) ///
        (sp_mr : 1 - _b[0.pa#3.msrmnt]) ///
        , post
. estimates store m_indep
```

	indep: p	[95%	CI]	dep: p	[95%	CI]
se_f	0.75	0.71	0.79	0.75	0.71	0.79
se_ga	0.79	0.76	0.83	0.79	0.76	0.83
se_mr	0.50	0.45	0.55	0.53	0.48	0.57
sp_f	0.73	0.68	0.78	0.73	0.68	0.78
sp_ga	0.65	0.60	0.70	0.65	0.60	0.70
sp_mr	0.92	0.89	0.95	0.92	0.89	0.95

Three measurements on the same ID (dependence by random intercept)

```
. qui mixed tst bn.pa#i.msrmnt, nocons ||id:, vce(robust)
. nlcom ///
        (se_f : _b[1.pa#1.msrmnt]) ///
        (sp_f : 1 - _b[0.pa#1.msrmnt]) ///
        (se_ga : _b[1.pa#2.msrmnt]) ///
        (sp_ga : 1 - _b[0.pa#2.msrmnt]) ///
        (se_mr : _b[1.pa#3.msrmnt]) ///
        (sp_mr : 1 - _b[0.pa#3.msrmnt]) ///
        , post
. estimates store m_dep
```

NORTH DENMARK REGION

Correlations in the case of dependence

- The correlation between instruments for sensitivities and specificities:
  - Is based on shared observations of positive agreement cases
  - Correlations reflect shared variation, not agreement in the sense of reliability.

```
. estat vce, correlation
Correlation matrix of coefficients of nlcom model
```

e(V)	se_f	sp_f	se_ga	sp_ga	se_mr	sp_mr
se_f	1.0000					
sp_f	0.0000	1.0000				
se_ga	**0.8801**	-0.0000	1.0000			
sp_ga	0.0000	0.4252	0.0000	1.0000		
se_mr	0.3642	0.0000	**0.5135**	0.0000	1.0000	
sp_mr	0.0000	0.5006	0.0000	0.4112	0.0000	1.0000

Comparing instruments ga and mr

Similar point estimates, but different confidence intervals

```
. estimates restore m_indep
. nlcom (se_ga_vs_mr: _b[se_ga]-_b[se_mr]) ///
        (sp_ga_vs_mr: _b[sp_ga]-_b[sp_mr])
. estimates restore m_dep
. nlcom (se_ga_vs_mr: _b[se_ga]-_b[se_mr]) ///
        (sp_ga_vs_mr: _b[sp_ga]-_b[sp_mr])
```

		diff	[95%	CI]
se_ga_vs_mr	indep	0.29	0.23	0.35
	dep	0.27	0.23	0.31
sp_ga_vs_mr	indep	-0.27	-0.33	-0.21
	dep	-0.27	-0.32	-0.22

Comparing instruments ga and f

Similar estimates, smaller confidence intervals

```
. estimates restore m_indep
. nlcom (se_ga_vs_f: _b[se_ga]-_b[se_f]) ///
        (sp_ga_vs_f: _b[sp_ga]-_b[sp_f])
. estimates restore m_dep
. nlcom (se_ga_vs_f: _b[se_ga]-_b[se_f]) ///
        (sp_ga_vs_f: _b[sp_ga]-_b[sp_f])
```

		diff	[95%	CI]
se_ga_vs_f	indep	0.04	-0.01	0.10
	dep	0.04	0.03	0.06
sp_ga_vs_f	indep	-0.08	-0.16	-0.01
	dep	-0.08	-0.14	-0.03

confreg

Title

confreg – Confusion matrix (Accuracy measures) estimated by regression and nlcom

Syntax

confreg varlist(min=2 max=3) [if] [, options]

options - Description

id(passthru) - If modalities are measured on the same id, specify the id variable. The used model becomes a mixed random intercepts by id.

randomeffect(string) - String for adding random effects to the model. Possibly succeeded by the random effect of the ids (||id:).

adjustment(varlist fv) - Add adjustment variables to the model.

coleq(string) - Add coleq text to the stored matrices.

prevalence(numlist max=1 >0 <1) - Specify prevalence to use. Default is the sample prevalence.

vce(passthru) - Set vce options.

stub(string) - Stub to add the names of the returned estimates.

scale(#) - Default value is 100.

NORTH DENMARK REGION

Confreg output with sample prevalence 1/3

Remember option id for correlated data

```
. confreg pa tst msrmnt, id(id) vce(robust)
                                  |          N           p       [95%         CI]
----------------------------------+-----------------------------------------------
f                                 |
           Sensitivity, P(TP|C+)  |        489    74.84663    70.9985    78.69475
           Specificity, P(TN|C-)  |        311     73.3119    68.3928      78.231
           AUCbin, (sens+spec)/2  |        800    74.07926    70.95654   77.20199
----------------------------------+-----------------------------------------------
ga                                |
           Sensitivity, P(TP|C+)  |        489     79.3456    75.75528   82.93592
           Specificity, P(TN|C-)  |        311    64.95177    59.64576   70.25778
           AUCbin, (sens+spec)/2  |        800    72.14869     68.9454   75.35197
----------------------------------+-----------------------------------------------
mr                                |
           Sensitivity, P(TP|C+)  |        489    52.59722    47.97447   57.21998
           Specificity, P(TN|C-)  |        311    91.63987    88.56173   94.71801
           AUCbin, (sens+spec)/2  |        800    72.11855    69.34164   74.89545
```

NORTH DENMARK REGION

Confreg output with sample prevalence 2/3

Returned by confreg

