Improved estimation of Area under the ROC curve under a more efficient sampling design

Jingjing Yin and Yi Hao

Georgia Southern University

Clinical Trials 2015, July 27

Medical diagnostic test



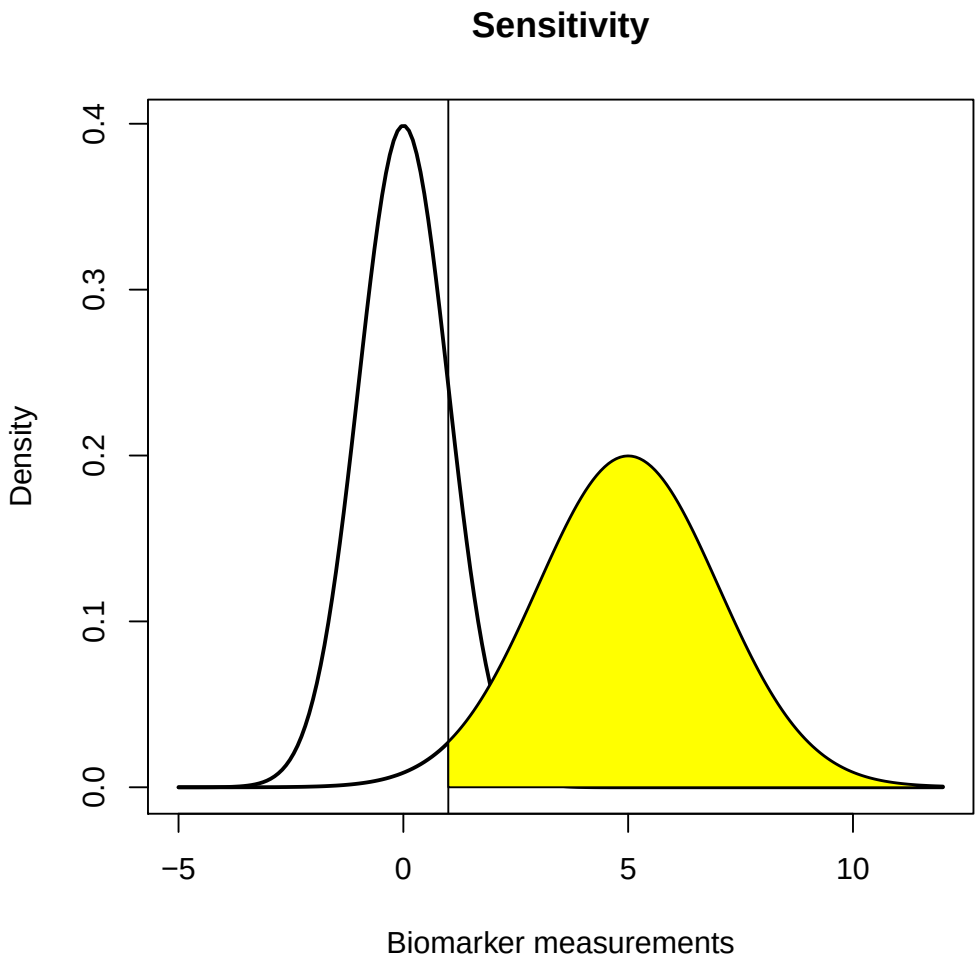
- Objective: to identify diseased individuals through biomarker measurements.
- Continuous marker as the predictor for a binary outcome: a threshold value c is needed. Assume diseased subjects have larger marker values.
  - For a random marker measurement Y, when $Y > c$, classified as diseased, otherwise, healthy (E.g., Systolic $>$ 140 indicates high blood pressure).

