Calculating diagnostic test characteristics in the presence of clustering

A practical guide

Katalin Tamási, PhD

k.tamasi@umcg.nl

Departments of Epidemiology and Neurosurgery

Unit of Medical Statistics and Decision Making

Nov 23, 2021

Important concepts

$$\text{Probability P(A)} = \frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$$

- Prob. of rolling a 6 $P(6)$?

Important concepts

$$\text{Probability P(A)} = \frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?

Important concepts

$$\text{Probability P(A)} = \frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

Important concepts

$$\text{Probability P(A)} = \frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

- $P(SIDS)$ = 1/8500

Important concepts

$$\text{Probability P(A)} = \frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

- $P(SIDS) = 1/8500$
- $P(SIDS|SIDS) = ?$

Important concepts

$$\text{Probability P(A)} = \frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

- $P(SIDS) = 1/8500$
- $P(SIDS|SIDS) = ?$
- $P(SIDS) * P(SIDS|SIDS) = ?$

Important concepts

Probability P(A) = $\frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

- $P(SIDS)$ = 1/8500
- $P(SIDS|SIDS)$ = ?
- $P(SIDS) * P(SIDS|SIDS)$ = ?

Independence assumption

- $P(6) = P(6|6)$

Important concepts

Probability P(A) = $\frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

- $P(SIDS)$ = 1/8500
- $P(SIDS|SIDS)$ = ?
- $P(SIDS) * P(SIDS|SIDS)$ = ?

Independence assumption

- $P(6) = P(6|6)$
- $P(SIDS) \neq P(SIDS|SIDS)$

Important concepts

Probability P(A) = $\frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

- $P(SIDS)$ = 1/8500
- $P(SIDS|SIDS)$ = ?
- $P(SIDS) * P(SIDS|SIDS)$ = ?

Independence assumption

- $P(6) = P(6|6)$
- $P(SIDS) \neq P(SIDS|SIDS)$
- $P(D) \neq P(D|H_0)$

Important concepts

Probability P(A) = $\frac{\text{number of outcomes in favor of A}}{\text{total number of possible outcomes}}$

- Prob. of rolling a 6 $P(6)$?
- Prob. of rolling a 6 *after having rolled a 6* $P(6|6)$?
- Prob. of rolling two sixes, one after the other $P(6) * P(6|6)$?

- $P(SIDS)$ = 1/8500
- $P(SIDS|SIDS)$ = ?
- $P(SIDS) * P(SIDS|SIDS)$ = ?

Independence assumption

- $P(6) = P(6|6)$
- $P(SIDS) \neq P(SIDS|SIDS)$
- $P(D) \neq P(D|H_0)$

Law of total probability

- $P(1) + P(2) + ... + P(6) = 1$

Screening example

Suppose we have a patient who we want to test for disease D using test T. We know that the sensitivity of the test is 80% and the specificity is 95%. If the test shows up positive, what's the probability that the patient has the disease?

Screening example

Suppose we have a patient who we want to test for disease D using test T. We know that the sensitivity of the test is 80% and the specificity is 95%. If the test shows up positive, what's the probability that the patient has the disease?

Not enough information to answer the question!

Screening example

Suppose we have a patient who we want to test for disease D using test T. We know that the sensitivity of the test is 80% and the specificity is 95%. If the test shows up positive, what's the probability that the patient has the disease?

Not enough information to answer the question!

- Sensitivity: $P(T|D) = 0.8$

Screening example

Suppose we have a patient who we want to test for disease D using test T. We know that the sensitivity of the test is 80% and the specificity is 95%. If the test shows up positive, what's the probability that the patient has the disease?

Not enough information to answer the question!

- Sensitivity: $P(T|D) = 0.8$
- Specificity: $P(-T|-D) = 0.95$

Screening example

Suppose we have a patient who we want to test for disease D using test T. We know that the sensitivity of the test is 80% and the specificity is 95%. If the test shows up positive, what's the probability that the patient has the disease?

Not enough information to answer the question!

- Sensitivity: $P(T|D) = 0.8$
- Specificity: $P(-T|-D) = 0.95$
- Positive predictive value: $P(D|T) = ?$

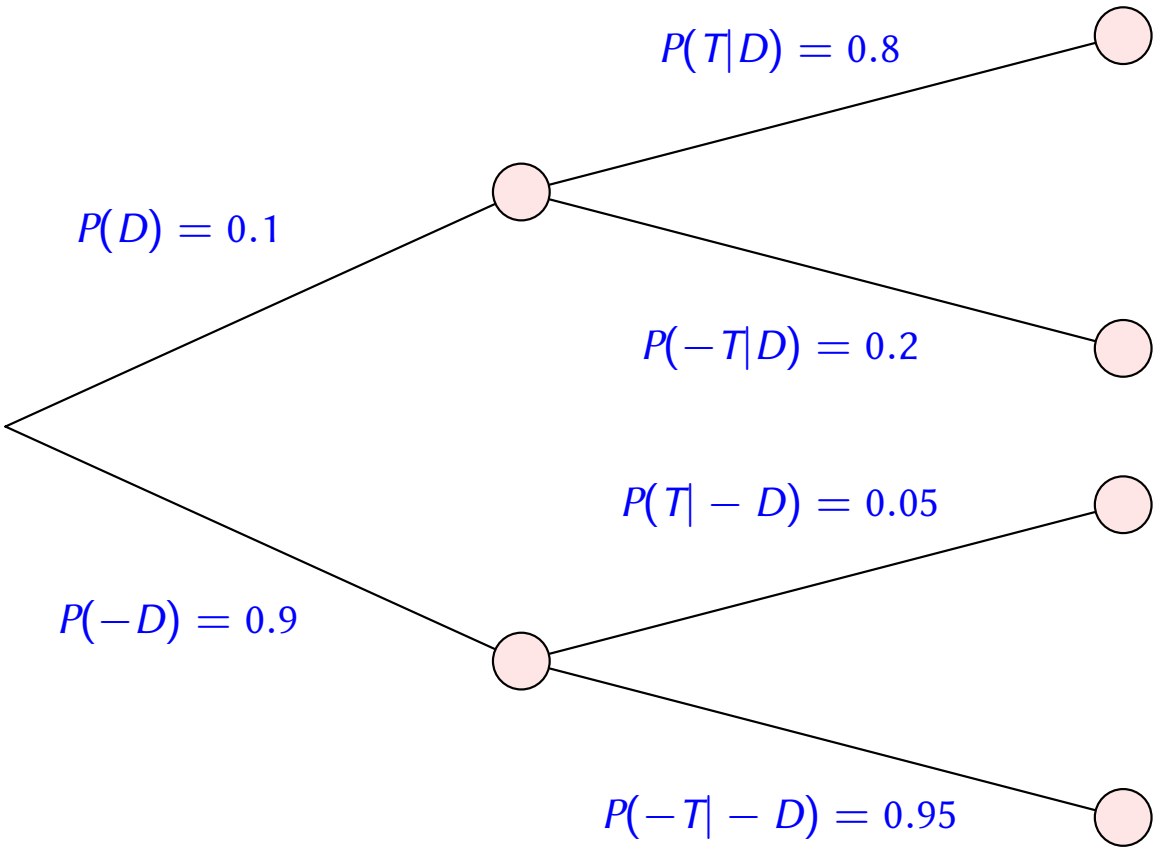
Screening example

Suppose we have a patient who we want to test for disease D using test T. We know that the sensitivity of the test is 80% and the specificity is 95%. If the test shows up positive, what's the probability that the patient has the disease?

