Metadta: A Stata Command for Meta-analysis and Meta-regression of Diagnostic Test Accuracy Data – A Tutorial.

Victoria Nyawira Nyaga (Victoria.NyawiraNyaga@sciensano.be)
Sciensano https://orcid.org/0000-0002-8381-9964
Marc Arbyn
Sciensano

Methodology

Keywords: Meta-analysis, meta-regression, diagnostic test accuracy, Stata, metadta

Posted Date: July 22nd, 2021

DOI: https://doi.org/10.21203/rs.3.rs-721114/v1

Version of Record: A version of this preprint was published at Archives of Public Health on March 29th, 2022. See the published version at https://doi.org/10.1186/s13690-021-00747-5.

metadta: A Stata command for meta-analysis and meta-regression of diagnostic test accuracy data – a tutorial.

Victoria Nyawira Nyaga[1*] and Marc Arbyn[1]

[1]Unit of Cancer Epidemiology - Belgian Cancer Centre, Sciensano, Juliette Wytsmanstraat 14, 1050 Brussels, Belgium

*Correspondence: victoria.nyawiranyaga@sciensano.be

Abstract

Background: Although statistical procedures for pooling of several epidemiological metrics are generally available in statistical packages, those for meta-analysis of diagnostic test accuracy studies including options for multivariate regression are lacking. Fitting regression models and the processing of the estimates often entails lengthy and tedious calculations. Therefore, packaging appropriate statistical procedures in a robust and user-friendly program is of great interest to the scientific community.

Methods*: metadta* is a statistical program for pooling of diagnostic accuracy test data in Stata. It implements both the bivariate random-effects and fixed-effects model, allows for meta-regression, and presents the results in tables, a forest plot and/or summary receiver operating characteristic (SROC) plot. For a model without covariates, it also quantifies heterogeneity using an I^2 statistic that accounts for the mean-variance relationship, and correlation between sensitivity and specificity, a typical characteristic of diagnostic data. To demonstrate *metadta,* we applied the program on two published meta-analyses on: 1) the sensitivity and specificity of cytology and other markers including telomerase for

primary diagnosis of bladder cancer; and 2) the accuracy of human papillomavirus testing on self-collected versus clinician-collected samples to detect cervical precancer.

Results: Without requiring a continuity correction, *metadta* generated a pooled sensitivity and specificity of 0.77 [95% CI: 0.70, 0.82] and 0.91 [95% CI: 0.75, 0.97] respectively of telomerase for the diagnosis of primary bladder cancer. *metadta* allowed to assess the relative accuracy of human Papilloma virus (HPV) testing on self- versus clinician-taken specimens in matched studies taking into account two covariates. Under the condition of using assays based on target-amplification, HPV tests were similarly sensitive to detect cervical pre-cancer, irrespective of clinical setting.

Conclusion: The *metadta* program implements state of art statistical procedures in an attempt to close the gap between methodological statisticians and systematic reviewers. With *metadta*, we hope to popularize even further, the use of appropriate statistical methods for diagnostic meta-analysis.

Keywords: Meta-analysis, meta-regression, diagnostic test accuracy, Stata, *metadta*.

Background

Meta-analysis of diagnostic test accuracy (DTA) data using approximate methods impose several challenges. These include poor statistical properties when sensitivity and/or specificity are close to the margins, when the sample sizes or when the number of studies are small. Moreover, the sample variance of sensitivity/specificity is a function of the sample mean and ignoring this mean-variance relationship may bias the summary estimate and its variance. Generalized linear mixed models (GLMM) (1) are therefore recommended (2). These models are relatively complex. Scientists in the field public health, epidemiology or clinical research often do not have access to advanced statistical and/or programming skills. Expertise in both in GLMMs and statistical programming is required. Hence, availability and dissemination of appropriate and optimal statistical methods in a robust and user-friendly program is quintessential.

The two most commonly used statistical models for pooling of diagnostic test accuracy data include the hierarchical summary receiver operation characteristic model (HSROC)(3) and the bivariate random-effects meta-analysis model (BRMA) (2). The two models incorporate covariates differently though they have been shown to be equivalent when no covariate are included.

The proportion of total variation caused by inter-study heterogeneity is usually quantified using the I^2 statistic (4), a measure based on the normal-normal model. This I^2 statistic represent the proportion of variation due to inter-study. It was defined for univariate meta-analysis and therefore with diagnostic data set, a separate statistic for sensitivity and specificity is computed. The fact that diagnostic data is binomial implies that the within-study variance in sensitivity and specificity parameters is a function of the mean

parameters. Hence, heterogeneity statistics based on the normal-normal model tend to underestimate the expected value of the within-study variance resulting in high values of I^2. This could lead to an incorrect conclusion of very high heterogeneity (4).

The reasons for the substantial heterogeneity in the null mixed-effects model can be explored by relating study level co-variates to the latent sensitivity and specificity. This is called meta-regression.

Stata is a popular software for meta-analysis among epidemiologists, clinicians, researchers and other stakeholders. Stata's *metand*i command (5) fits both the HSROC and the BRMA model, displays the summary accuracy measures, and plots the SROC curve, its prediction region, the summary point and its confidence region. However, the command does not allow meta-regression. *midas*(5) is another Stata command implementing the bivariate random-effects model. It is a more comprehensive package with more graphical output to explore goodness of fit, publication bias and other precision-related biases. Nonetheless, the command only allows for univariate meta-regression and quantification of heterogeneity using the I^2 statistic based on the normal-normal model.

In this paper, we present the new Stata command *metadta* which implements the bivariate random-effects and the univariate the fixed-effects model as a special case of the bivariate model. In addition, the command allows for univariate and bivariate meta-regression. *metadta* presents the results in tables, a forest and/or SROC plot. With paired data (two tests performed on the same subjects/study), a forest plot of relative sensitivity and specificity can be displayed. For the model without covariates, it also quantifies

heterogeneity using I^2 that accounts for the mean-variance relationship, and correlation between sensitivity and specificity, an intrinsic characteristic of diagnostic data.

Methods

Diagnostic test accuracy data

Data from DTA studies usually result from a 2 x 2 cross-tabulation of index versus reference test results (see Table 1). The data in the four cells represent the true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The sum of TP and FN is the total with disease, and the sum of TN and FP the total without disease.

		Disease Status	
		+	-
Index Test	+	True positive (TP) = Y_{i1}	False Positive (FP)
	-	False Negative (FN)	True Negative (TN) = Y_{i2}
	Total	Diseased = N_{i1}	Non-diseased = N_{i2}

Table 1. Cross-tabulation of index test results by the disease status in study *i*.

In a comparative study, there will be a pair of 2 x 2 cross-tables, one for the index test and the second for the comparator test, and type of the test will be treated as a covariate.

The logistic regression model

Consider a meta-analysis of K studies. For each study *i* (i = 1, …, K), let Y_{i1} be the number of true positive, Y_{i2} be the number of true negatives, N_{i1} the total number of

subjects with the disease, and N_{i2} the total number of subjects without the disease. Then, the fixed-effects model is formulated as follows;

$Y_{ij} \sim binomial\left(p_{ij}, n_{ij}\right)$ for $i = 1, \ldots, K$ and $j = 1,2,$

$$p_{ij} = \frac{\exp(\beta_0^j + \beta_1^j X_{ij}^1 \ldots \beta_P^j X_{ij}^P)}{1 + \exp(\beta_0^j + \beta_1^j X_{ij}^1 \ldots \beta_P^j X_{ij}^P)},$$

where p_{i1} and p_{i2} are parameters denoting sensitivity and specificity in study *i* respectively. β_0^j are log-odds while $\beta_1^j \ldots \beta_P^j$ are log odds ratios. $X_{ij}^p$is the value of the *p'th* predictor variable in study *i* for logit sensitivity(j=1) and logit specificity (j=2).

The random-effects model has (un)correlated random components in the mean predictor as follows;

$$p_{ij} = \pi(x) = \frac{\exp(\beta_0^j + \beta_1^j X_{ij}^1 \ldots \beta_P^j X_{ij}^P + \delta_{ij})}{1 + \exp(\beta_0^j + \beta_1^j X_{ij}^1 \ldots \beta_P^j X_{ij}^P + \delta_{ij})},$$

$$\begin{pmatrix}\delta_{i1}\\ \delta_{i2}\end{pmatrix} \sim N\left(\begin{pmatrix}0\\ 0\end{pmatrix}, \boldsymbol{\Sigma}\right),$$

where$\delta_{ij}$are the study-specific random-effects for the logit sensitivity(j=1) and logit specificity (j=2). The variability in the two random-effects and their correlation is represented by $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ can be any of the four variance-covariance structures: unstructured $\begin{pmatrix}\tau_1^2 & \tau_{12}\\ \tau_{12} & \tau_2^2\end{pmatrix}$, independent $\begin{pmatrix}\tau_1^2 & 0\\ 0 & \tau_2^2\end{pmatrix}$, exchangeable $\begin{pmatrix}\tau^2 & \tau_{12}\\ \tau_{12} & \tau^2\end{pmatrix}$ or identity $\begin{pmatrix}\tau^2 & 0\\ 0 & \tau^2\end{pmatrix}$. The BRMA imposes the unstructured variance-covariance structure. The most complex and relaxed model with the unstructured variance covariance has at least five parameters in the model to be identified. Hence, at least five studies would be required to enable

parameter identification. Other structures are more restrictive and require less number of studies (at least 3) for identifiability.