Diagnostic probabilities

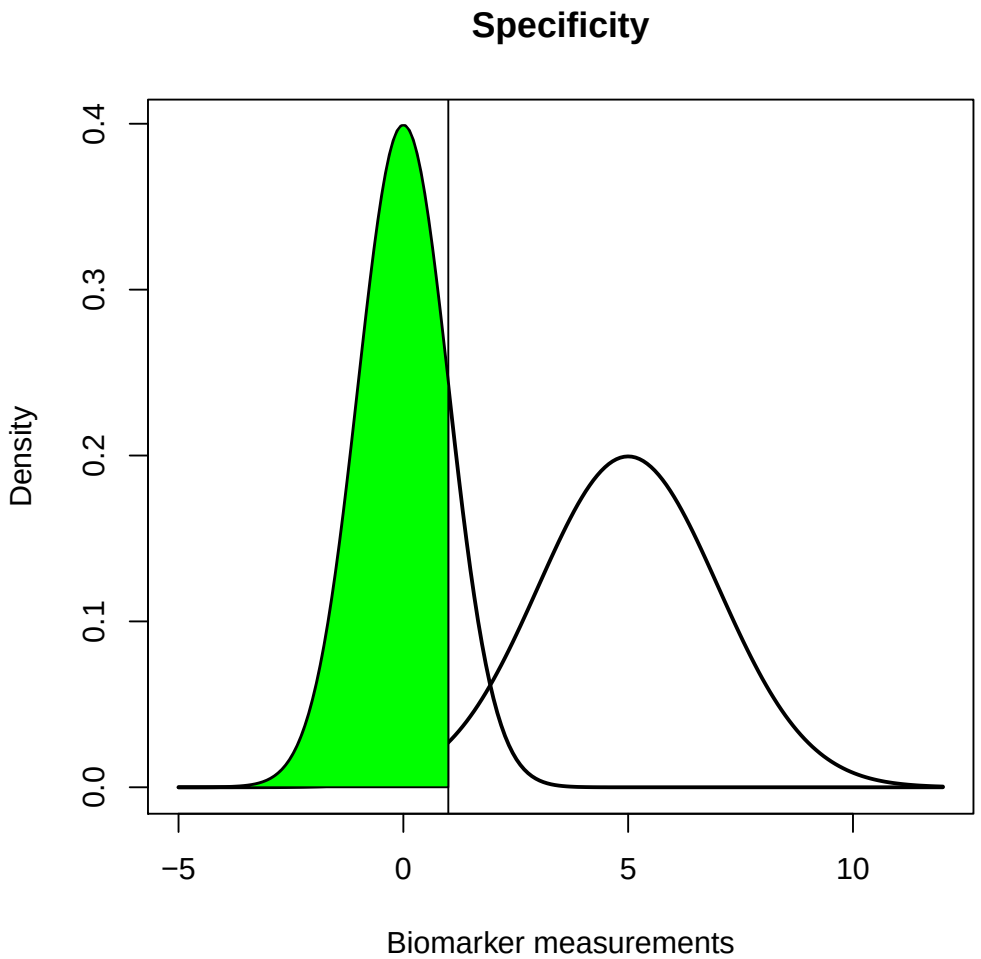
With a gold standard, the true classification is known.