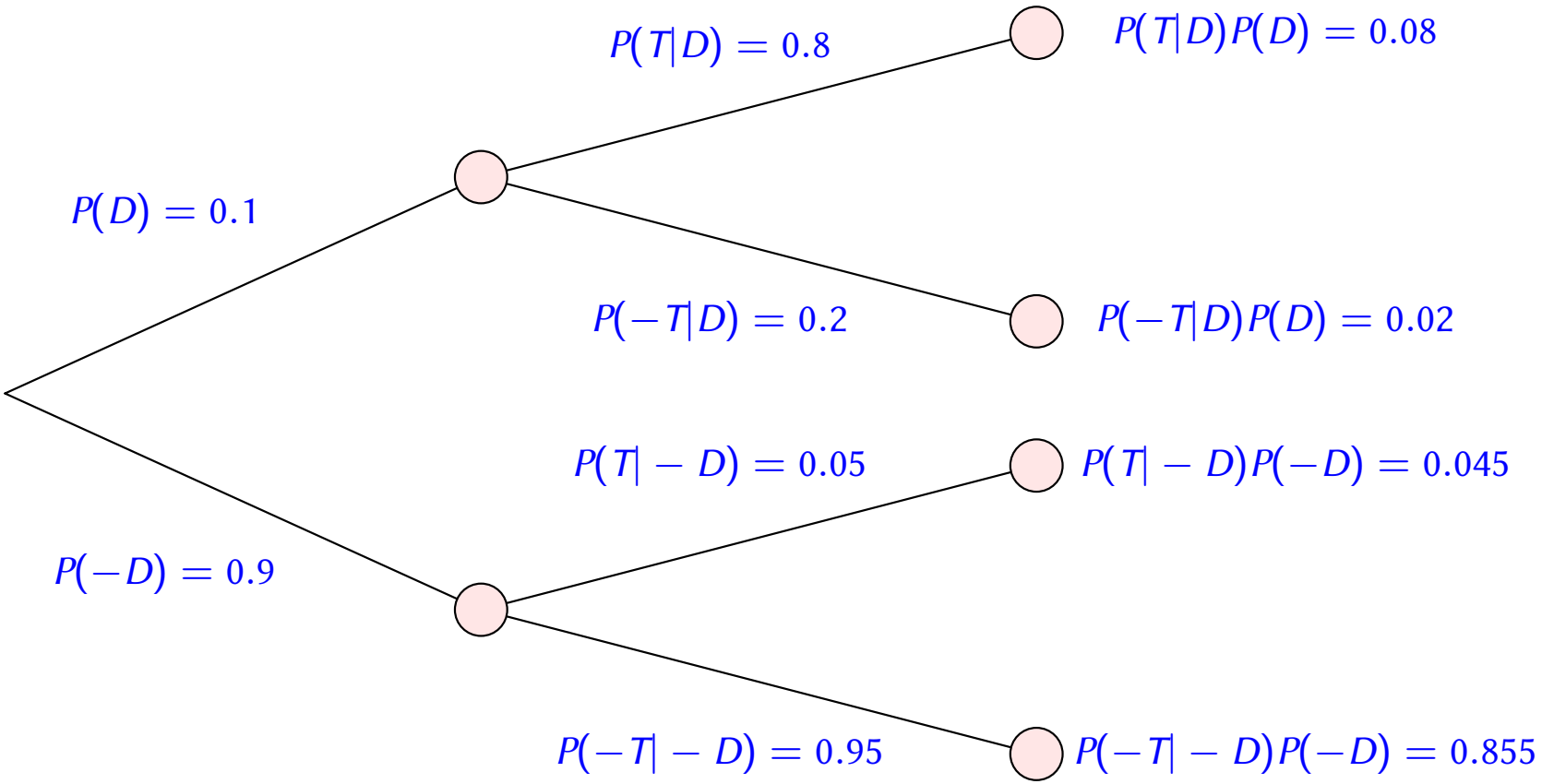
Not enough information to answer the question!

- Sensitivity: $P(T|D) = 0.8$
- Specificity: $P(-T|-D) = 0.95$
- Positive predictive value: $P(D|T) = ?$
- Probability of having the disease (prevalence / incidence risk): $P(D) = 0.1$

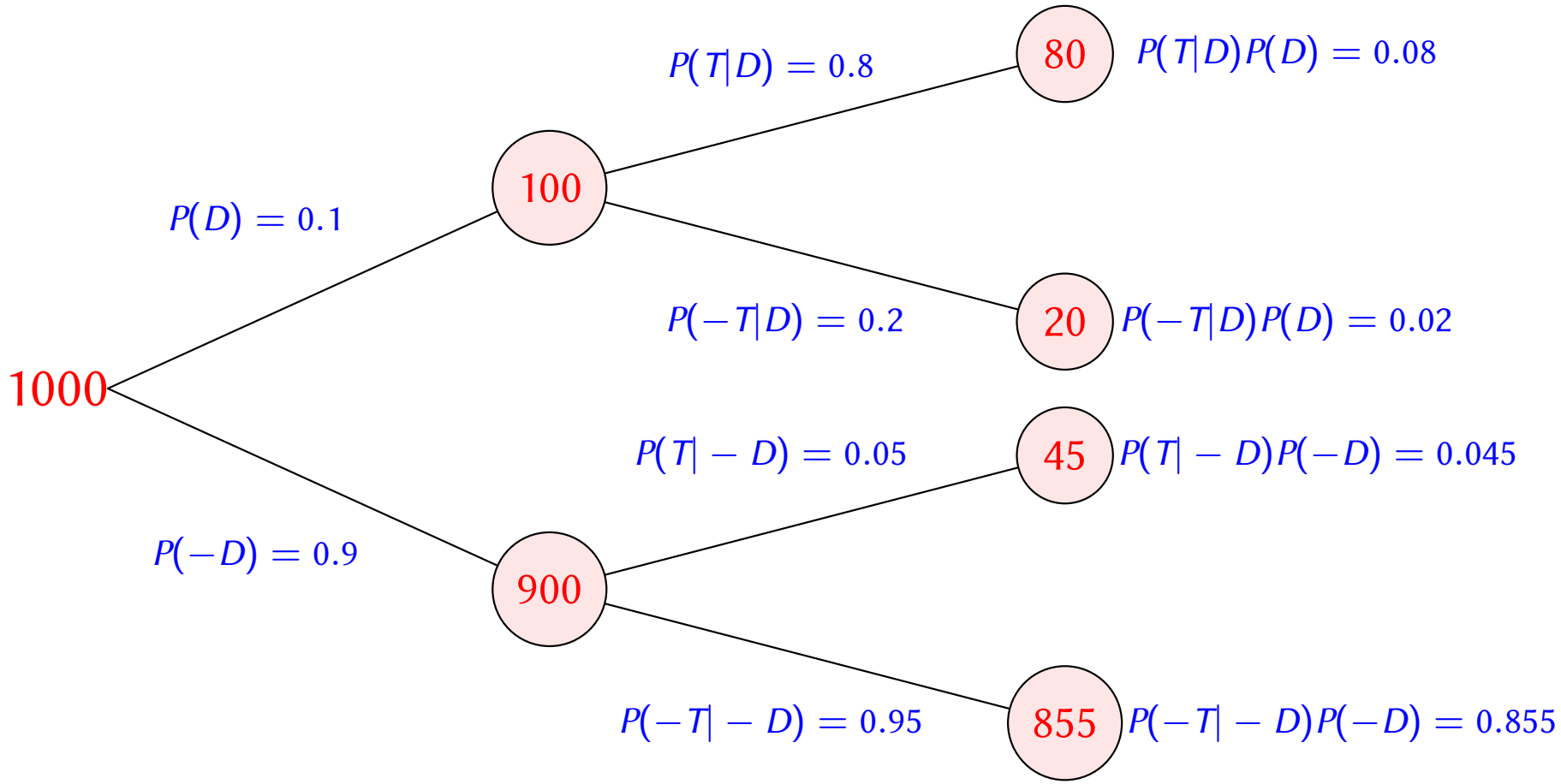
Probability tree



Probability tree



Make it concrete with numbers



$$P(D|T) = \frac{\text{P(T|D)P(D)}}{\text{P(T|D)P(D) + P(T|-D)P(-D)}} = \frac{80}{80+45} = 64\%$$

Screening example cont'd

Before the test, our best estimate if the patient has the disease is:

- Population prevalence $P(D) = 0.1$

After a positive test, we update the estimate to:

- Positive predictive value: $P(D|T) = 0.64$

The estimate is updated again when new information becomes available

$\rightarrow$ Bayesian updating

Screening example cont'd

Suppose a new patient comes in. What's the probability that they *don't* have the disease?