```
. return list
matrices:
            r(confreg) :  19 x 4
         r(acc_ppv_npv) :  10 x 4
      r(se_sp_auc_corr) :  9 x 9
           r(se_sp_auc) :  9 x 4


. estimates dir
-------------------------------------------------------
             |           Dependent  Number of
        Name | Command    variable     param.  Title
-------------+-----------------------------------------
  _se_sp_auc | nlcom     no depvar           9
_acc_ppv_npv | nlcom     no depvar           9
-------------------------------------------------------
```

NORTH DENMARK REGION

Confreg output with sample prevalence 3/3

Remember option id for correlated data

```
. confreg pa tst msrmnt, id(id) vce(robust)
. matprint r(acc_ppv_npv), decimals((0,2))
```

		N	p	[95%	CI]
	Prevalence, C+/N		0.61		
f	Accuracy, P(TP + TN)	800	74.25	71.22	77.28
	PPV, P(TP|P+)	449	81.51	78.63	84.40
	NPV, P(TN|P-)	351	64.96	61.15	68.76
ga	Accuracy, P(TP + TN)	800	73.75	70.74	76.76
	PPV, P(TP|P+)	497	78.07	75.36	80.77
	NPV, P(TN|P-)	303	66.67	62.40	70.93
mr	Accuracy, P(TP + TN)	734	67.78	64.71	70.84
	PPV, P(TP|P+)	238	90.82	87.66	93.98
	NPV, P(TN|P-)	496	55.15	52.60	57.70

NORTH DENMARK REGION

Confreg output with population prevalence = 0.3

Remember option id for correlated data

```
. confreg pa tst msrmnt, id(id) vce(robust) prevalence(0.3)
. matprint r(acc_ppv_npv), decimals((0,2))

------------------------------------------------------
                              N      p    [95%    CI]
------------------------------------------------------
    Prevalence, C+/N                0.30
f   Accuracy, P(TP + TN)    800   73.77   70.14   77.40
    PPV, P(TP|P+)           449   54.59   49.84   59.33
    NPV, P(TN|P-)           351   87.18   85.31   89.05
ga  Accuracy, P(TP + TN)    800   69.27   65.40   73.14
    PPV, P(TP|P+)           497   49.24   45.30   53.19
    NPV, P(TN|P-)           303   88.01   85.98   90.03
mr  Accuracy, P(TP + TN)    734   79.93   77.36   82.49
    PPV, P(TP|P+)           238   72.95   65.48   80.42
    NPV, P(TN|P-)           496   81.85   80.32   83.39
------------------------------------------------------
```

Summary

- A regression approach for estimating diagnostic accuracy is presented
  - regression models handles confounding and stratifications
- Only sensitivity (TPR) and specificity (1 - FPR) needs to be estimated
  - LRp, LRm, AUC, PPV, NPV, and accuracy are derived using -nlcom-
- PPV, NPV, and accuracy are dependent on the prevalence (What-if)
- The prevalence is more volatile than the sensitivity and the specificity
  - reuse sensitivity and specificity for different prevalences (What-if)
- Consider designs estimating TPR (FPR) in subpopulations dominated by positive (negative) cases
  - smaller samples (real world?)
  - risk of bias
  - replace sample prevalence with population prevalence
- Eckert and Vach (2023) demonstrates a slightly different approach

NORTH DENMARK REGION

References on the next page

References

- Bland, Martin. 2015. *An Introduction to Medical Statistics*. Fourth edition. Oxford University Press.
- Buderer, Nancy M. Fenn. 1996. "Statistical Methodology: I. Incorporating the Prevalence of Disease into the Sample Size Calculation for Sensitivity and Specificity." *Academic Emergency Medicine* 3 (9): 895–900. https://doi.org/https://doi.org/10.1111/j.1553-2712.1996.tb03538.x.
- Cohen, Jérémie F, Daniël A Korevaar, Douglas G Altman, et al. 2016. "STARD 2015 Guidelines for Reporting Diagnostic Accuracy Studies: Explanation and Elaboration." *BMJ Open* (LONDON) 6 (11): e012799–.
- Cummings, Peter. 2009. "Methods for Estimating Adjusted Risk Ratios." *The Stata Journal* 9 (2): 175–96. https://doi.org/10.1177/1536867X0900900201.
- Eckert, Maren, and Werner Vach. 2023. "Visualizing Uncertainty in a Two-Dimensional Estimate Using Confidence and Comparison Regions." *The Stata Journal* (Los Angeles, CA) 23 (2): 455–90.
- Pepe, M. S. 2003. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford Statistical Science Series. OUP Oxford. https://books.google.dk/books?id=oiiQDwAAQBAJ.
- Rothman, K. J., J. E. Gallacher, and E. E. Hatch. 2013. "Why Representativeness Should Be Avoided." *International Journal of Epidemiology* (OXFORD) 42 (4): 1012–14.
- Whiting, Penny F. 2011. "QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies." *Annals of Internal Medicine* (PHILADELPHIA) 155 (8): 529–36.
- Whiting, Penny F., Anne W. S. Rutjes, Marie E. Westwood, and Susan Mallett. 2013. "A Systematic Review Classifies Sources of Bias and Variation in Diagnostic Test Accuracy Studies." *Journal of Clinical Epidemiology* (NEW YORK) 66 (10): 1093–104.
- Whiting, Penny, Anne W. S. Rutjes, Johannes B. Reitsma, Afina S. Glas, Patrick M. M. Bossuyt, and Jos Kleijnen. 2004. "Sources of Variation and Bias in Studies of Diagnostic Accuracy: A Systematic Review." *Annals of Internal Medicine* (PHILADELPHIA) 140 (3): 189–202.