		Truth	
		Diseased	Healthy
Diagnosis	Diseased	☺	FP
	Healthy	FN	☺

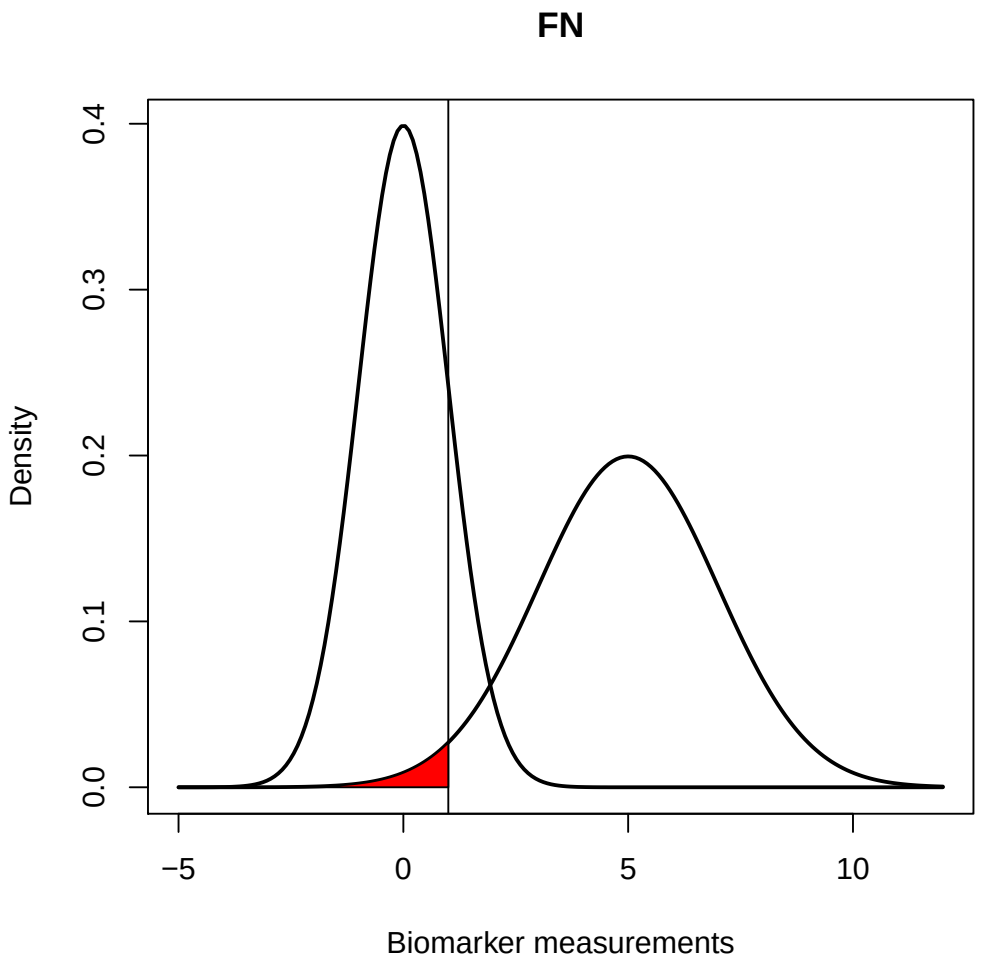
Diagnostic probabilities



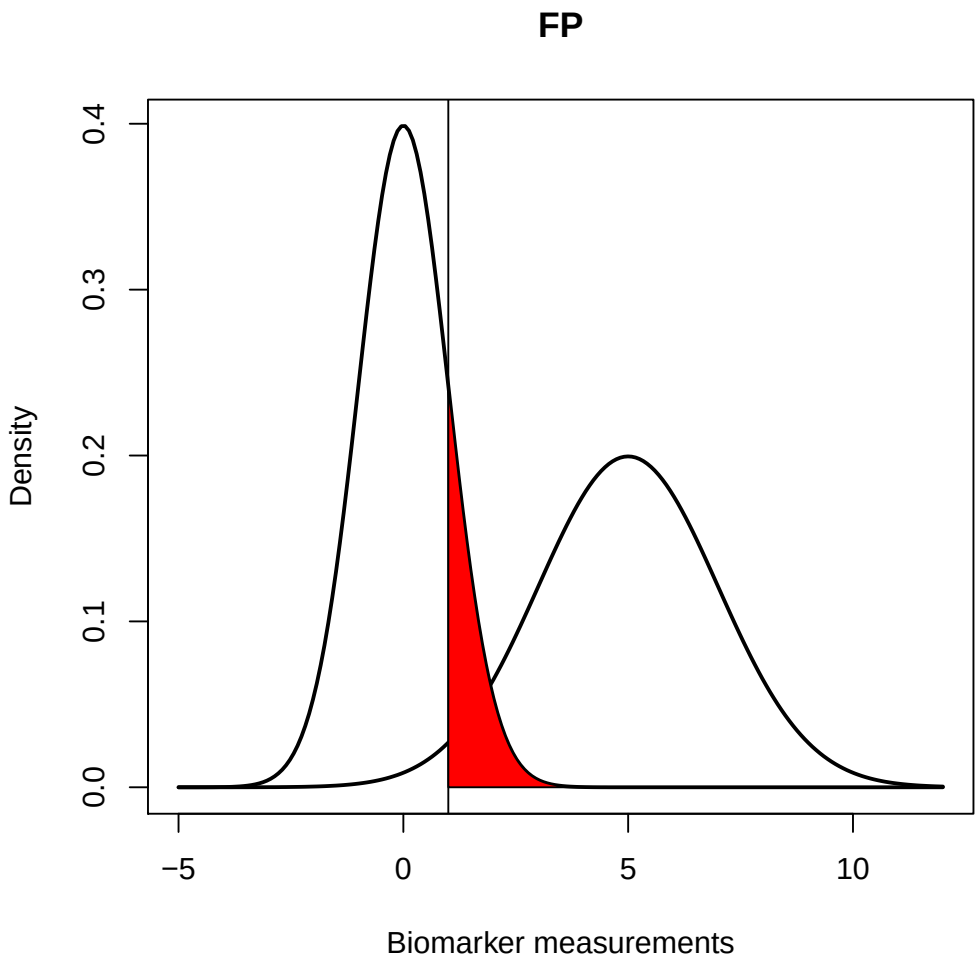
Diagnostic probabilities



Diagnostic probabilities