When there are two observations per study the linear predictor for p_{ij} is modified to account for dependence (6). The fixed-effects component in the linear predictor can be extended to include interaction terms between the first covariate, say T and the remaining covariates as follows;

$$\beta_0^j + \alpha_j T_{it} + \beta_1^j X_{ij}^1 \ldots \beta_P^j X_{ij}^P + \vartheta_1^j T_{it} X_{ij}^1 \ldots \vartheta_P^j T_{it} X_{ij}^P.$$

β_0^j are the baseline log-odds, α_j the log-odds ratio for test and β_p^j the log-odds ratio for the confounding factor p. ϑ_p^j is a log ratio of odds ratio for test in the two levels of factor p. A log-likelihood ratio test is then conducted comparing the model with and without the interaction terms to give an answer as to whether differences exists in log-odds ratios given the confounding factors. The fixed-effect model is fitted using the Stata command *binreg* while the random-effects model uses the command *meqrlogit*.

Summary estimates

By using the Stata command *margins*, we obtained an average of the predicted probabilities from the model. Also called marginal probabilities, they are said to be standardized to the distribution of the covariate variables (7). Model-adjusted probability ratios(5) are then computed as a ratio of the marginal probabilities. The margins and their non-linear combinations were obtained either in the logit or the log scale and then transformed back to the probability ratio scale (relative sensitivity and relative specificity). Stata uses the delta method to compute the variances of the computed statistics.

For the model without covariates, the I^2 statistic developed by Zhou & Dendukuri (4) is presented. The statistic accounts for the mean-variance relationship across studies and the correlation between sensitivity and specificity yielding a joint measure of heterogeneity. The univariate I^2 for logit sensitivity and specificity are presented as well. It is worth noting that the I^2 statistic will almost always be much lower than the heterogeneity statistic based on the approaches of Higgins and Thompson (8), and Jackson et al.(9)

Forest and SROC plots

The Stata command *ci proportions* command was used to compute the study-specific confidence intervals for proportions (absolute sensitivity and specificity). The command allows for six different confidence intervals for proportions; the Wald, Wilson, Agresti-Coull, Jeffreys, and exact confidence intervals. The exact confidence intervals are the default.

For comparative and paired studies, the Koopman confidence intervals (10) intervals for the relative sensitivity and specificity are presented. Such data is typically presented as in Table 2 and Table 3. The two 2 x 2 tables are displayed below;

		Disease Status	
		+	-
Index Test	+	True positive (TP1)	False Positive (FP1)
	-	False Negative (FN1)	True Negative (TN1)

Table 2: Cross-tabulation of index test results by the disease status in study *i*.

		Disease Status	
		+	-
Comparator Test	+	True positive (TP2)	False Positive (FP2)
	-	False Negative (FN2)	True Negative (TN2)

Table 3: Cross-tabulation of comparator test results by the disease status in study *i*.

In comparative studies, that the index and the comparator test should be similar in all studies. In Figure 1, the origin and the arrow head are the comparator and index test respectively. In comparative studies, that the index and the comparator test should be similar in all studies. This implies that we only consider 1 comparison in Figure 1. In paired studies, the index and comparators tests can differ among the studies and hence the all the comparisons in Figure 1 are considered.

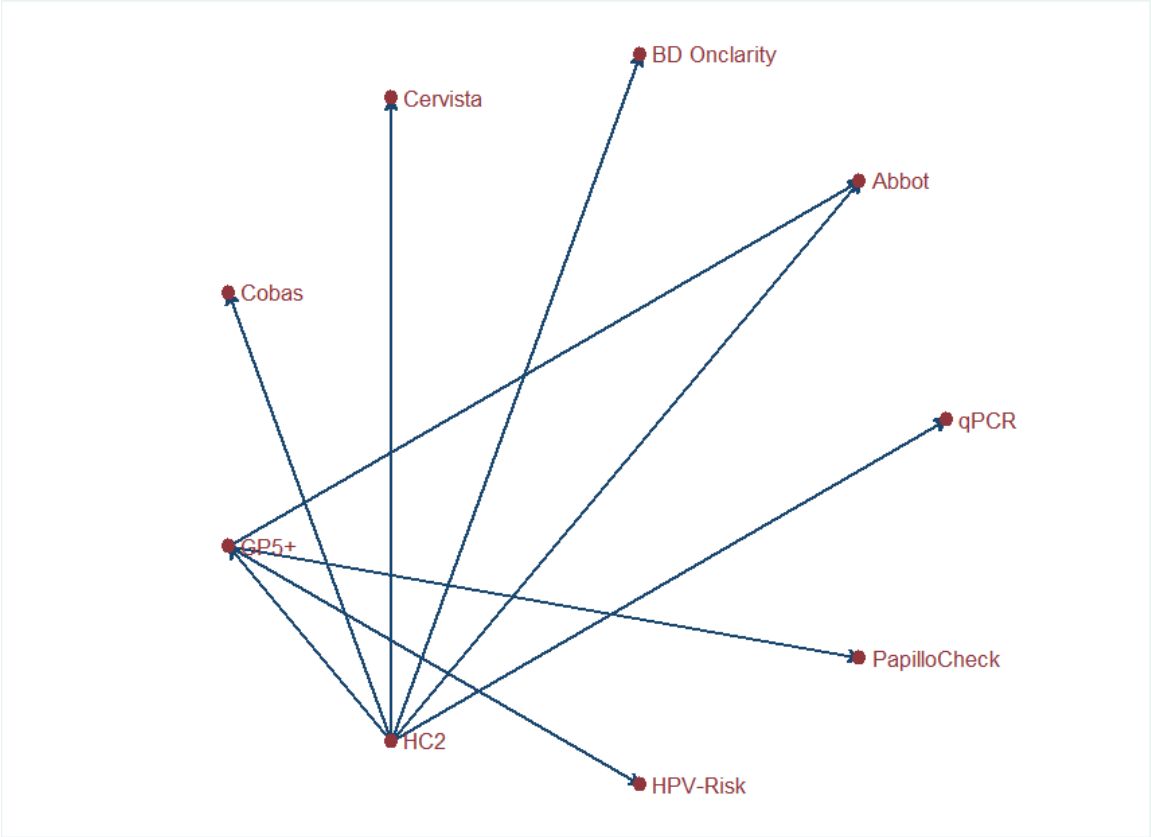


Figure 1: A network plot in a hypothetical meta-analysis including studies that compare either HC2 or GP5+ with other high risk HPV tests for cervical cancer screening.

When the number of studies is not sufficient to fit the random-effects model or when the fixed-effects model is explicitly used, cross-hairs indicating the confidence intervals of the summary estimates are presented in the SROC plot. Otherwise, the confidence and prediction regions are presented. In presence of more than one covariates, the SROC plot presents the overall summary estimates and the corresponding confidence and/or prediction region. When the focus is on the probability ratio, the SROC plot is not presented.

Software installation

The *metadta* command was developed in Stata 14.2. The program along with the help files and three demonstration datasets are publicly available for downloading at https://ideas.repec.org/c/boc/bocode/s458794.html. When connected to the internet, the command can be directly installed within Stata by typing **`ssc install metadta.`**

Syntax

```
by varlist: metadta tp fp fn tn tp2 fp2 fn2 tn2 covariates,
      studyid(varname)
      label(namevar=varname, yearvar=varname) by(byvar)
      dp(integer 2) power(integer 0) model(fixed|random)
      sortby(varlist) alphasort cveffect(se|sp|sesp)
      interaction(se|sp|sp) level(integer 95) paired comparative
      cimethod(string) cov(string) sumtable(string) nomc progress
      nofplot nosroc noitable nooverall nosubgroup summaryonly
      foptions(astext(integer 50) arrowopt(options)
            ciopt(options) diamopt(options) double lcols(varlist)
            noovline nostats olineopt(options) outplot(abs|rr)
            plotstat(string) subline texts(real 1)
            xlabel(string) xtick(string) grid
            graphsave(filelocation) logscale *)
      soptions(colorpalette(string) noprediction bubbles bubbleid
                  spointopt(options) opointopt(options)
                  curveopt(options) ciopt(options)
```

predciopt(options) bubopt(options)

*bidopt(options) graphsave(filelocation) *)*

The five parameters in bold are required in the *metadta* command. The other parameters in italics are optional. The required parameters are; the variable identifying the studies **studyid**, and four dependent variables **tp fp fn tn** from the 2 x 2 cross-tabulation Table 1.

Once installed, typing **help metadta** should display the help window. The help file provides a detailed description of all the command parameters. Some of the parameters worth mentioning here include;

bysort varlist: allows for separate meta-analyses for each level of by variable or each combination of the by variables.

covariates indicates one or more variables to be used as covariates for meta-regression.

comparative indicates whether the data supplied is from comparative studies. This option requires the first covariate supplied to be categorical with two levels for the index and the comparator test.

paired indicates whether the data supplied is from two tests applied on the same study. This option requires at least 10 variables in the following order *tp1 fp1 fn1 tn1 tp2 fp2 fn2 tn2 index comparator*.