Screening example cont'd

Suppose a new patient comes in. What's the probability that they *don't* have the disease?

- $1 - P(D) = P(-D) = 0.9$

Screening example cont'd

Suppose a new patient comes in. What's the probability that they *don't* have the disease?

- $1 - P(D) = P(-D) = 0.9$

Their test turns out negative. Now what's the probability that they don't have the disease?

Screening example cont'd

Suppose a new patient comes in. What's the probability that they *don't* have the disease?

- $1 - P(D) = P(-D) = 0.9$

Their test turns out negative. Now what's the probability that they don't have the disease?

- Negative predictive value

Screening example cont'd

Suppose a new patient comes in. What's the probability that they *don't* have the disease?

- $1 - P(D) = P(-D) = 0.9$

Their test turns out negative. Now what's the probability that they don't have the disease?

- Negative predictive value
- $P(-D|-T) = \frac{\text{P(-T|-D)P(-D)}}{\text{P(-T|-D)P(-D) + P(-T|D)P(D)}} = \frac{855}{855+20} \approx 98\%$

Screening example cont'd

Suppose a new patient comes in. What's the probability that they *don't* have the disease?

- $1 - P(D) = P(-D) = 0.9$

Their test turns out negative. Now what's the probability that they don't have the disease?

- Negative predictive value
- $P(-D|-T) = \frac{\text{P(-T|-D)P(-D)}}{\text{P(-T|-D)P(-D) + P(-T|D)P(D)}} = \frac{855}{855+20} \approx 98\%$

Without any specific information, we estimated 90%, with the negative test, we updated the estimate to $\approx 98\%$

Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Not enough information to answer the question!

Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Not enough information to answer the question!

- True positive (power): $P(E|H_1) = 0.8$

Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Not enough information to answer the question!

- True positive (power): $P(E|H_1) = 0.8$
- False negative (type II error): $P(-E|H_1) = 0.2$

Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Not enough information to answer the question!

- True positive (power): $P(E|H_1) = 0.8$
- False negative (type II error): $P(-E|H_1) = 0.2$
- True negative (correct retainment of H_0): $P(-E|H_0) = 0.95$

Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Not enough information to answer the question!

- True positive (power): $P(E|H_1) = 0.8$
- False negative (type II error): $P(-E|H_1) = 0.2$
- True negative (correct retainment of H_0): $P(-E|H_0) = 0.95$
- False positive (type I error): $P(E|H_0) = 0.05$

Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Not enough information to answer the question!

- True positive (power): $P(E|H_1) = 0.8$
- False negative (type II error): $P(-E|H_1) = 0.2$
- True negative (correct retainment of H_0): $P(-E|H_0) = 0.95$
- False positive (type I error): $P(E|H_0) = 0.05$
- Prob. of H_1 being true given the evidence: $P(H_1|E) = ?$

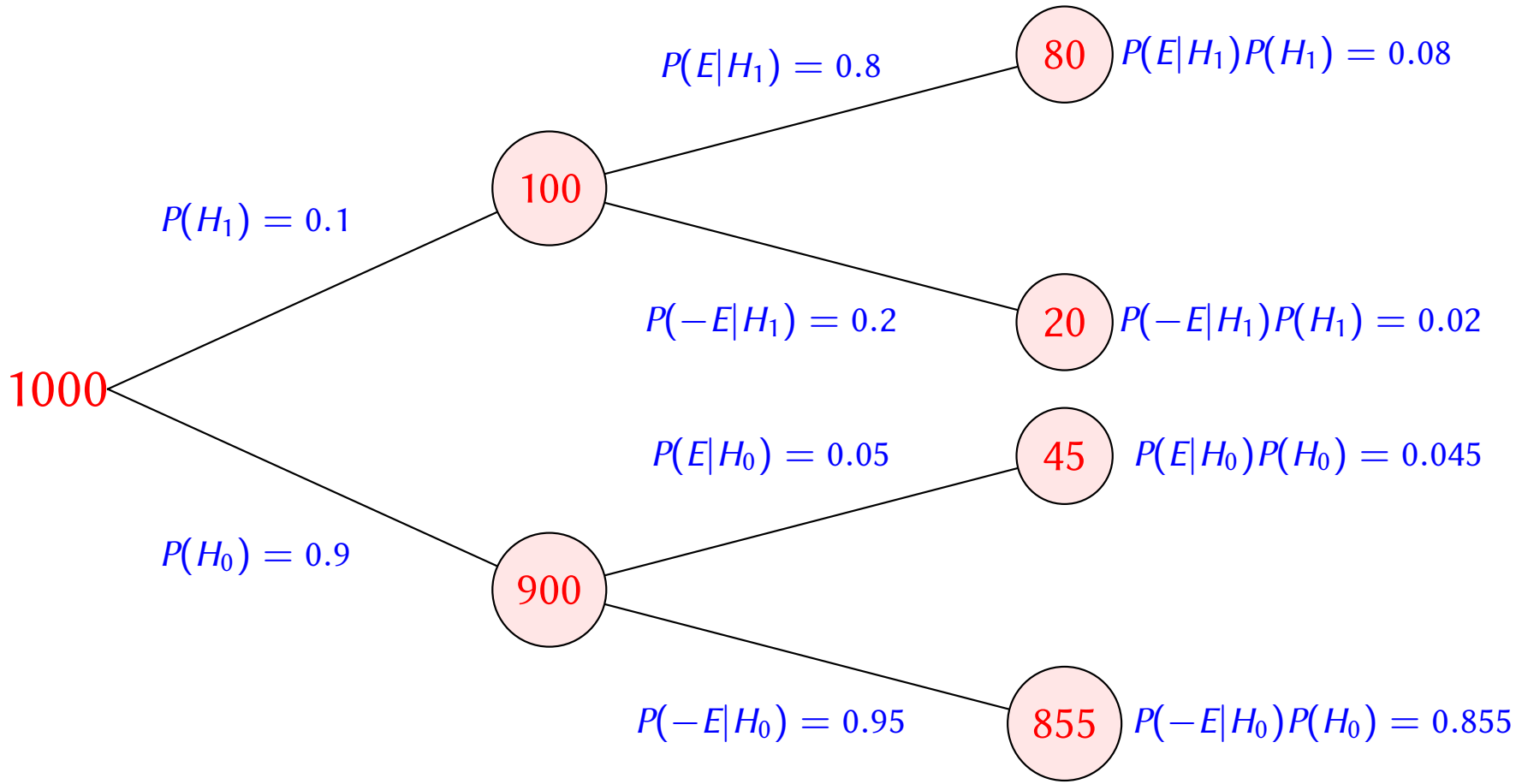
Side note: Hypothesis testing example

H_0: A new intervention is not more effective than the standard.
H_1: A new intervention is more effective than the standard.
E: Evidence supporting H_1
Suppose that our data supports the alternative hypothesis that a new intervention is more effective. What's the probability that the alternative hypothesis is correct given the evidence?

Not enough information to answer the question!

- True positive (power): $P(E|H_1) = 0.8$
- False negative (type II error): $P(-E|H_1) = 0.2$
- True negative (correct retainment of H_0): $P(-E|H_0) = 0.95$
- False positive (type I error): $P(E|H_0) = 0.05$
- Prob. of H_1 being true given the evidence: $P(H_1|E) = ?$
- Prob. of H_1 being true: $P(H_1) = 0.1(?)$ estimated by Iohannidis (2005)

Make it concrete with numbers



$$P(H_1|E) = \frac{P(E|H_1)P(H_1)}{P(E|H_1)P(H_1)+P(E|H_0)P(H_0)} = \frac{80}{80+45} \approx 64\%$$

Hypothesis testing example cont'd

Before the experiment, our best estimate for H_1 being true is:

- $P(H_1) = 0.1$

After a experiment, we update the estimate to:

- Prob. of H_1 being true given the evidence: $P(H_1|E) = 0.64$
- Still not very certain!