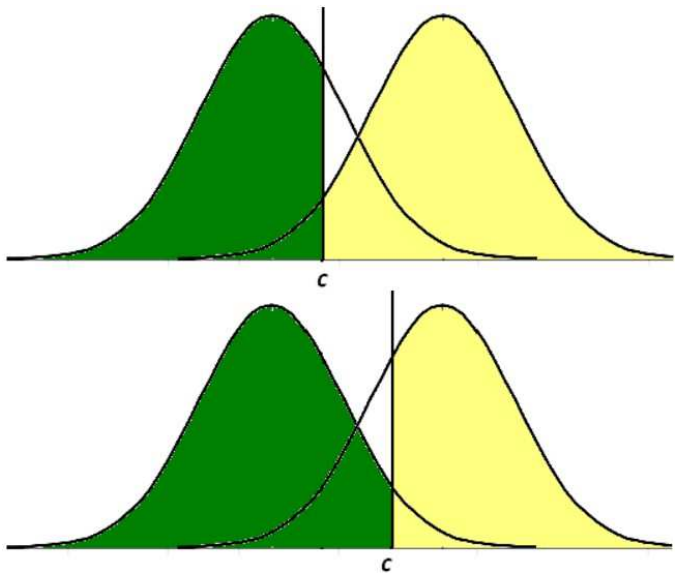
Diagnostic probabilities



Trade-off between sensitivity and specificity

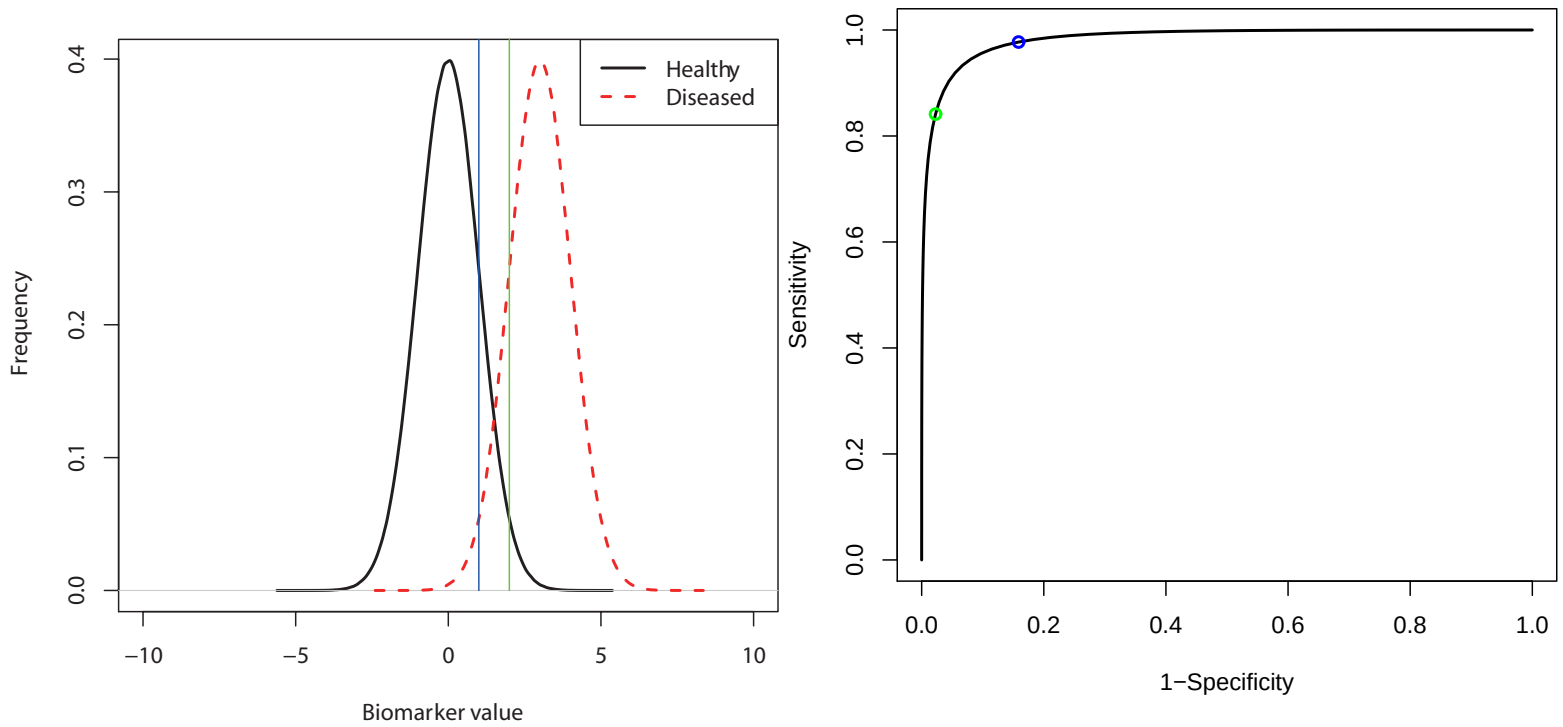
As threshold value increases, sensitivity decreases while specificity increases.