* are graphical options native in Stata to change the aesthetics of the forest plot and/or sROC plot.

Application

Example one – Random-effects model with no covariates

Glas et al. (11) assessed in a meta-analysis the sensitivity and specificity of cytology and other markers including telomerase to diagnose bladder cancer. They fitted a bivariate normal distribution to the logit transformed sensitivity and specificity across the studies allowing for heterogeneity between the studies. Because the seventh study had an estimated specificity equal to one, the authors used a continuity correction of 0.5 to enable parameter estimation. From the included 10 studies, they reported that telomerase had a sensitivity and specificity of 0.75 [95% CI: 0.66, 0.74] and 0.86 [95% CI: 0.71, 0.94] respectively. They concluded that telomerase was not sensitive enough to be recommended for daily use. This dataset is provided along with the installation files.

```
. use "http://fmwww.bc.edu/repec/bocode/t/telomerase.dta", clear

. list, noobs clean
            study   tp   fp   fn    tn
          Ito1998   25    1    8    25
        Rahat1999   17    3    4    11
      Kavaler1998   88   16   16    31
      Yoshida1997   16    3   10    80
    Ramakumar1999   40    1   17   137
      Landman1998   38    6    9    24
    Kinoshita1997   23    0   19    12
      Gelmini2000   27    2    6    18
        Cheng2000   14    3    3    29
       Cassel2001   37   22    7     7
```

`study` is the study identifier, `tp, fp, fn,` and `tn` are the variable names for the true positives, false positives, false negatives and true negatives respectively.

```
.  metadta tp fp fn tn,                                          ///
  studyid(study) model(random) dp(2) sumtable(all)               ///
  soptions(xtitle("False positive rate")                         ///
     ciopt(lpattern(dash_dot))                                   ///
     xlabel(0(0.2)1) xscale(range(0 1))                          ///
     ytitle("Sensitivity") yscale(range(0 1))                    ///
     ylabel(0(0.2)1, nogrid)                                     ///
     graphregion(color(white)) plotregion(margin(medium))        ///
          xsize(15) ysize(15)                                    ///
          legend(order(1 "Summary" 5 "Observed data"             ///
          2 "SROC" 3                                             ///
          "Confidence region" 4 "Prediction region")             ///
           cols(1) ring(0) bplacement(6)))                       ///
    foptions(graphregion(color(white)) texts(2.5)                ///
          xlabel(0, 0.5, 1) diamopt(color(red))                  ///
          olineopt(color(red) lpattern(dash)))
```

Apart from the 4 required main arguments (`tp fp fn tn`), we specify the following options: `model(random)` to request for fitting of a random-effects model, this specification is actually not required since *metadta* fits a random-effects by default unless when the number of studies is less than 3. `dp(2)` requests the results of all estimates to be displayed with 2 decimal places (except the p-values for which the decimals places are fixed at 4). `sumtable(all)` requests for the `all` available summary tables, i.e summary

estimates on the logit and probability scale. With `soptions()` and `foptions()` we could refine the display of the forest and SROC plots.

The first part of the output displays the fitted model and the number of observations and studies in the meta-analysis as shown below;

```
************** Fitted model*********************************
    tp ~ binomial(logit(se), tp + fn)
    tn ~ binomial(logit(sp), tn + fp)
    logit(se) = mu_se + study_se
    logit(sp) = mu_sp + study_sp
    logit(study_se), logit(study_sp) ~ biv.normal(0, sigma)
    Number of observations = 10
    Number of studies = 10
```

We requested all output tables with `sumtable(all)`, where all summary statistics are displayed on the logodds scale (logodds) and on the probability (abs) scale. They are presented as follows;

```
*********************************************************************
            Conditional summary measures of test accuracy : Log_odds
*********************************************************************
---------------------------------------------------------------------
  Parameter | Log odds    SE        z      P>|z|    Lower     Upper
------------+--------------------------------------------------------
Sensitivity | 1.19       0.18      6.61     0.0000   0.84      1.55
--------------------+------------------------------------------------
Specificity | 2.34       0.63      3.69     0.0002    1.10      3.58
---------------------------------------------------------------------
```

```
*****************************************************************
            Conditional summary measures of test accuracy : Proportion
*****************************************************************
-----------------------------------------------------------------
  Parameter | Proportion  SE(logit) z(logit)  P>|z|   Lower     Upper
------------+----------------------------------------------------
Sensitivity | 0.77        0.18      6.61      0.0000  0.70      0.82
------------+----------------------------------------------------
Specificity | 0.91        0.63      3.69      0.0002  0.75      0.97
-----------------------------------------------------------------
NOTE: H0: p = 0.5 vs. H1: P != 0.5
```

The first table presents the summary statistics on the logit scale. The mean logit sensitivity and specificity were 1.19 [95% CI: 084, 1.55] and 2.34 [95 % CI: 1.10, 3.58]. The p-values after testing whether the logit sensitivity or logit specificity is 0 are both < 0.01. Thus the logits are significantly different from zero.

The second table presents the same summary statistics but on the probability scale. The means and the confidence intervals are transformed back but the standard errors, the z-statistic and the p-values are reported in the logit scale. Since the transformation from the logit to the probability scale is non-linear, the table presents the median sensitivity and median specificity. Translated into the probability scale, the p-values are for testing whether the median sensitivity and specificity are both 0.5. If needed, one can use the delta method to compute the standard errors in the probability scale.

Results above show that telomerase in urine as a tumour marker for the diagnosis of primary bladder cancer has a pooled sensitivity and specificity of 0.77 [95% CI: 0.70, 0.82] and 0.91 [95% CI: 0.75, 0.97] respectively. Our results are different from the

original publication in that the authors used the linear mixed model which required the use of a 0.5 continuity correction in the seventh study, whereas in *metadta* no continuity correction was needed.

The third table is generated by default and presents the study-specific sensitivity and specificity and their corresponding 95% exact confidence intervals. After the last study, the overall sensitivity and specificity as estimated by the model is presented as defined by the `sumtable()` option.

```
*********************************************************************
            Study specific test accuracy sensitivity and specificity
*********************************************************************

             |  Sensitivity                    |    Specificity
 Study       | Estimate  [95% Conf. Interval] | Estimate [95% Conf. Interval]
-------------+---------------------------------+-----------------------
Ito1998      | 0.76        0.58      0.89     | 0.96       0.80      1.00
Rahat1999    | 0.81        0.58      0.95     | 0.79       0.49      0.95
Kavaler1998  | 0.85        0.76      0.91     | 0.66       0.51      0.79
Yoshida1997  | 0.62        0.41      0.80     | 0.96       0.90      0.99
Ramakumar1999| 0.70        0.57      0.82     | 0.99       0.96      1.00
Landman1998  | 0.81        0.67      0.91     | 0.80       0.61      0.92
Kinoshita1997| 0.55        0.39      0.70     | 1.00       0.74      1.00
Gelmini2000  | 0.82        0.65      0.93     | 0.90       0.68      0.99
Cheng2000    | 0.82        0.57      0.96     | 0.91       0.75      0.98
Cassel2001   | 0.84        0.70      0.93     | 0.24       0.10      0.44
Overall      | 0.77        0.70      0.82     | 0.91       0.75      0.97
-------------+---------------------------------+-----------------------------
*********************************************************************************
```

The last part of the output presents the heterogeneity statistics. By default, the unstructured covariance structures is imposed. The correlation between sensitivity and specificity in the logit scale is -1 (`rho`). `I^2 > 0` indicates higher between-study heterogeneity across the studies compared to the expected within-study variability.

There is more heterogeneity in specificity ($\sigma^2 = 3.32, I^2 = 60.29\%$) than in sensitivity $\sigma^2 = 0.18, I^2 = 50.62\%$). The generalized `Tau.sq = 0` summarizes the variance in both logit sensitivity and specificity while accounting for the correlation between them. Despite presence of heterogeneity in both dimensions, it may look surprising that the

bivariate `I^2 = 0.02`. This is because the generalized between-study variance goes to zero with (nearly) perfect correlation (indicated by `rho = -1.00`), and the higher the correlation, the lower the bivariate `I^2`.

```
Between-study heterogeneity

                           rho
                           -1.00
                           Tau.sq           I^2(%)
Generalized                 0.00             0.02
Sensitivity                 0.18            50.62
Specificity                 3.32            60.29


                            Chi2        degrees of
                           statistic      freedom        p-val
LR Test: RE vs FE model     104.33           3            0.0000
```

When a random-effects is fitted, the results of the likelihood ratio (`LR Test`) test compares the random- to the fixed-effects model. The test has 3 degree of freedom since the unstructured covariance matrix imposed by default has 3 parameters. The p-value is < 0.0001 indicating that the random-effects is a better fit than the fixed effects model.

Figure 2 and Figure 3 (left) presents the forest and the SROC plots respectively. The default plots are already visually appealing but the aesthetics can be optimized by options `soptions()` and `foptions()` for the SROC plot and the forest plot, respectively. The forest plot displays graphically the content of the table with the study-specific estimates. It is worth noting that the order of the data in the forest plot, the table with the study-specific and the dataset provided is preserved. Say it is preferred to order the studies by year of publication, a variable with the year of publication (say `year`) should be included

in the data, by using the `sortby(year)` option to order the whole dataset. It is worth noting that the SROC curve is restricted to the range of observed specificities in the dataset.