The estimate is updated again when new information becomes available
$\rightarrow$ Bayesian updating

Fundamental concepts in diagnostic testing: Review

1. Prevalence-independent characteristics
   - Sensitivity $P(T|D)$
   - Specificity $P(-T|-D)$
2. Prevalence-dependent characteristics
   - Contingent on the underlying (disease) prevalence $P(D)$
     - Not always appropriate to calculate from study, e.g. case-control studies, oversampling
   - Positive predictive value $P(D|T)$
   - Negative predictive value $P(-D|-T)$
   - Can be calculated from Bayes Theorem (/ natural frequency trees) using prevalence, sensitivity, and specificity
   - Prevalence ↑: PPV ↑ NPV ↓
3. Other measures (not discussed today)
   - ROC curves / AUC (continuous outcomes)
   - Positive / negative likelihood ratio
   - Accuracy, etc.

Possible forms of clustering / multiplicity

- Multiple observations per patient
  - Repeated testing: multiple simultaneous segments, e.g., teeth
  - Serial testing: longitudinal data, e.g., in monitoring studies
- Multiple sites per patient
  - E.g., left / right artery
- Multiple tests per patient (can be paired study)
  - Parallel testing: reference test, index test 1, index test 2
- Multiple raters / surgical teams / centers etc.

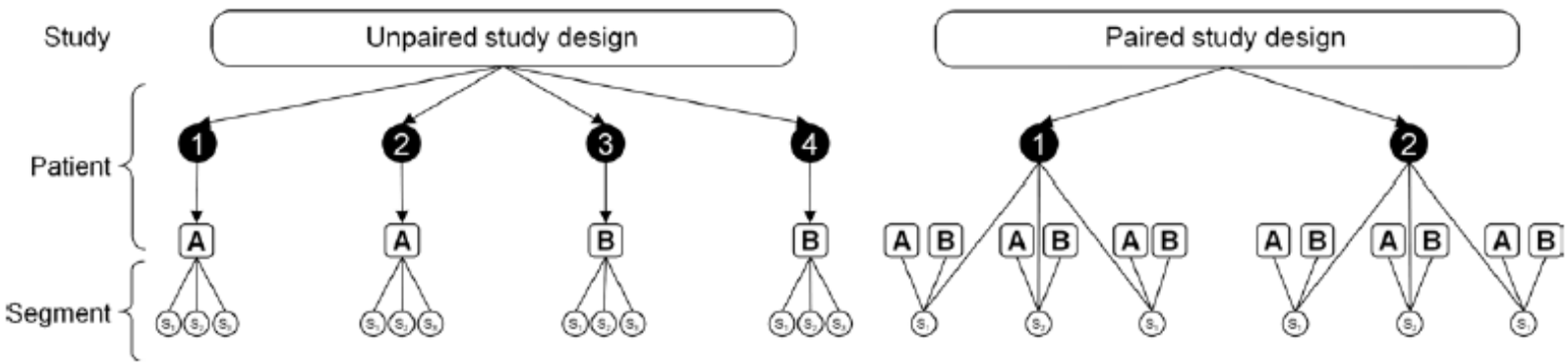


Figure 1: Schematic representation of the unpaired and paired study design. Black circles = individual patients, *A* = test A, and *B* = test B. s_1, s_2, and s_3 = multiple segments that are observed per patient. In a paired design, each patient undergoes test A and B. Therefore, not only the correlation between segments in patients, but also the correlation of two test results for each segment should be accounted for. The same concept applies to a design with two readers, two scan protocols, and so forth.

Genders *et al.* (2012)

Why is adjusting for clustering important?

- Within-cluster observations more similar than between-cluster ones
  - E.g., observations from a patient more likely to resemble each other due to patient characteristics
  - Indicated by a positive intraclass-correlation coefficient (ICC)
- Possible conditional dependence
  - Updating our best estimate with each new information
  - Serial testing: $P(T_2|T_1 \cap D) \neq P(T_2|D)$
  - Parallel testing: $P(T_B|T_A \cap D) \neq P(T_B|D)$
    - Tests taken closer in time / similar tests prone to give correlated results
    - Sufficiently distinct tests: low correlation Shen *et al.* (2001)
- Disregarding clustering = assuming (conditional) independence of observations or tests
  - Inflates sample size, leads to biased estimates and too-small CI's
  - Especially problematic with
    - Highly correlated observations
    - Complex (multi-level) clustering

Motivation

Even though diagnostic testing fundamental to medicine:

- Require custom coding in SPSS (only ROC curves produced)
- No option to adjust for clustering in MedCalc
- Clustering issue routinely ignored in medical literature

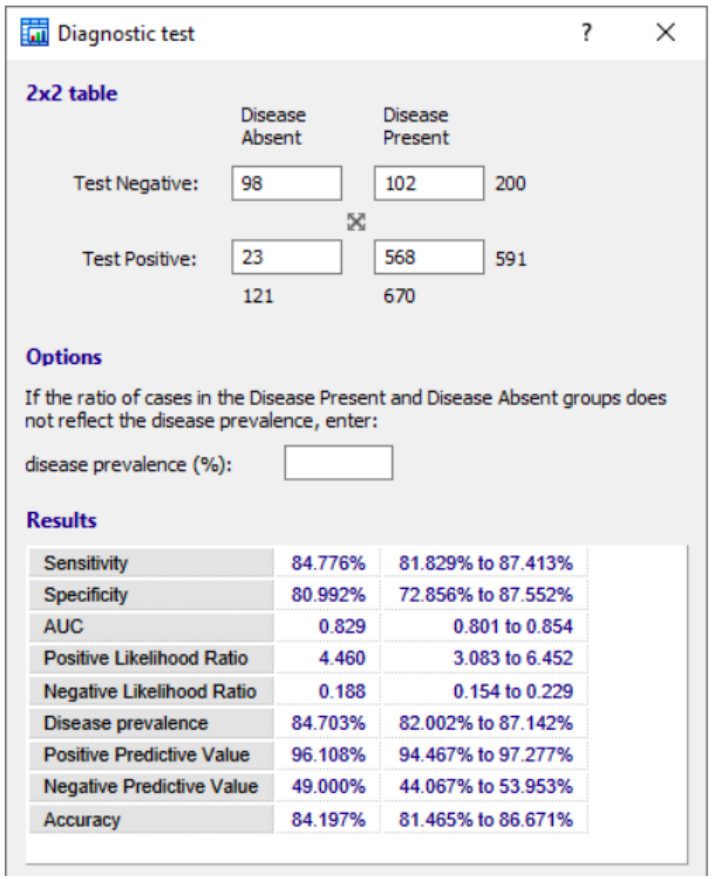


Figure: Medcalc input window. Medcalc Software Ltd. (2021)

Goals

- Spread awareness on issues that clustering creates
- Mediating between medical practice and statistical applications
- Giving healthcare professionals tools in
  - Educational contexts
  - Medical decision making
    - Mitigate the chances of over / underdiagnosis and over / undertreatment
- Determine how to best deal with clustering under what conditions
- Long-term: Get reliable estimates on diagnostic test values conditional on other tests, if usually done in conjunction