- What if a positive diagnosis triggers an invasive procedure, e.g. Breast cancer?
  –High specificity
- What if the disease is life threatening and an inexpensive and effective treatment is available, e.g. Cervical cancer?
  –High sensitivity

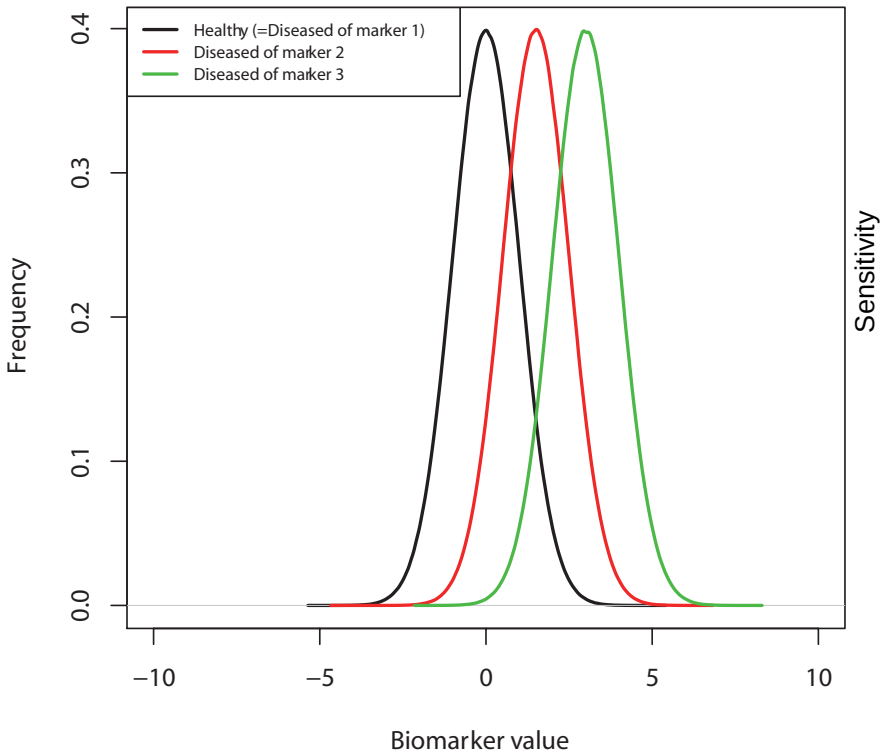


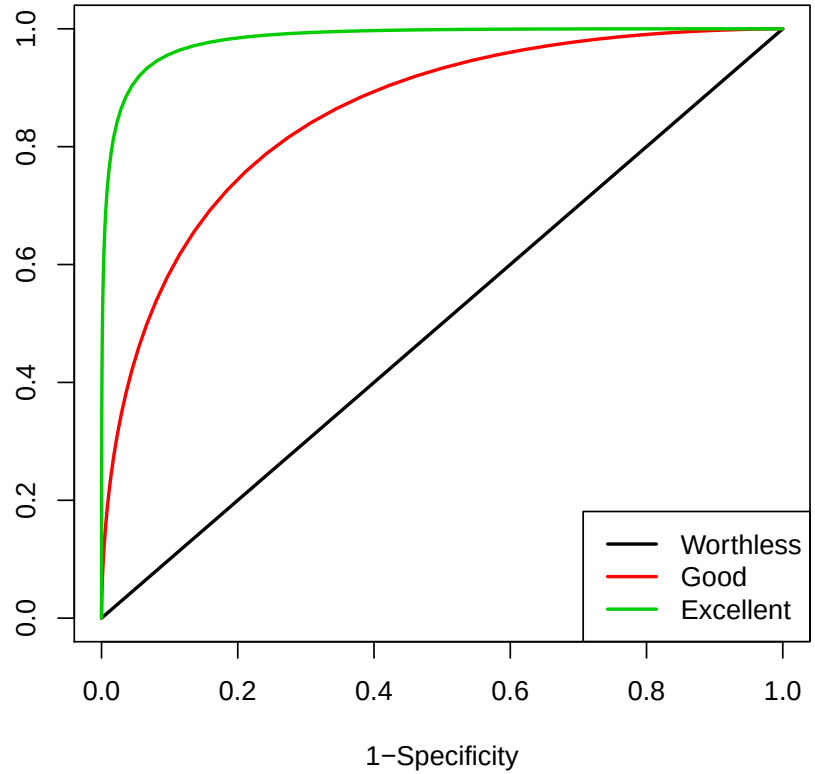
The ROC curve demonstrates such trade-off across all thresholds

The *ROC* curve is a plot of $se(c)$ versus $FPR(c)$ $(= 1 - sp(c))$ across all thresholds (c).