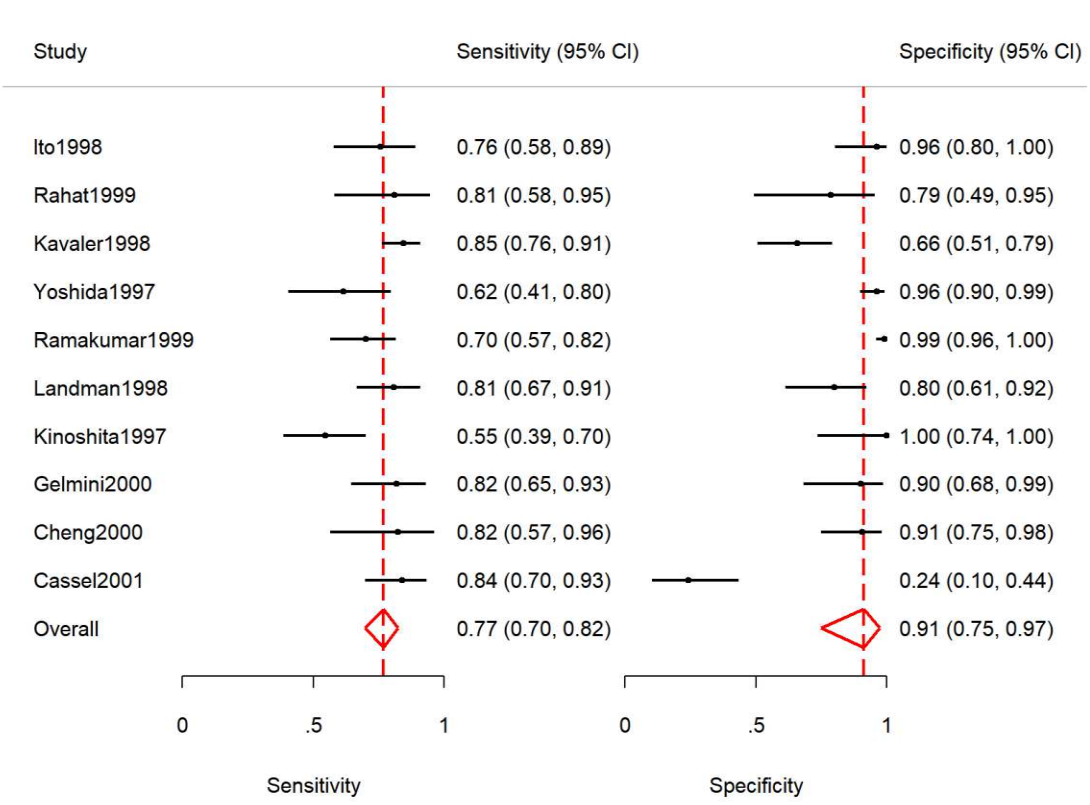


Figure 2. Forest plot - meta-analysis of diagnostic accuracy of telomerase for the diagnosis of bladder cancer.

Model comparison

Imposing a different covariance structure (e.g. independence) and picking up the most parsimonious model proceeds as following;

1. Restore the model estimates by typing `estimates restore metadta_modest`. (The estimates of the fitted model are always stored as `metadta_modest`). Once

restored, use the command `estat ic` to display the Akaike (AIC) and Bayesian information criteria (BIC) (12). The output is;

```
-----------------------------------------------------------------------------
       Model |  Obs  ll(null)  ll(model)  df   AIC       BIC
-------------+---------------------------------------------------------------
metadta_mo~t |   20   .       -50.38657    5   110.7731 115.7518
-----------------------------------------------------------------------------
       Note: N=Obs used in calculating BIC; see [R] BIC note.
```

2. Using the command `estimates store` store the estimates with a different name, say unstructured. i.e `estimates store unstructured.`
3. Fit a the model again with the `cov(independent)` option imposing the independence between logit sensitivity and logit specificity.
4. Repeat step 1 above to be able to display the information criteria once more;

```
-----------------------------------------------------------------------------
       Model | Obs  ll(null)  ll(model) df    AIC       BIC
-------------+---------------------------------------------------------------
metadta_mo~t | 20         .  -54.66482  4    117.3296   121.3126
-----------------------------------------------------------------------------
        Note: N=Obs used in calculating BIC; see [R] BIC note.
```

The models compared using information criteria need not be nested but need to be fitted to the same data. The model with a smaller criterion fits the data better. From the output in steps 1 and 4, the model with the correlation parameter (see step 1) fits the data better since both the Akaike information criterion (AIC) and Bayesian information criterion (BIC) (12) are lower.

In this example, the AIC and BIC give the same conclusion although sometimes they can lead to conflicting conclusions. The difference between AIC and BIC is in measuring the model complexity. Model complexity is measured either as 2*q or ln (K)*q, where q is the number of parameters estimated and K is the number of observations. Explicitly,

AIC = -2 x ln(likelihood) + 2 x q

BIC = - 2 x ln(likelihood) + ln(K) x q

In short, BIC is seen as more conservative than AIC by overweighting the model complexity. There is a debate on the value for K in computing the BIC, should it be the total number of observations or the number of independent groups in the data?

Implication of independence

The parameter estimates for the summary statistics in the logit scale under the independence assumption (1.18 (s.e = 0.18)) and 2.22 (s.e = 0.58)) are very similar to those under unstructured assumptions. The estimates for the between-study variability in logit sensitivity (`Tau.sq = 0.15, I2=46.81%`) and logit specificity (`Tau.sq = 2.75, I2=65.01%`) are also very similar. However, the heterogeneity statistics are much more different, logit sensitivity (I^2`=46.81%`) and logit specificity (I^2`=65.01%`). The estimate for the generalized variability is also much higher (`Tau.sq = 0.43`).

This confirms that the location parameters always tend to be consistent independent of the assumed variance-covariance structure. However, the confidence and prediction region under the two assumptions are different. Ignoring correlation yields wider regions as displayed in Figure 3 (right).

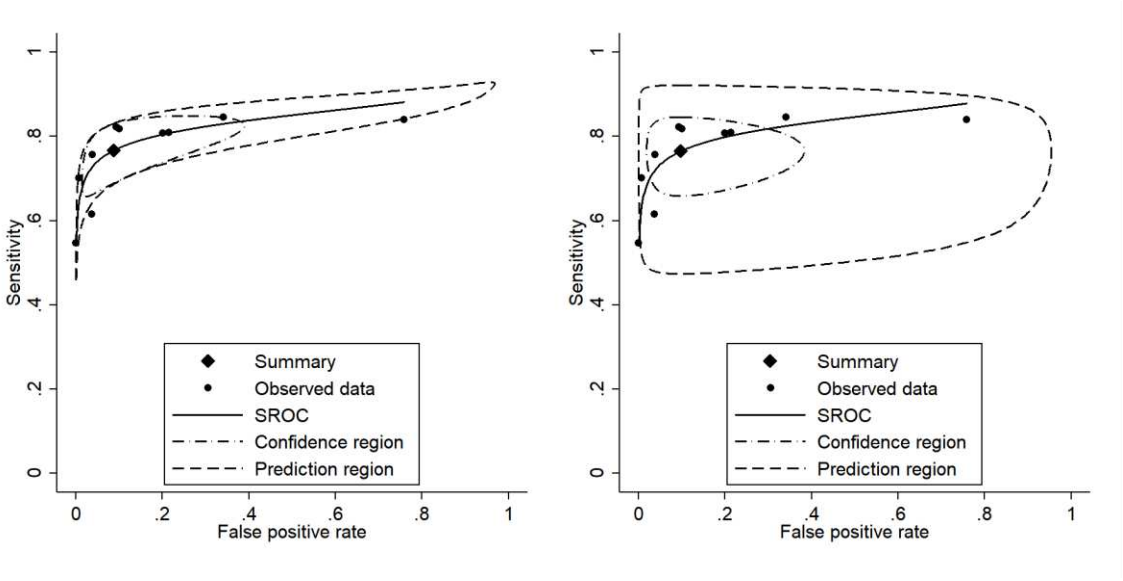


Figure 3. SROC plots - Left: unstructured covariance and right: independent covariance.

Example two – Random-effects meta-regression

Arbyn et al.(13) published a meta-analysis on the accuracy of human papillomavirus testing on self-collected versus clinician-collected samples. In the review, they sought to find whether an HPV test on a vaginal self-sample was as good as the test on a cervical sample taken by a clinician to detect cervical precancer (cervical intraepithelial neoplasia of grade 2 or worse [CIN2+]). *Metandi*(5) was used to report the pooled absolute sensitivity and specificity and *metadas*(14) was used to obtain the relative sensitivity and specificity (self-samples vs clinician samples) with type of the test (signal or amplification based) as a covariate.

The studies included in the meta-analysis came from three clinical settings: 1) cervical cancer screening, 2) testing of high-risk women, and 3) colposcopy where women were referred to because of previous positive screening results. Information on the study participants, tests, sampling device were recorded also.