Role of visualization

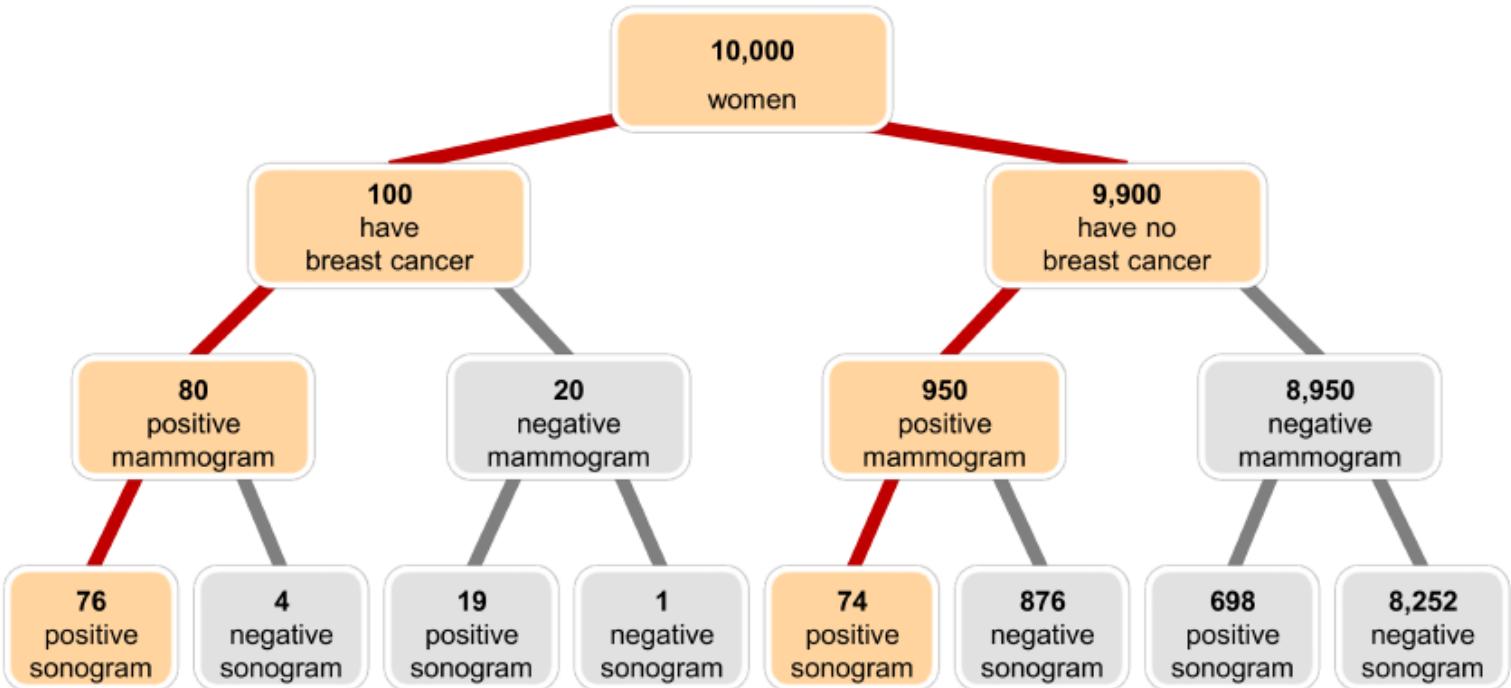
Example

Imagine that you are a physician in a mammography screening center where women without symptoms are screened for breast cancer. In addition to mammograms, you frequently use sonograms as a supplementary medical test to detect breast cancer. At the moment, you are advising a woman who has no symptoms but who has received a positive result from her mammogram as well as a positive result from her sonogram. This woman wants to know what these results mean for her.

Binder *et al.* (2018)

Role of visualization

How many of the women with both positive mammogram and positive sonogram actually have breast cancer?



Binder *et al.* (2018)

- Using natural frequency trees
- Highlighting relevant information

→ 66–68% of medical students can arrive at the correct inference (as opposed to 0–13% of students given probability trees and/or text)

At what level to analyze?

Depends on:

- Analysis stage
  - E.g., sample size calculation usually on patient-level
- The nature of the intervention / clinical consequence
  1. Patient-level intervention: chemotherapy
  2. Segment-level intervention: stent placement vs. balloon angioplasty
  3. Not clear-cut: radiotherapy
- Comparing two (or multiple) levels of analysis
  - Still advisable to pre-specify a primary level of analysis
- Higher-level / aggregate analyses: independence of experimental units can be reasonable assumption
  - E.g., patients without additional clustering
- Lower-level analyses: independence cannot be assumed
  - E.g., sites within patients

e.g., Hujoel, Moulton, Loesche (1990)

Levels of analysis

Patient-level contingency table

Test	Disease		
	+	−	Total
+	35	5	40
−	3	7	10
Total	38	12	50

Segment-level contingency table

Test	Disease		
	+	−	Total
+	50	71	121
−	43	538	581
Total	93	609	702

→ Naive point estimates:

- Sensitivity: 92.1% (35/38)
- Specificity: 58.3% (7/12)
- PPV: 87.5% (35/40)
- NPV: 70% (7/10)

- Sensitivity: 53.8% (50/93)
- Specificity: 88.3% (538/609)
- PPV: 41.3% (50/121)
- NPV: 92.6% (538/581)

Genders *et al.* (2012)

Levels of analysis

Depending on the level of analysis, estimates can vary wildly

Sensitivity / PPV

- Patient-level $>$ Segment-level
- At least 1 positive test + disease $\rightarrow$ Patient-level TP

Specificity / NPV

- Patient-level $<$ Segment-level
- No positive tests + no disease $\rightarrow$ Patient-level TN

Levels of analysis

Remarks

- The higher the intra-cluster correlation, the closer the corresponding estimates of the different levels
  - ICC = 0: independent observations
  - ICC = 1: patient-level = segment-level estimates, though segment-level more precisely estimated
- One contingency table cannot be generated from the other
  - Breakdown by patients needed

Wide format (1 row per patient)

ID	TP	FN	TN	FP
2	1	1	9	2
28	0	0	14	0
31	0	2	10	2
...				
Total	35	3	7	5

Long format (1 row per segment)

ID	Seg.	TP	FN	TN	FP
2	1	1	0	0	0
2	2	0	1	0	0
2	3	0	0	1	0
...					
28	1	0	0	1	0
...					
Total		50	43	71	538

Methods for dealing with clustered data

1. Patient-level analysis (no adjustment for clustering necessary)
   - Binomial proportion (Exact vs. Wald approximation / Logit-transformed / Continuity-corrected)
   - Logistic regression
2. Segment-level analysis
   - No adjustment for clustering
     - Binomial proportion
     - Logistic regression
   - Adjustment for clustering
     - Variance adjustment
     - Logistic mixed effects modeling
     - Generalized Estimating Equations (GEE)
     - Cluster bootstrap

Single (vs. separate) models

Advantages

- Including all available information in one model
  - Disease status as an independent variable (vs. subsetting by disease status)
- Modeling together
  - Sensitivity + specificity
  - PPV + NPV
- Can be performed with logistic regression, logistic mixed effects modeling, GEE, cluster bootstrap

Leisenring *et al.* (1997), Ronco & Biggeri (1999)

Disadvantages

- Assuming common ICC for diseased and healthy populations: not necessarily valid, $ICC_{diseased} > ICC_{healthy}$
- Harder to calculate and interpret than separate models

Genders et al. (2012)

Binomial model + Logistic regression

Simple models, all assume independence of observations

- Diagnostic test values estimated directly as proportions (π)
  - Exact binomial distribution discrete and skewed
    $\rightarrow$ Wald approximation / continuity correction
- Diagnostic test values estimated indirectly
  Binomial proportion bounded: [0, 1] (0 and 100%)
  - Problem with extreme estimates & CI's

  1. Converting to odds $\pi/(1-\pi)$: $[0, \infty)$
  2. Converting to log odds $\log(\pi/(1-\pi))$: $(-\infty, \infty)$

  $\rightarrow$ Logit transformation: CI's calculated for transformed estimates that can be converted back to probabilities
  $\rightarrow$ Logistic regression
  - e.g., Sensitivity: Log odds of testing positive (continuous outcome) regressed on presence of disease that can be converted back to probabilities