The ROC curve indicates the diagnostic performance of a marker





Summary indices of the ROC curve

The summary indices of the *ROC* curve are more useful and direct than the simple visual interpretation of the *ROC* curve alone. Some popular indices are

- Area under the *ROC* curve (*AUC*)
- Partial *AUC* (*pAUC*) given a range of specificity

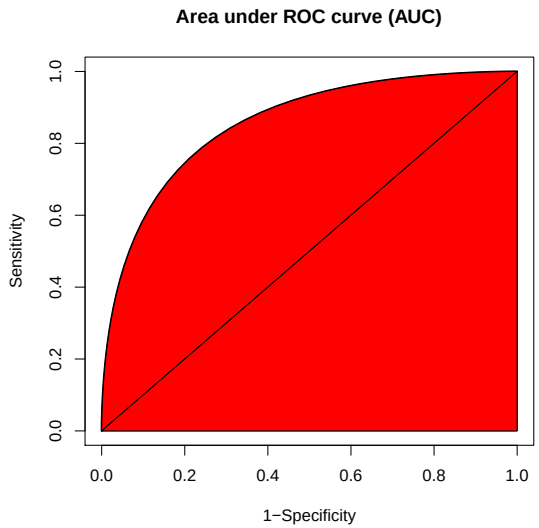
Area under the ROC curve (AUC)

AUC calculation

$$AUC = \int_0^1 ROC(t)\,dt = \int_0^1 1 - F_{Y_1}(F_{Y_2}^{-1}(t))\,dt,$$

where t stands for specificity.

- $AUC = Pr(Y_1 \geq Y_2)$.
- For clinical diagnosis, $AUC \in [0.5, 1]$ and $AUC > 0.8$ is determined to be good.



Background

Introduction and motivation

- Motivating example

Motivating Example

Sometimes, the cost of taking the full measurements on some diagnostic biomarker is very high while drawing samples and ranking them can be easy. An example:

- The study of prostate cancer biomaker free-PSA (fpsa): to obtain the optimal diagnostic cut-off point corresponds to the largest sum of sensitivity and specificity for diagnosing prostate cancer patient.
- Each PSA blood test costs from $100 to $400 and takes 2 to 3 days to get the results.

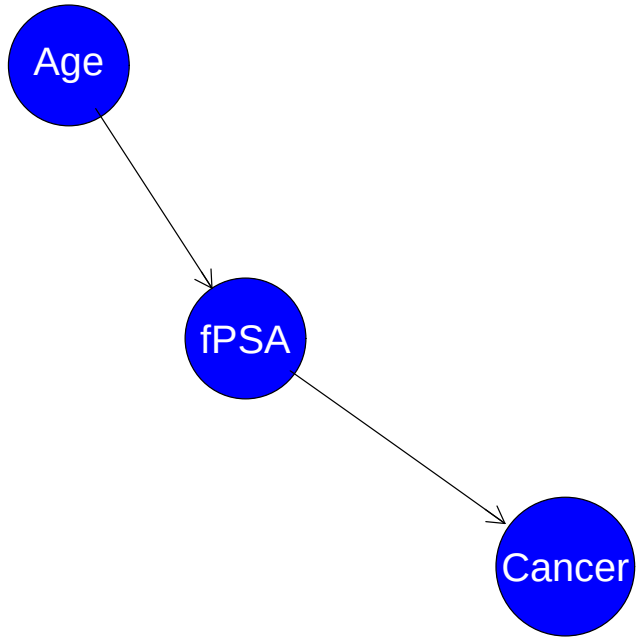
Can we design a diagnostic test evaluation study to save the cost and time for taking measurements of the biomarker by selecting a more representative sample resulting in more efficient estimation of the optimal threshold?

– Yes! Select fpsa **“judgment-ranked” samples**.

Age is an auxiliary variable

We can select the fpsa “judgment-ranked” samples by the ranking of an auxiliary variable age.

- Age is a matching variable for case and control
- Age is significantly associated with fpsa (p=0.008, r=0.25)
- Age information is easy to obtain and almost zero cost



Background

Introduction and motivation

- Sampling procedures

Estimation from a sample

- A valid sample should be representative of the population, otherwise, “Even the best statistical methods cannot help a poorly collected data”.



Sampling methods

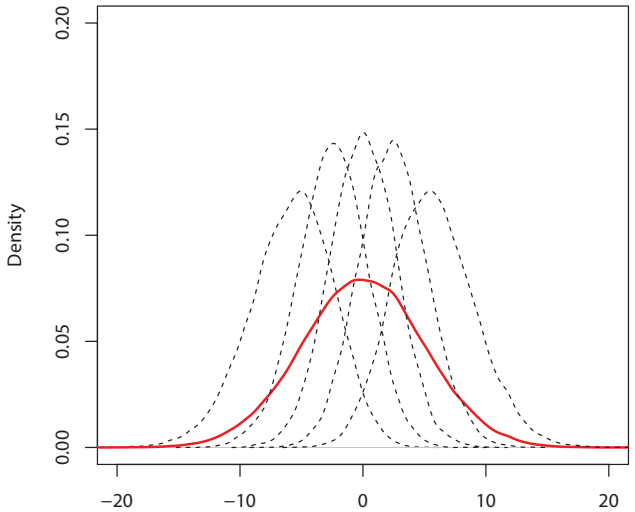
- Most common sampling approach is **simple random sample (SRS)**:
  - A set of $(X_1, X_2, ..., X_n)$ i.i.d distributed
- Some techniques to improve SRS:
  - Stratified sampling: divide population into subgroups and employ SRS within each subgroups. However, need prior knowledge of the structure of the underlying population before collecting data.
  - **Randed set sampling (RSS)** (McIntyre (1952)): collect observations based on the ranks so that the samples are more likely to span the full range of values in the population.