For demonstration purpose, we will only use a sample of the published dataset. The first 10 observations of the dataset of the sample dataset are as below;

```
. use "https://github.com/VNyaga/Metadta/blob/master/clinselfdemo.dta?raw=true", clear

. list in 1/10, noobs clean
sample year  study           tp   fn    fp     tn   setting    ta
clin   2003  Salmeron, 2003  94    7   539   6694   screening  SA
clin   2000  Wright, 2000    47    9   228   1081   screening  SA
self   2007  Szarewski, 2007 17    4   160    739   screening  SA
self   2000  Wright, 2000    37   19   243   1066   screening  SA
clin   2006  Holanda, 2006    8    1   243    626   screening  SA
clin   2006  Girianelli, 2006 32   3   171   1569   screening  SA
clin   2007  Szarewski, 2007  21   0   139    760   screening  SA
clin   2012  Belinson, 2012  216   17  806   7517   screening  TA
clin   2012  Belinson, 2012  217   16  734   7589   screening  SA
self   2013  Nieves, 2013     24   17   179   1829  screening  SA
```

`study` is the study identifier, `tp, fp, fn,` and `tn` are the variable names for the true positives, false positives, false negatives and true negatives respectively. `sample` is a categorical variable with two classes, `clin` and `self` for clinician-sample and self-sample respectively. `year` indicates the year of study publication. `setting` is categorical variable with two classes, `screening` and `follow-up` indicating the clinical setup of the study. `ta` is a categorical with two levels; `TA` and `SA` indicating the type of test, signal-amplification or test-amplification respectively. *metadta* requires that categorical variables be string variables otherwise they are treated as continuous variables. The user can use the Stata commands *decode* to make factor variable strings. Also, the covariate names should not contain the underscore(_) characters, otherwise the program quits since this character is reserved for the program. If some of the variables

contain the underscore character, the user can use the Stata command *rename* to prepare the data.

Exploratory analysis

We first fitted four models for each category in setting and amplification method with sample as a covariate. We then inspected the tabular and graphical output to see whether the absolute sensitivity and specificity between self- and clinician sample differ by the amplification method and setting. The first model was as follows;

```
. gsort sample
. metadta tp fp fn tn sample if ta=="SA" & setting=="follow-up"   ///
      studyid(study) sortby(year study) nooverall                 ///
      noitable sumtable(abs rr) nosroc                            ///
      foptions(pointopt(msize(1.5)) diamopt(color(red))            ///
            olineopt(color(red) lpattern(dash)) outplot(abs)       ///
            graphregion(color(white)) texts(2.5)                    ///
            tit("SA, follow-up, by sample, CIN2+")                 ///
            graphsave("f1.gph") xlabel(0, .2, .4, .6, .8, 1)       ///
            arrowopt(msize(1)))
```

We submitted the following variables; `tp tn fp fn sample`, of which the first four are the required diagnostic outcome variables and `sample` is a covariate. (`studyid(study)`) identifies the studies.

From the output below, there are 6 observations from 3 studies in the meta-analysis.

```
*************** Fitted model *********************************
    tp ~ binomial(logit(se), tp + fn)
    tn ~ binomial(logit(sp), tn + fp)
```

```
    logit(se) = mu_se + setting + study_se
    logit(sp) = mu_sp + setting + study_sp
    study_se, study_sp ~ biv.normal(0, sigma)
    Number of observations = 6
    Number of studies = 3
```

The second part of the output shows the assigned base levels for the categorical covariate used in the model. The base level in `setting` is `screening`.

```
   Variable -- Base Level
   sample -- clin
```

The order of the data is preserved and therefore, the levels in each of the categorical variable are decoded based on their order in the data, i.e first-come-first-assignment. The Stata command *gsort* sorts the data to alphabetically such that base level for `sample` is `clin` and the second level is `self`.

If data were not sorted, the option `alphasort` assigns the levels based on their alphabetical order (A to Z) i.e the first level alphabetically is assigned the base level while still preserving the order of the data.

The next part of the output below displays the representation of several simpler models fitted to the data for model comparison. The fitting of each model takes some time, and the time increases with the number of covariates. Save time by skipping this with the option `nomc`.

```
Just a moment - Fitting reduced models for comparisons
Ommitted : sample logit(se)
    logit(se) = mu_se + study_se
    logit(sp) = mu_sp + sample + study_sp
```

```
Ommitted : sample in logit(sp)
     logit(se) = mu_se + study_se
     logit(sp) = mu_sp + sample + study_sp
```

From the output above, two more models are fitted in addition to the specified model. The two simpler models leave out the `sample` effect term in each of the two predictor equations. If there were interaction terms present, one interaction term would be left out each time instead.

The option `noitable` suppressed the table with the study-specific estimates, and `sumtable(abs rr)` requested to display the absolute and relative specificity and relative sensitivity only (summary tables in the logit scale are suppressed).

The table below shows the absolute sensitivity and specificity of signal-based amplification HPV testing on self- and clinician samples in follow-up setting and overall:

```
*****************************************************************
     Marginal summary measures of test accuracy : Proportion
*****************************************************************
```

Parameter	Proportion	SE(logit)	z(logit)	P>|z|	Lower	Upper
--Sensitivity--						
sample						
clin	0.91	0.55	4.26	0.0000	0.78	0.97
self	0.84	0.51	3.23	0.0012	0.66	0.93
Overall	0.88	0.49	4.12	0.0000	0.74	0.95
--Specificity--						

```
----------+------------------------------------------------------------------
sample    |
     clin |       0.63       0.19      2.71    0.0067      0.54      0.71
     self |       0.57       0.19      1.49    0.1375      0.48      0.66
          |
  Overall |       0.60       0.18      2.28    0.0223      0.51      0.68
------------------------------------------------------------------------------
NOTE: H0: p = 0.5 vs. H1: P != 0.5
```

The next table (below) shows the comparison the relative sensitivity and specificity of signal-based amplification HPV testing on self- vs clinician-samples in follow-up setting and overall:

```
*******************************************************************************
           Marginal summary measures of test accuracy : Ratio
*******************************************************************************
-------------------------------------------------------------------------------
     Parameter | Rel Ratio   SE(lor)    z(lor)     P>|z|     Lower     Upper
---------------+---------------------------------------------------------------
 Relative Sensitivity
       Overall |      0.92      0.06     -1.39    0.1640      0.81      1.04
---------------+---------------------------------------------------------------
 Relative Specificity
       Overall |      0.91      0.06     -1.56    0.1186      0.81      1.02
-------------------------------------------------------------------------------
```

The overall relative sensitivity and relative specificity are based on the ratio of the average marginal predictions for `sample=clin` vs `sample=self`. For each level of `sample`, a prediction model is fitted where each study is treated as if all studies were in the same `sample` group.

The overall relative sensitivity and specificity was 0.92 [95% CI: 0.81, 1.04] and 0.91 [95% CI: 0.81, 1.02] respectively. This implies that signal-based self-sample and clinician sample tests were had similar sensitivity and specificity in the follow-up setting.

The next part of the output reports the several tests for heterogeneity. In meta-regression, the I^2 is not reported. From the output below, there is more heterogeneity in the logit sensitivities (`Tau.sq = 0.56`) than in the logit specificity (`Tau.sq = 0.06`). The generalized variability is even less (`Tau.sq = 0.03`) after accounting for the correlation (`rho = 0.13`) between the logits. Comparing the fixed-effects model, the random-effects meta-regression model is highly significant (p = 0.0018). The reported p-value is an upper bound of the actual p-value because the hypothesis test is on parameters on the boundary of the parameter space. Comparing the specified model to a model without covariates either on the linear predictor for logit sensitivity or logit specificity. The output indicates that the null model without covariate is more parsimonious (p-values are > 0.05 in both cases).

```
Between-study heterogeneity

                           rho
                           0.13
                           Tau.sq
Generalized                 0.03
Sensitivity                 0.56
Specificity                 0.06


                             Chi2        degrees of
                           statistic      freedom       p-val
LR Test: RE vs FE model      14.97          3          0.0018
```

```
LR Test: Full Model vs Intercept-only Model

Sensitivity                    2.79          1          0.1000

Specificity                    2.50          1          0.1100
```

The other three model are fitted by changing `if ta=="SA" & setting=="follow-up"` to include all the combinations of amplification method and setting.

In Figure 4, we observed differences in absolute specificity by sample, amplification and setting.

Confounding factors

Our interest is in answering whether a HPV test on a vaginal self-sample is as good as the test on a cervical sample taken by a clinician. The studies used different test amplification method and were conducted in different settings. The forest plots show that specificity consistently has low values in follow-up settings (range 50-63%) and substantially higher values in the screening setting (range 84-88%) suggesting that the absolute accuracy differed by settings.

To assess differences by `setting`, we entered the second (confounding factor) covariate `setting`. We also do the same for the amplification method `ta`. In `foptions()` we requested for a forest plot of the relative specificity and relative sensitivity with `outplot(rr)`. Generating such plots requires the option `comparative`, where the studies included applied the same test on a self-sample and a clinician-sample from the same women. The option `comparative` indicates that data is from studies that provide two rows; one for the `clin` and the second for the `self`. The the first covariate `sample` identify the two rows (`clin` or `self`). An SROC plot is not plotted with `outplot(rr)` option.

In Figure 5,we observe that the relative sensitivity of SA testing on self- vs clinician samples is consistently lower than unity irrespective of setting (top left). However, the relative sensitivity of TA testing on self- vs clinician samples consistently includes unity (bottom left) in screening and follow-up. Also the relative pooled specificities (at right) show limited variation by setting. These findings demonstrate that it is reasonable to report the "pooled" or the overall relative accuracy estimates without regard to setting.

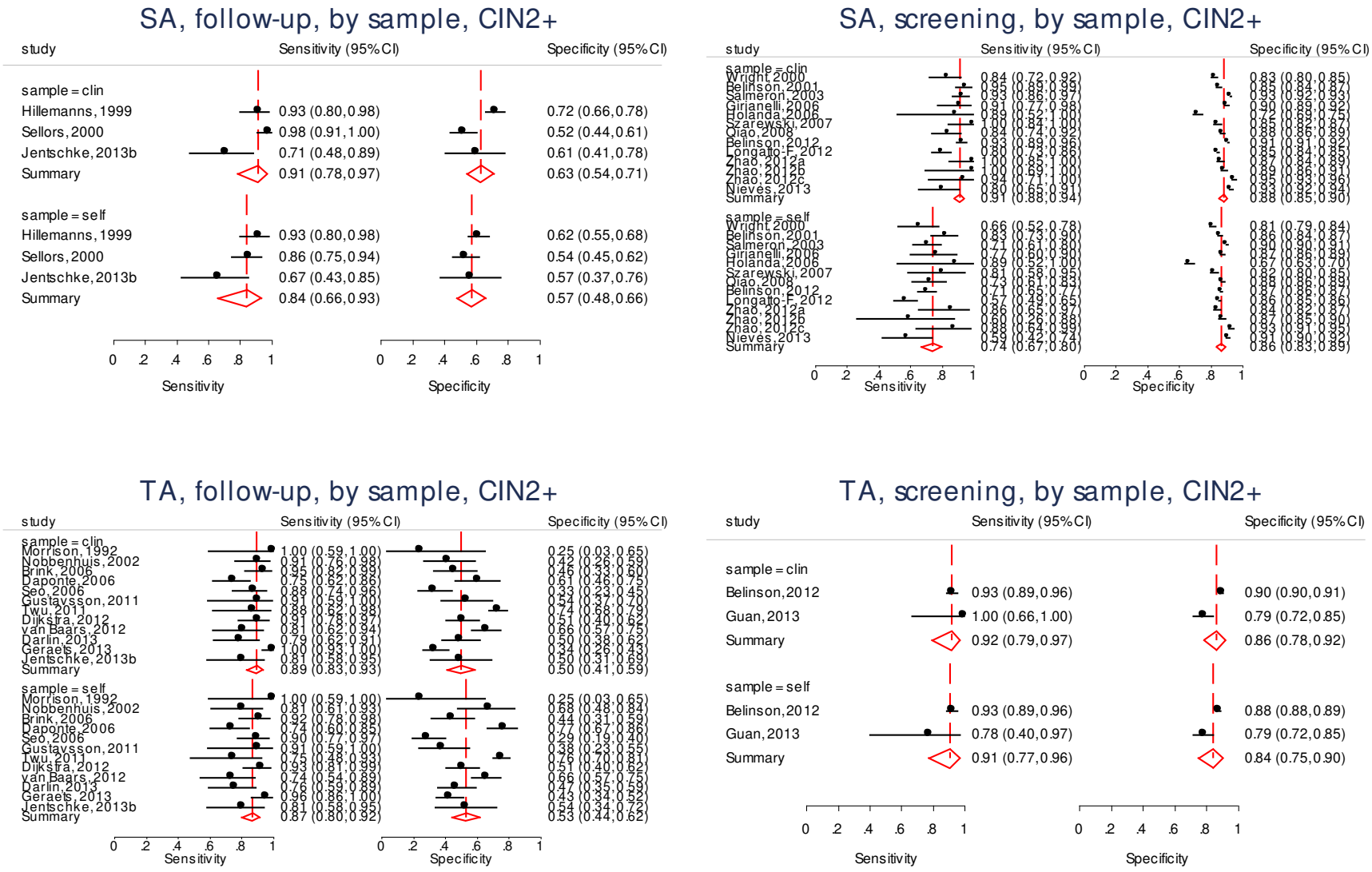


Figure 4. Absolute accuracy for CIN2+ of HPV testing on self-samples and on clinician-samples using signal amplification-based (SA) assays (top) or target amplification-based (TA) assays (bottom) in follow-up setting (left) and screening setting (right)

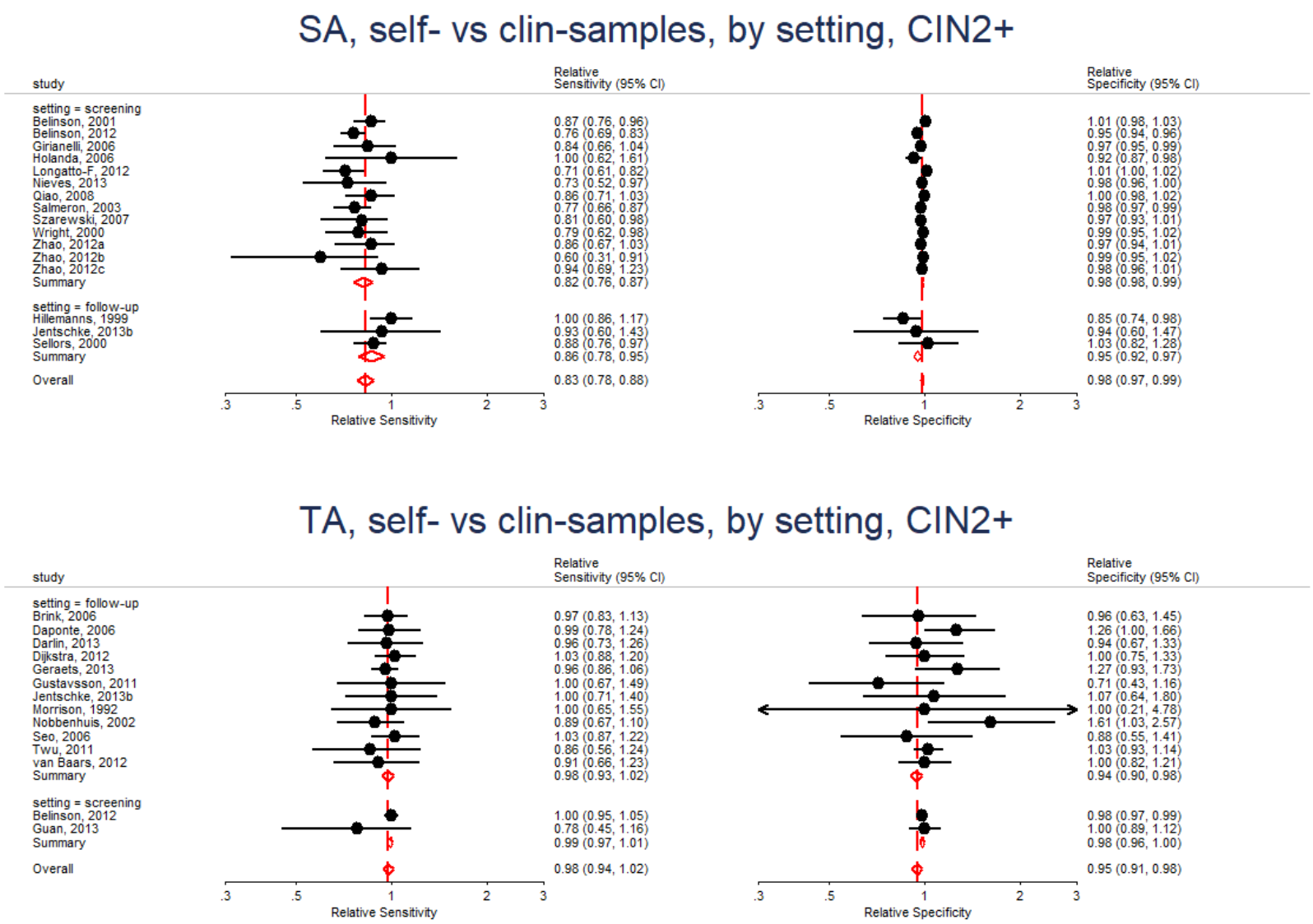


Figure 5. Relative sensitivity and specificity for CIN2+ of HPV testing on self- vs clinician-samples by setting with SA assays and (top) or TA assays (bottom).

The graphics in Figure 4 and Figure 5 suggest that both setting and amplification method significantly influences the absolute accuracy but that setting has minor influence on the relative accuracy.

Interaction variables

To examine the how `setting` and `ta` modify the relative sensitivity and specificity for CIN2+ of HPV testing on self- and clinician we used the option `interaction(sesp)` The interactions terms are placed on the both the linear predictor equations for logit sensitivity and specificity. `interaction(se)` or `interaction(sp)` would place the interaction terms in the linear predictor equation for logit sensitivity or logit specificity respectively. The interactions terms are between the first covariate `sample` and each of the rest of the covariates i.e `ta` and `setting`.

```
. sort ta setting study sample
. metadta tp fp fn tn sample ta setting,                          ///
    studyid(study) interaction(sesp)                              ///
    comparative noitable sumtable(rr)                             ///
    foptions(diamopt(color(red)) olineopt(color(red)              ///
           lpattern(dash)) outplot(rr)                            ///
           graphregion(color(white)) texts(1.6) lcol(setting)     ///
           xlabel(0.3, 0.5, 1, 2, 3) logscale arrowopt(msize(1)))
```

The dataset in the meta-analysis comprised of 60 observations from 28.

```
    Number of observations = 60
    Number of studies = 28
```

Controlling for the type of the test amplification, the relative sensitivity in the screening and follow-up groups were 0.91 [95% CI: 0.86, 0.95] and 0.92 [95% CI: 0.86, 0.98] respectively. The two confidence intervals over-lap suggesting that there might be little or no difference in the relative sensitivities. Similarly, controlling for the clinical setting, the relative sensitivity in test-amplification and signal-amplification tests were 0.82 [95% CI: 0.76, 0.89] and 0.98 [95% CI:0.95, 1.01] respectively. The confidence intervals do not overlap suggesting differences by test amplification method. The relative specificity can be interpreted in a similar manner.

A Wald-type test for non-linear hypothesis can be conducted using the model estimates to formally test whether the relative sensitivities and relative specificities are similar in all levels of the covariates. The results are displayed as follows;

```
**************************************************************************
Wald-type test for nonlinear hypothesis
    H0: All (log)RR equal vs. H1: Some (log)RR are different


------------------------------------------------------
            Parameter |     chi2        df         p
----------------------+-------------------------------
Relative Sensitivity  |
                   ta |    16.00         1    0.0001
              setting |     0.14         1    0.7049
----------------------+-------------------------------
Relative Specificity  |
                   ta |     0.16         1    0.6909
              setting |     0.50         1    0.4799
------------------------------------------------------
```

From the output above, even though there are two levels in each covariate, there one degree of freedom because there is only contrast, e.g. for `setting`, the contrasts is `RR.screening= RR.follow-up.`

Since the hypothesis test is conducted on the model-adjusted estimates, it is reasonable to conclude that after controlling for the type of test, the relative sensitivities were similar (p = 0.7049) across the two clinical settings. However, there were differences (p = 0.0001) in relative sensitivities between the two types of test after controlling for the clinical setting. There were no differences in relative specificities by clinical setting (p =0.4799) and type of test (p = 0.6909) after controlling for the type of test and the clinical setting respectively.

We also tested whether the interaction terms were significant by leaving one interaction term out in each of the predictor equation at a time. According to the output below, the two interaction terms in the predictor equation for logit specificity are not significant.

```
Leave-one-out LR Tests: Model comparisons
-----------------------------------------------
Excluded Effect |     chi2        df         p
----------------+------------------------------
Sensitivity     |
      ta*sample |     0.78         1    0.3800
 setting*sample |     9.95         1    0.0000
----------------+------------------------------
Specificity     |
      ta*sample |     3.56         1    0.0600
 setting*sample |     0.18         1    0.6700
-----------------------------------------------
```

Having established that the model without interaction terms in the predictor equation for logit specificity is not worse, we fitted the model with only `interaction(se)` option. The option instructs the program to fit interaction terms only in the predictor equation for logit specificity.

Before proceeding with the model fitting, we stored the model estimates for use later to compare the current model (`interaction(sesp)`) with the next model (`interaction(se)`).

First, we restored the model estimates by typing `estimates restore metadta_modest`. (The estimates of the fitted model are always stored as `metadta_modest`). Using the command `estimates store` we then stored the estimates under a different name, say full. i.e `estimates store full`.

The model with interaction only on the logit sensitivity predictor equation is as follows;

```
    logit(se) = mu_se + sample + ta + ta*sample + setting +
                setting*sample + study_se
    logit(sp) = mu_sp + sample + ta + setting + study_sp
```

We compared the new model to other simple models and the results are as follows;

```
Leave-one-out LR Tests: Model comparisons


-----------------------------------------------
Excluded Effect |     chi2        df         p
----------------+------------------------------
Sensitivity     |
      ta*sample |     0.78         1    0.3800
 setting*sample |     9.95         1    0.0000
```

```
---------------+--------------------------------
Specificity    |
             ta |    37.91         1    0.0000
        setting |     0.82         1    0.3700
-----------------------------------------------
```

From the output above, leaving out the `ta*sample` (p = 0. 3800) or `setting` (p = 0. 3700) from the linear predictor of the logit sensitivity and logit specificity would yield a more parsimonious model.

Model comparison

We use the following command to restore and store the model estimates with a different name;

```
estimates restore metadta_modest
estimates store reduced
```

The command `estat ic` displays AIC and BIC(12) of the current model. The output was;

```
------------------------------------------------------------------------------
 Model  |   Obs  ll(null)  ll(model)      df         AIC        BIC
--------+---------------------------------------------------------------------
reduced |   120    .        -440.8005      13    907.6009   943.8383
------------------------------------------------------------------------------
      Note: N=Obs used in calculating BIC; see [R] BIC note.
```

We restored the estimates from the previous model and requested for the AIC and BIC follows;

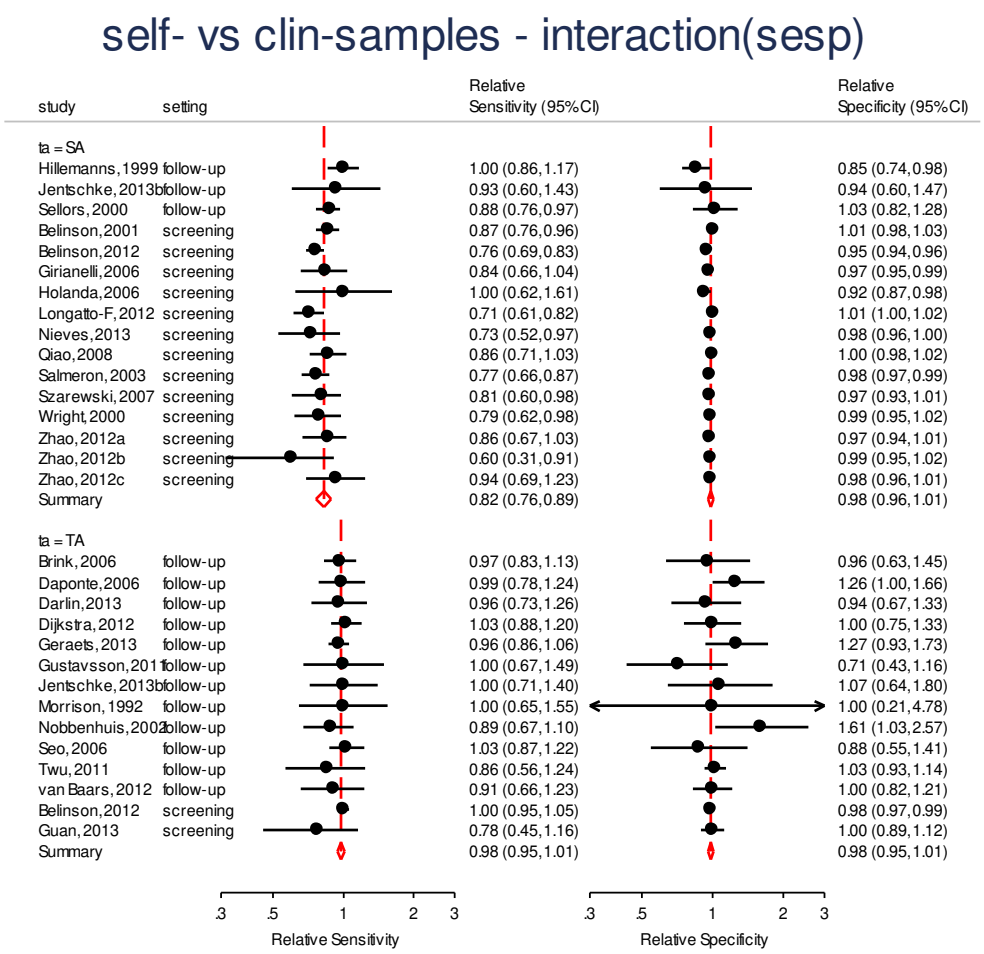
```
estimates restore full
estat ic
```

```
-----------------------------------------------------------------------------
Model |     Obs  ll(null)  ll(model)      df         AIC        BIC
------+----------------------------------------------------------------------
 full |     120    .        -439.0055      15    908.0111   949.8235
-----------------------------------------------------------------------------
     Note: N=Obs used in calculating BIC; see [R] BIC note.
```

Both the AIC and BIC conclude that the `reduced` is more parsimonious.

As mentioned earlier, the `reduced` model could be improved further by removing the more terms from the predictor equations. However, *metadta* makes interaction terms between the first covariate and the other covariates, the program cannot be used to fit the model without `ta*sample` while keeping `setting*sample`. It is also impossible to fit a model that includes the interactions terms on the logit sensitivity but leave out `setting` on the predictor equation of the logit specificity. Nonetheless, this "refined" model can be fitted via the native Stata command *meqrlogit.*

The forest plot from the `full`(on the left) and the `reduced` model (on the right) are displayed in Figure 6. The differences between the two plots is unnoticeable.



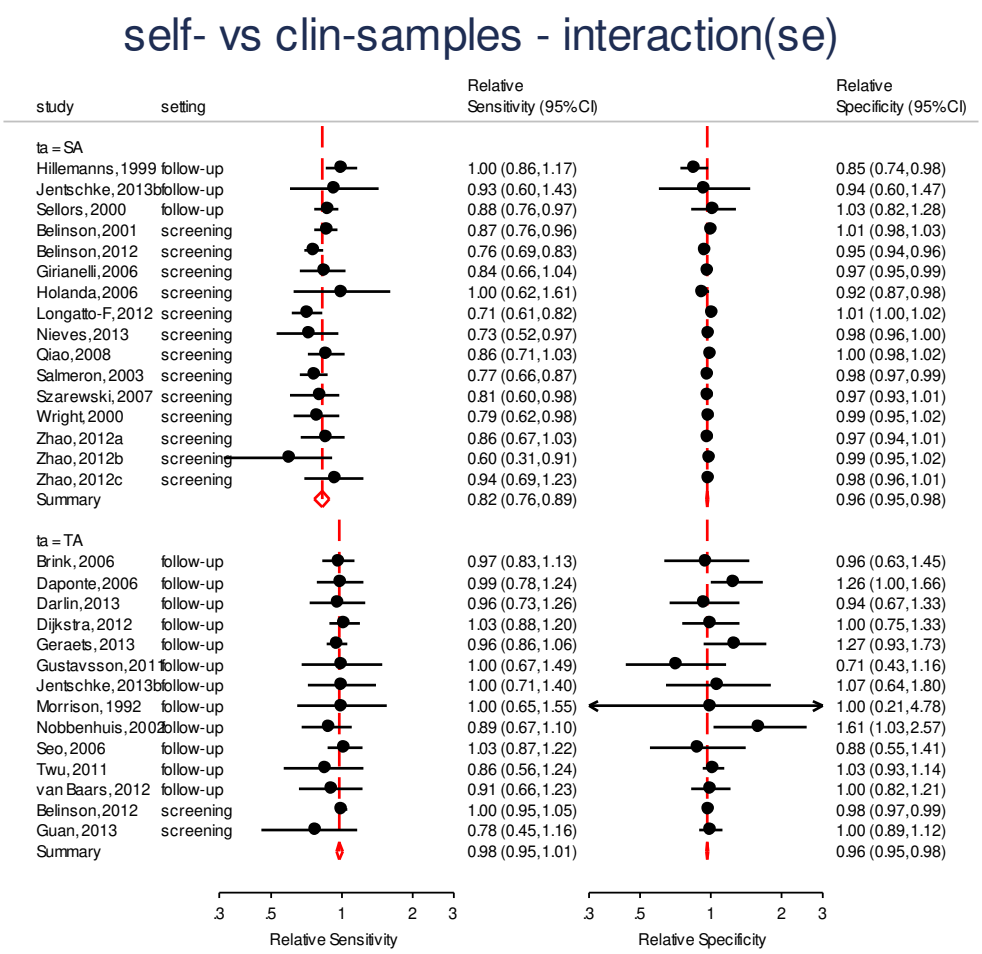


Figure 6. Relative sensitivity and specificity for CIN2+ of HPV testing on self- vs clinician-samples controlling for setting and test amplification method.

Discussion

This tutorial demonstrates how to perform meta-analysis and meta-regression of diagnostic test accuracy data in Stata using *metadta*. We developed m*etadta* to provide procedures specific for independent or comparative diagnostic test accuracy studies. Both fixed-effects and random-effects models with and without covariates were fitted using logistic regression. The model estimates are processed further to estimate model-adjusted summaries for absolute and relative sensitivity and specificity. The summaries are presented as tables or graphically in a forest and/or SROC plot. We demonstrated certain capabilities of *metadta* in this article using two published well-known meta-analyses. We encourage users of our program to explore the help file and run the demonstration examples, e.g for paired data where a two diagnostic tests are used on same patient to familiarize further with *metadta*.

For the random-intercept model, we present a bivariate heterogeneity statistic that takes into account the correlation between logit sensitivity and specificity and the sizes of the studies in the meta-analysis. *metadt*a also allows for model comparison with simpler leave-one-out models or the null models using tests likelihood ratio tests. With data from comparative studies, a hypothesis test is performed to test similarity in relative sensitivity and relative specificity between groups. The estimates of the specified model are stored and can be restored for further manipulation.

Compared to approximate methods that transform the date and use the normal distribution, logistic regression models binomial data appropriately while leaving the data intact. The binomial distribution is used to model the within-study variation appropriately without the need for continuity correction or excluding studies when sensitivity and/or specificity are zero or 100%.

Conclusion

metadta implements state of art statistical procedures in an attempt to close the gap between expert statisticians and end-users. With *metadta,* we hope to popularize even further, the use of appropriate statistical methods for meta-analysis of diagnostic test accuracy data.

List of Abbreviations

SROC - summary receiver operating characteristic

HPV - human Papilloma virus

DTA - diagnostic test accuracy

GLMM - generalized linear mixed models

HSROC- hierarchical summary receiver operation characteristic model

BRMA- bivariate random-effects meta-analysis

TP - true positive

FP - false positive

FN - false negative

TN - true negative

AIC - Akaike information criterion

BIC - Bayesian information criterion

CIN2+ - cervical intraepithelial neoplasia of grade 2 or worse

TA - signal-amplification

SA - test-amplification

Declaration

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The *metadta* program was developed in Stata 14.2. The code, the help files used herein are publicly available for download at https://ideas.repec.org/c/boc/bocode/s458794.html.

Competing interests: The authors declare that they have no competing interests.

Funding: Financial support was received from Horizon 2020 Framework Programme for Research and Innovation of the European Commission, through the RISCC Network (Grant No. 847845), the Belgian Cancer Centre, European Society of Gynaecological Oncology (ESGO), National Institute of Public Health and the Environment (Bilthoven, the Netherlands) and the German Guideline Program in Oncology (German Cancer Aid project)

Authors' contribution: VNN wrote the *metadta* program, analysed the data and drafted the manuscript. MA conceptualized the project, tested the program, and edited the manuscript.

Acknowledgements: Not applicable.

References

1. Breslow NE, Clayton DG. Approximate Inference in Generalized Linear Mixed Models. J Am Stat Assoc. 1993;88(421):9.

2. Chu H, Cole SR. Bivariate meta-analysis of sensitivity and specificity with sparse data: a generalized linear mixed model approach. J Clin Epidemiol. 2006;59(12):1331–2.

3. Rutter CM, Gatsonis CA. A hierarchical regression approach to meta-analysis of

diagnostic test accuracy evaluations. Stat Med. 2001;20(19):2865–84.

4. Zhou Y, Dendukuri N. Statistics for quantifying heterogeneity in univariate and bivariate meta-analyses of binary data: The case of meta-analyses of diagnostic accuracy. Stat Med. 2014;33(16):2701–17.

5. Harbord RM, Whiting P. Metandi: Meta-analysis of diagnostic accuracy using hierarchical logistic regression. Stata J. 2009;9(2):211–29.

6. Nyaga VN, Arbyn M. Advanced statistical methods for meta-analysis and meta-regression of diagnostic test accuracy data: a comparison of metadta with other statistical packages. Submitted. 2021;

7. Muller CJ, Maclehose RF. Estimating predicted probabilities from logistic regression: Different methods correspond to different target populations. Int J Epidemiol. 2014;43(3):962–70.

8. Higgins JPT, Thompson SG. Quantifying heterogeneity in a meta-analysis. Stat Med. 2002;21(11):1539–58.

9. Jackson D, White IR, Thompson SG. Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. Stat Med [Internet]. 2010 May 30 [cited 2020 Sep 11];29(12):1282–97. Available from: https://onlinelibrary.wiley.com/doi/full/10.1002/sim.3602

10. Koopman PAR. Confidence Intervals for the Ratio of Two Binomial Proportions. Biometrics. 1984;40(2):513.

11. Glas AS, Roos D, Deutekom M, Zwinderman AH, Bossuyt PMM, Kurth KH. Tumor markers in the diagnosis of primary bladder cancer. A systematic review. J Urol.

2003;169(6):1975–82.

12. Aho K, Derryberry D, Peterson T. Model selection for ecologists : the worldviews of AIC and BIC Author (s): Ken Aho , DeWayne Derryberry and Teri Peterson Stable URL : http://www.jstor.org/stable/43495189 Model selection for ecologi. Ecology [Internet]. 2014;95(3):631–6. Available from: http://www.esajournals.org/doi/full/10.1890/13-1452.1

13. Arbyn M, Verdoodt F, Snijders PJF, Verhoef VMJ, Suonio E, Dillner L, et al. Accuracy of human papillomavirus testing on self-collected versus clinician-collected samples: A meta-analysis. Lancet Oncol [Internet]. 2014;15(2):172–83. Available from: http://dx.doi.org/10.1016/S1470-2045(13)70570-9

14. Takwoingi Y, Deeks J. METADAS: A SAS macro for meta-analysis of diagnostic accuracy studies. Quick reference and worked example. https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/MetaDAS%20Quick%20Reference%20v13%20May%202012.pdf [Internet]. 2010;(July):Version 1.3. Available from: https://methods.cochrane.org/sites/methods.cochrane.org.sdt/files/public/uploads/MetaDAS Quick Reference v1.3 May 2012.pdf