Variance adjustment

Widen variance → widen CI's to compensate for clustering

Different methods available

- Point estimates not affected: only appropriate if estimates unbiased
  - Computationally least intensive → Previously common
  - Ratio estimator (sens / spec only)
  - Variance inflation factor (sens / spec only) Genders *et al.* (2012)
- Point estimates may be affected
  - Logistic regression with sandwich estimator (all test values)
    - General, no regard for cluster type: only appropriate if clustering information cannot be recovered
    - Specific, correctly representing clustering

A note on point estimates

Naive (exact) point estimates likely unbiased if

- No missing values
  - Alternatively: data missing completely at random (MCAR), e.g., equipment malfunction Rubin (1976)
- Balanced data
  - Participants contribute a similar number of observations

Genders et al. (2012), Hujoel et al. (1990), Ying et al. (2020)

Model-adjusted point estimates

- Unbiased in correctly specified model (e.g., no confounding)

Comparison of mixed effects models vs. GEE

Logistic mixed effects models

- Estimates effect conditional on cluster / aspects of the effect that varies by cluster
- Variability can be of interest
- Coeff for sensitivity: (log) odds of testing + for a previously healthy, now diseased patient
- Can accommodate covariates
- Can handle smaller samples
- Coeff more extreme than marginal due to non-linearity
- Missing data can be MAR (more permissible)

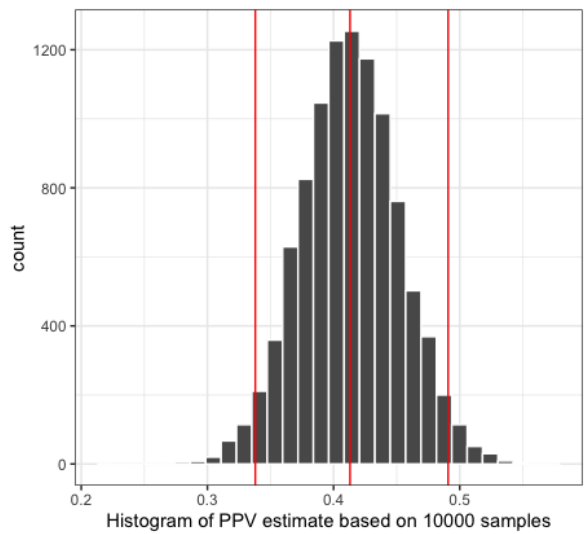
GEE

- Estimates the non-varying (average/marginal) effect in the presence of clustering
- Variability is nuisance factor
- Coeff for sensitivity: (log) odds of testing + in diseased vs. healthy populations
- Can accommodate covariates
- Requires larger samples
- Sandwich estimators: asymptotically valid CI's even if corr. structure misspecified
- Missing data has to be MCAR

e.g., Zeger & Liang (1992)

Cluster bootstrap

1. Resampling technique: sampling participants (units of cluster) a large number of times with replacement
2. Calculating sensitivity and specificity within each sample
3. Calculating predictive values using sensitivity and specificity within each sample
4. Coefficients and CI's determined by their respective distributions
5. Computationally intensive with large datasets



Ying *et al.* (2020)

When to use which method?

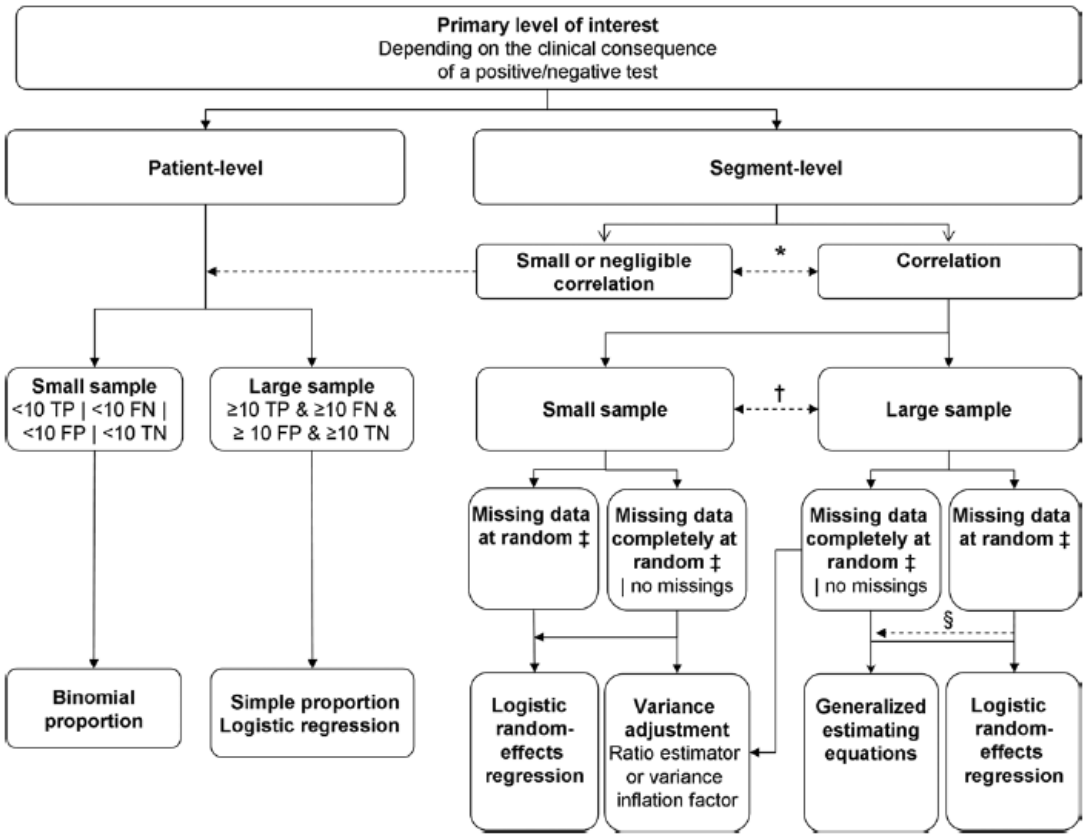


Figure 2: Chart shows guidance for choosing between methods for adjusting for correlation. This figure gives an indication but is not intended to be fully exhaustive. * = Magnitude of the correlation is often not known. Whether the correlation is strong enough to influence the results can often only be determined by applying one of the methods that adjusts for correlation and to compare it with the results without adjustment. † = There are no clear definitions for small and large samples with respect to considering random-effects logistic regression and GEEs. ‡ = Appendix E1 (online) has more details on missing data mechanisms. § = Use-weighted GEEs. *FN* = false-negative, *FP* = false-positive, *TN* = true-negative, *TP* = true-positive.

Genders et al. (2012)

Will soon be revised to include: predictive values, logit transform, continuity correction, cluster bootstrap, covariates, extreme values, balancedness, single models etc.

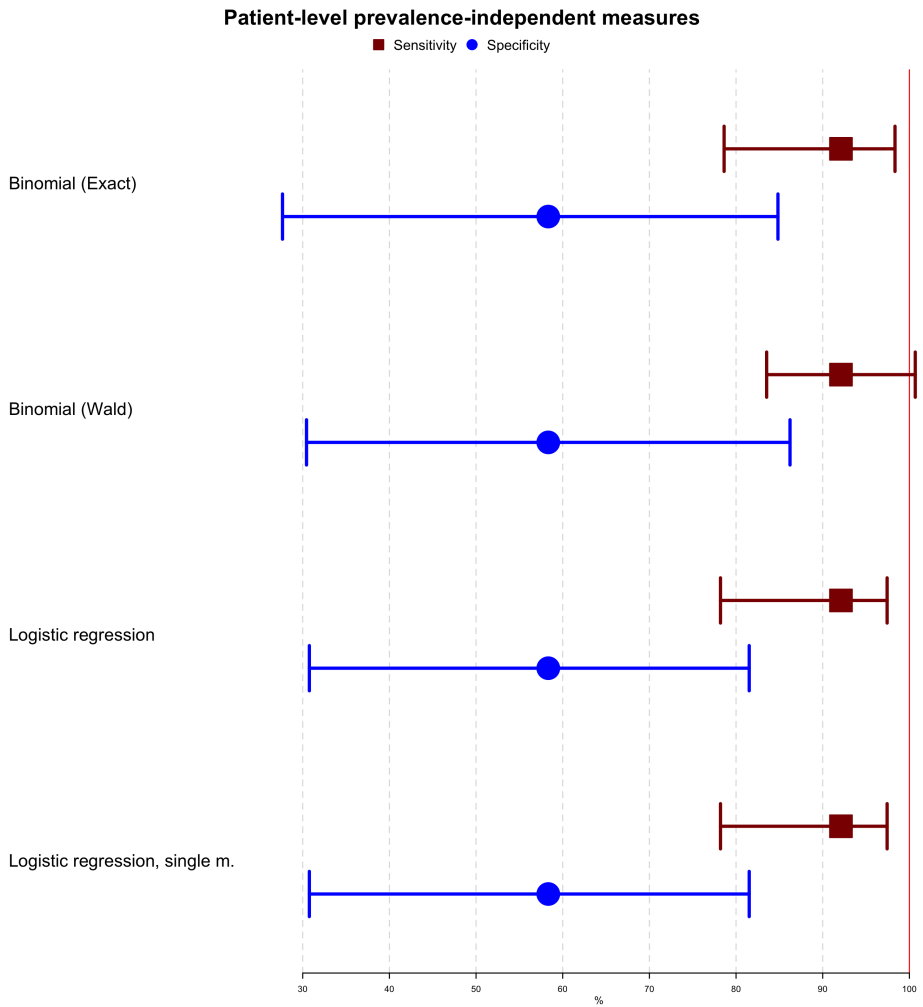
Results

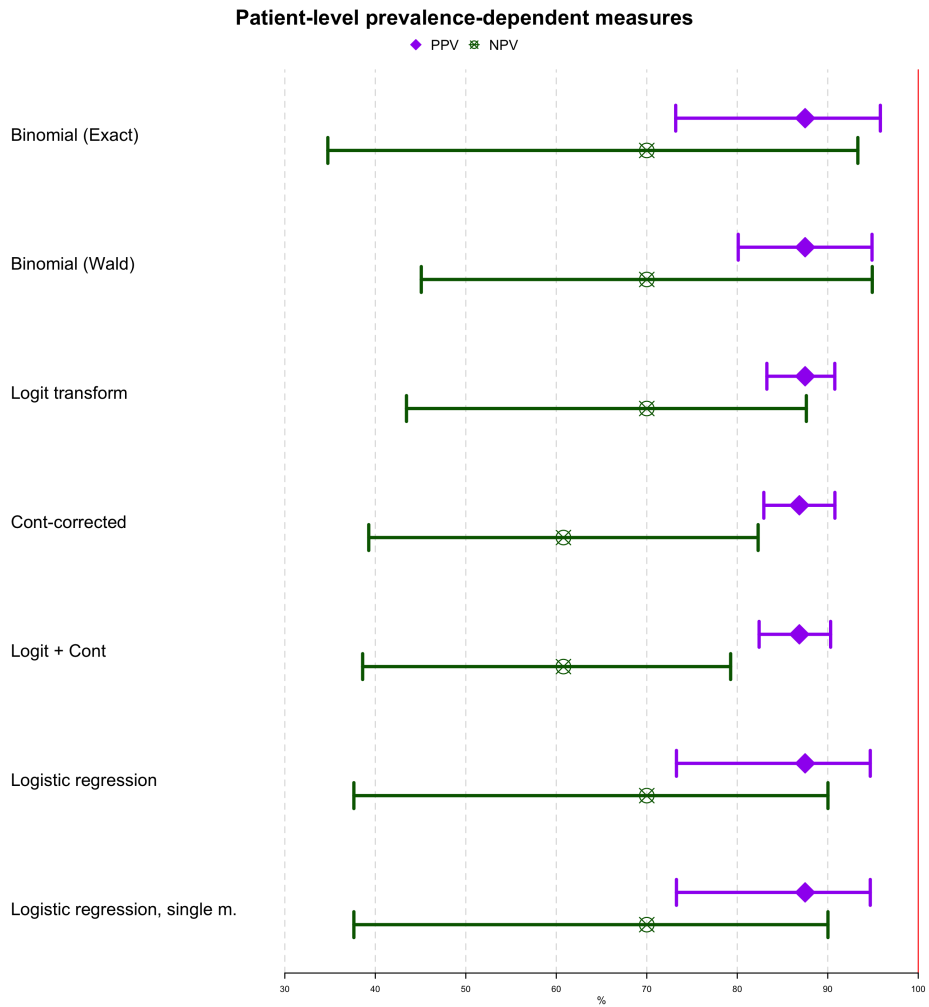
Developed R script that performs all necessary calculations wrt every method discussed

- Script will be open-access & freely adaptable
- Input: data in wide format
- Output:
  - Contingency tables and natural frequency trees at each level of analysis
  - Summary tables of each applicable diagnostic test value and CI's using every method
  - Forest plots of each diagnostic test value at each level of analysis

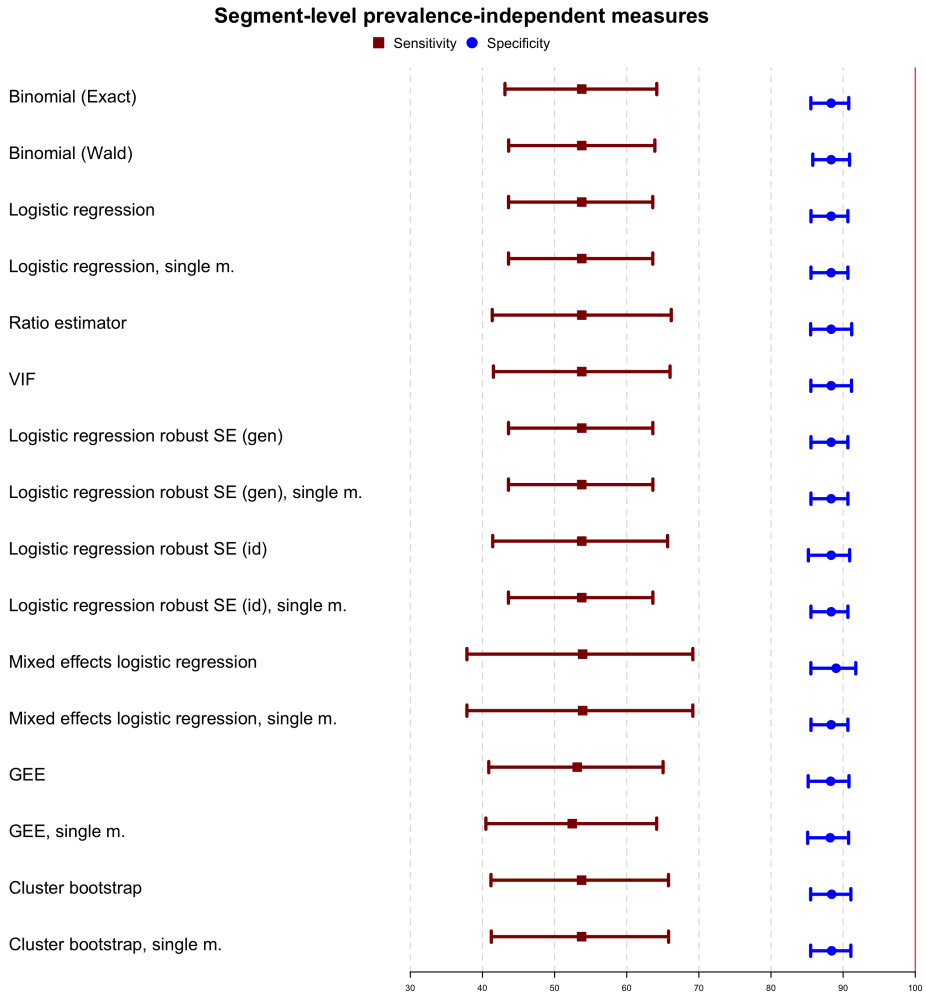
Patient-level measures

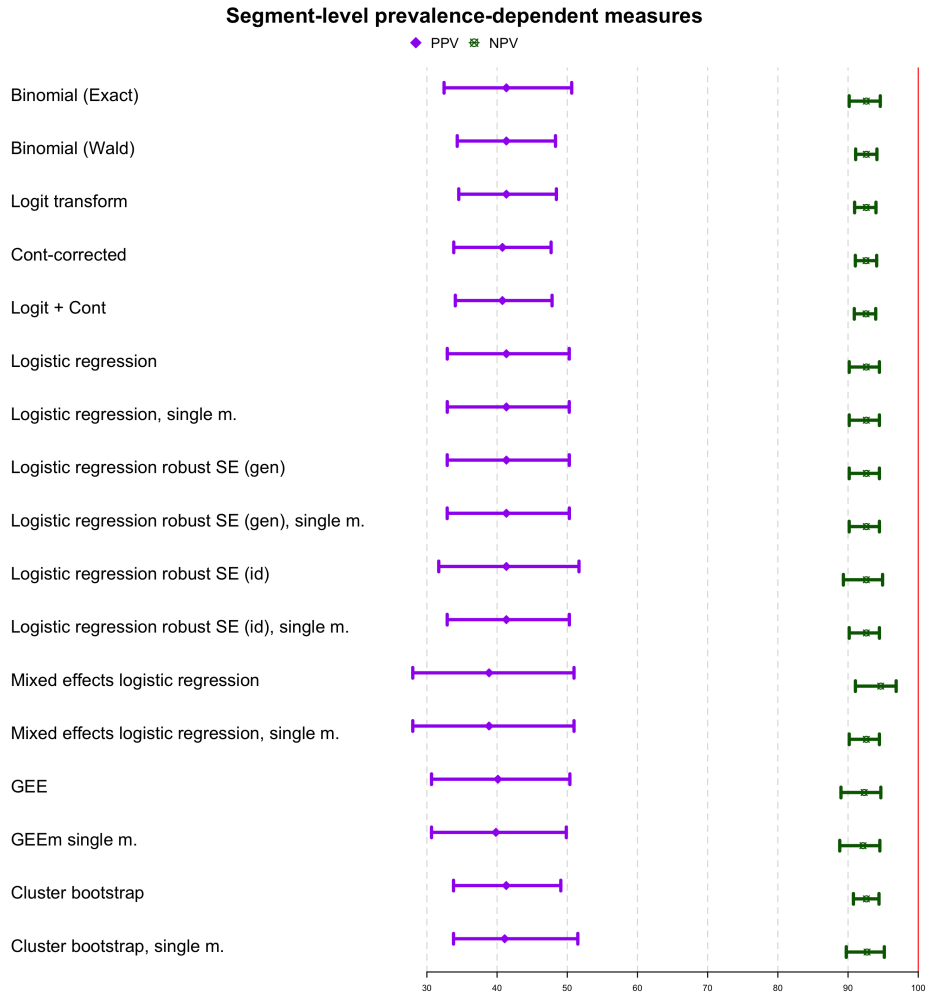
Using Genders *et al.* (2012)'s data set as input





Segment-level measures





Discussion

- "Multiverse" analysis e.g., Steegen et al. (2016)
  - Forest plots can be used to assess robustness of estimates & CI's
- $ICC_{diseased}$: 0.18 > $ICC_{healthy}$: 0.02 (general pattern)
  - Differential impact on CI's
- Both ICC's relatively small
  - Adjusted estimates and CI's similar to unadjusted
- The higher the ICC:
  - The more difference expected between unadjusted and adjusted values
  - The more closely patient- and segment-level estimates resemble each other

Possible sources of bias

- Verification bias: Application of reference (gold standard) test depends on index test outcome and/or other covariates
  - Can be corrected for if P(V|T) is known
- Selection bias: Number of tests per subject is affected by outcome of previous tests
  - Protocol should pre-specify number of tests
  - If not possible, ad-hoc solution: Stratify by number of tests

Ronco & Biggeri (1999)

- "Specificity bias": Artificially boosting (segment-level) specificity by “salami-slicing” / increasing the number of observations

Zwinderman *et al.* (2008)

  - Protocol should pre-specify number of observations per patient
- Etc.

Simulation study

Work-in-progress together with Prof. Gerton Lunter

Input parameters

- Number of patients
- Patient-level disease prevalence
- Number of observations per patient
- Segment-level disease prevalence
- Segment-level sensitivity & specificity

Derived parameters

- Segment-level PPV & NPV
- Patient-level sensitivity, specificity, PPV, & NPV

$\rightarrow$ True "population" diagnostic test values

Simulation study

1. Simulate a large number (10000) of patient populations given input parameters
2. Calculate diagnostic test values + 95% CI with each method
3. Rank methods based on
   1. Calibration: Do the 95% CIs contain the true parameters ∼95% of the time? = Nominal coverage probability
   2. Coverage probability: The closer to 95% the better
   3. CI length: The narrower the better
4. Declare the winner for each diagnostic test value at each level
5. Assess robustness of rankings as a function of input parameters
   - Clustering, missing data, balancedness, population/sample size, extreme values
6. Prepare recommendations based on results

Key takeaways

When reading a publication with diagnostic test values / Designing a screening/diagnostic study:

1. Determine if there is an issue with clustering
   - If so: What type? How does it affect analysis/interpretation?
2. Choose your (primary) level of analysis carefully
3. Identify what diagnostic test values can and cannot be estimated
   - Only contingency tables given?
   - Prevalence known to calculate predictive values?
4. Beware of how clustering impacts point estimates *and* CI's
   - Always report CI's and interpret estimates therewith
5. Look out for biases
6. Choose best possible type of adjustment for clustering
   - Stay tuned for specific recommendations
7. Visualize your (and other's) data and findings!

References

Binder, K., Krauss, S., Bruckmaier, G., & Marienhagen, J. (2018). Visualizing the Bayesian 2-test case: The effect of tree diagrams on medical decision making. *PLoS One, 13*(3), e0195029.

Genders, T. S., Spronk, S., Stijnen, T., Steyerberg, E. W., Lesaffre, E., & Hunink, M. M. (2012). Methods for calculating sensitivity and specificity of clustered data: A tutorial. *Radiology, 265*(3), 910-916.

Hujoel, P. P., Moulton, L. H., & Loesche, W. J. (1990). Estimation of sensitivity and specificity of site-specific diagnostic tests. *J of Periodontal Res, 25*(4), 193-196.

Leisenring, W., Pepe, M. S., & Longton, G. (1997). A marginal regression modelling framework for evaluating medical diagnostic tests. *Stat in Med, 16*(11), 1263-1281.

Ronco, G., & Biggeri, A. (1999). Estimating sensitivity and specificity when repeated tests are performed on the same subject. *J of Epi and Biostat, 4*(4), 329-336.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592.

Shen, Y., Wu, D., & Zelen, M. (2001). Testing the independence of two diagnostic tests. *Biometrics, 57*(4), 1009-1017.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psych Sci, 11*(5), 702-712.

Watkins S. J. (2000). Conviction by mathematical error? Doctors and lawyers should get probability theory right. *BMJ (Clinical research ed.), 320*(7226), 2–3. https://doi.org/10.1136/bmj.320.7226.2

Ying, G. S., Maguire, M. G., Glynn, R. J., & Rosner, B. (2020). Calculating sensitivity, specificity, and predictive values for correlated eye data. *Inv Ophth & Vis Sci, 61*(11), 29-29.

Zeger, S. L., & Liang, K. Y. (1992). An overview of methods for the analysis of longitudinal data. *Stat in Med, 11*(14-15), 1825-1839.

Zwinderman, A. H., Glas, A. S., Bossuyt, P. M., Florie, J., Bipat, S., & Stoker, J. (2008).