Structure of RSS data set

How to employ RSS with set size r and cycle size m?

Cycle 1	$X_{[1]1}$	$X_{[2]1}$	...	$X_{[r]1}$
Cycle 2	$X_{[1]2}$	$X_{[2]2}$	...	$X_{[r]2}$
:	:	:	:	:
Cycle m	$X_{[1]m}$	$X_{[2]m}$	...	$X_{[r]m}$

The RSS is still independent but no longer identically distributed.



Improved estimation of AUC on RSS

- Simulation

Selected results of *AUC* under normal

Sample size	auc rb				auc sd				auc rmse				red.%
	SRS	RSS			SRS	RSS			SRS	RSS			
		$r=2$	$r=4$	$r=5$		$r=2$	$r=4$	$r=5$		$r=2$	$r=4$	$r=5$	
$(\mu_1,\sigma_1^2)=(0.5622365,3.9250)$, $(\mu_2,\sigma_2^2)=(0,1)$													
$(20,20)$	-0.0101	-0.0141	-0.0138	-0.0143	0.0846	0.0706	0.0545	0.0501	0.0848	0.0712	0.0551	0.0508	0.64
$(40,40)$	-0.0115	-0.0127	-0.0118	-0.0121	0.0615	0.0497	0.0392	0.0348	0.0619	0.0503	0.0398	0.0355	0.67
$(60,40)$	-0.0140	-0.0085	-0.0119	-0.0098	0.0577	0.0486	0.0383	0.0336	0.0583	0.0488	0.0389	0.0342	0.66
$(60,60)$	-0.0111	-0.0092	-0.0122	-0.0111	0.0494	0.0400	0.0321	0.0299	0.0498	0.0404	0.0329	0.0307	0.62
$(60,80)$	-0.0095	-0.0110	-0.0099	-0.0097	0.0444	0.0367	0.0292	0.0259	0.0447	0.0373	0.0298	0.0266	0.65
$(80,80)$	-0.0086	-0.0108	-0.0100	-0.0089	0.0439	0.0361	0.0275	0.0254	0.0442	0.0367	0.0281	0.0260	0.65
$(100,100)$	-0.0072	-0.0079	-0.0078	-0.0085	0.0391	0.0313	0.0247	0.0228	0.0393	0.0316	0.0251	0.0234	0.65
$(\mu_1,\sigma_1^2)=(0.4316105,0.2547)$, $(\mu_2,\sigma_2^2)=(0,1)$													
$(20,20)$	-0.0171	-0.0182	-0.0185	-0.0173	0.0835	0.0695	0.0535	0.0516	0.0842	0.0705	0.0548	0.0528	0.61
$(40,40)$	-0.0154	-0.0149	-0.0147	-0.0153	0.0590	0.0483	0.0386	0.0338	0.0598	0.0493	0.0397	0.0352	0.65
$(60,40)$	-0.0127	-0.0142	-0.0094	-0.0153	0.0499	0.0420	0.0336	0.0311	0.0506	0.0430	0.0341	0.0326	0.58
$(60,60)$	-0.0117	-0.0118	-0.0128	-0.0144	0.0478	0.0404	0.0312	0.0298	0.0484	0.0411	0.0323	0.0312	0.58
$(60,80)$	-0.0130	-0.0151	-0.0122	-0.0132	0.0462	0.0392	0.0310	0.028	0.0470	0.0404	0.0320	0.0293	0.61
$(80,80)$	-0.0123	-0.0131	-0.0140	-0.0149	0.0425	0.0352	0.0276	0.0255	0.0433	0.0362	0.0291	0.0273	0.60
$(100,100)$	-0.0098	-0.0117	-0.0133	-0.0119	0.0373	0.0309	0.0243	0.023	0.0379	0.0319	0.0258	0.0243	0.59

Simulation summary

The following conclusions are universally true across all simulation settings:

- Larger the set size of RSS, smaller RMSE.
- RSS has similar bias (both are negligible) but much smaller RMSE compared to SRS.
- Smaller the sample size, more advantageous of RSS over SRS.
- RSS reduces MSE by approx. 60% (i.e., RSS saves 60% sample sizes to achieve same efficiency of estimators by SRS).

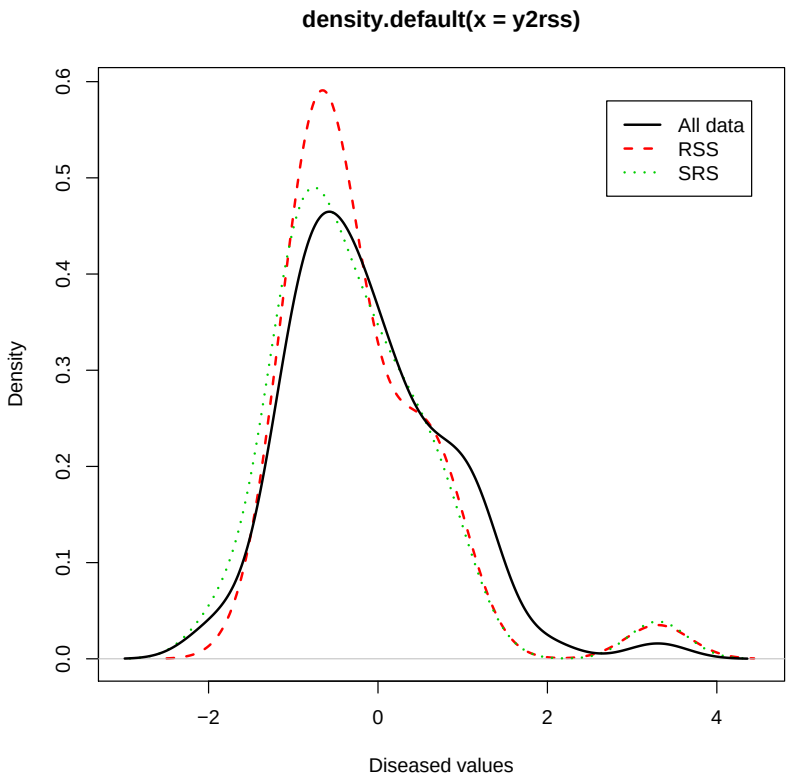
Improved estimation of Youden index on RSS

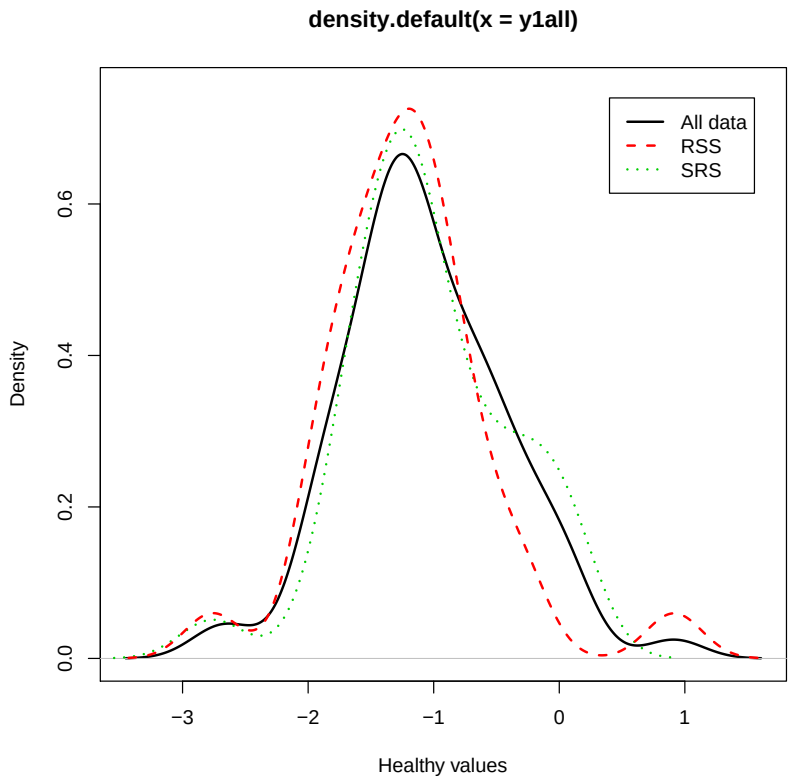
- Data example

PSA study revisited

- A case-control study of 71 prostate cancer patients and 70 controls
- Select two sub-samples, and RSS and SRS each is of size $(30, 30)$
- RSS set size $r = 5$

Density estimation of the total sample





Estimates for PSA

Methods	$\hat{AUC}$	$\hat{Var}(\hat{AUC})^*$	95%$C.I.$ of AUC
Total sample (Population)	0.7913		
RSS sub-sample	0.8317	0.0022	(0.7398,0.9236)
SRS sub-sample	0.7354	0.0033	(0.6228,0.8480)

Thank you! Questions?

.

References: