CAN-19-240: Supplemental information – reader statistics

Kelsey M. Kennedy[1,10]
Renate Zilkens[1,2]
Wes M. Allen[1,3]
Ken Y. Foo[1,3]
Qi Fang[1,3]
Lixin Chin[1,3]
Rowan W. Sanderson[1,3]
James Anstie[1,3]
Philip Wijesinghe[1,3,11]
Andrea Curatolo[1,3,12]
Hsern Ern I. Tan[2]
Narelle Morin[4]
Bindu Kunjuraman[5]
Chris Yeomans[6]
Synn Lynn Chin[5]
Helen DeJong[1,7]
Katharine Giles[8]
Benjamin F. Dessauvagie[2,6]
Bruce Latham[6]
Christobel M. Saunders[2,5,9]
Brendan F. Kennedy[1,3]

[1] BRITElab, Harry Perkins Institute of Medical Research, QEII Medical Centre, Nedlands, and Centre for Medical Research, The University of Western Australia, Perth, Australia
[2] School of Medicine, The University of Western Australia, Perth, Australia
[3] Department of Electrical, Electronic & Computer Engineering, School of Engineering, The University of Western Australia, Perth, Australia
[4] Sonowest, Perth, Australia
[5] Breast Centre, Fiona Stanley Hospital, Murdoch, Australia
[6] PathWest, Fiona Stanley Hospital, Murdoch, Australia
[7] School of Medical and Health Sciences, Edith Cowan University, Joondalup, Australia
[8] OncoRes Medical, Perth, Australia
[9] Breast Clinic, Royal Perth Hospital, Perth, Australia
[10] Current affiliation - Department of Biomedical Engineering, Columbia University, New York, NY
[11] Current affiliation - School of Physics and Astronomy, University of St Andrews, United Kingdom
[12] Current affiliation - VioBio, Instituto de Óptica "Daza de Valdés", Consejo Superior de Investigaciones Cientificas (IO, CSIC), Madrid, Spain

Document generated using: `R -e "rmarkdown::render('CAN-19-1240_Supplemental_Reader_Stats.Rmd', 'all', encoding='UTF-8')"`

For formatting the output data tables, we use the `kable` function in the `knitr` library . For inter-reader variability, we use the `kappam.fleiss` function in the `irr` library, version 0.84, obtained from CRAN. For conditional formatting of table entries (highlighting p-values), we use the `format_table` function in the `formattable` library, version 0.2.0.1, obtained from CRAN.

```
library("knitr")
#install.packages("irr", lib = "./lib")
#library(lpSolve, lib.loc = "./lib")
#library(irr, lib.loc = "./lib")
source("./lib/kappam.fleiss.R")
source("./lib/print.irrlist.R")
#install.packages("formattable", lib = "./lib")
library(formattable, lib.loc = "./lib")
```

1 Raw reader data

Loaded from `.csv` files. Each with column "ROI_id", "OCT", "QME", "OCT_QME".

"Truth" contains "ROI_id", and "Involved", and is a record of the histopathology report for each sample.

"QME_Auto" contains "ROI_id" and "QME", and holds the results of a preliminary automated classification based on the QME elasticity values.

```
Truth <- read.csv(file = "reader-results-truth.csv", header = TRUE)

Eng1  <- read.csv(file = "reader-eng1.csv",  header = TRUE)
Eng2  <- read.csv(file = "reader-eng2.csv",  header = TRUE)
Surg1 <- read.csv(file = "reader-surg1.csv", header = TRUE)
Surg2 <- read.csv(file = "reader-surg2.csv", header = TRUE)
Path  <- read.csv(file = "reader-path.csv",  header = TRUE)
Res   <- read.csv(file = "reader-res.csv",   header = TRUE)
Sonog <- read.csv(file = "reader-sonog.csv", header = TRUE)

readers <- list(Eng1, Eng2, Surg1, Surg2, Path, Res, Sonog)
reader_names <- c("Eng 1", "Eng 2", "Surg 1", "Surg 2", "Path", "Res", "Sonog")

QME_Auto <- read.csv(file = "reader-auto.csv",  header = TRUE)
auto_readers <- list(QME_Auto)
auto_reader_names <- c("QME_Auto")
```

2 Methods

2.1 Confusion matrix

From the raw data, calculate the elements of the confusion matrix, True/False Positive/Negative.

```
tp_tn_fp_fn <- function(
  truth, readers, reader_names, class=c("OCT", "QME", "OCT_QME")
) {
  # Assemble the dataframe
  d_frame <- data.frame(
    # Names of the readers
    reader = reader_names,
    # True positives
    tp = sapply(readers,
        function(r) sum(truth$Involved == "Positive" & r[, class] == "Cancer")),
    # True negatives
    tn = sapply(readers,
        function(r) sum(truth$Involved == "Negative" & r[, class] == "Not cancer")),
```

```
    # False positives
    fp = sapply(readers,
        function(r) sum(truth$Involved == "Negative" & r[, class] == "Cancer")),
    # False negatives
    fn = sapply(readers,
        function(r) sum(truth$Involved == "Positive" & r[, class] == "Not cancer"))
  )
  # Calculates the aggregate data and returns the combined data frame.
  if (nrow(d_frame) > 1) {
    agg_frame <- data.frame(
      reader = c("Aggregate"),
      tp = sum(d_frame$tp), tn = sum(d_frame$tn),
      fp = sum(d_frame$fp), fn = sum(d_frame$fn)
    )
    all_frame <- rbind(d_frame, agg_frame)
  } else {
    all_frame <- d_frame
  }
  return(all_frame)
}
```

2.2 Confidence interval estimation

There are many ways to estimate confidence interval (1,2). In medical statistics, this is most commonly given by the "Wald" interval (2,3). For a sample probability $\hat{p} = m/n$, and $100(1-\alpha)\%$ confidence interval, the Wald interval is given by:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}. \tag{1}$$

```
#' Calculates and formats a probabilty and confidence interval for display as
#' percentages
wald_p_ci <- function(m, n, alpha=0.05) {
  p <- m / n
  Za2 <- qnorm(1 - alpha / 2)
  ci <- Za2 * sqrt(p * (1 - p) / n)
  disp_p_ci <- sprintf("%2.1f±%2.1f%%", p * 100, ci * 100)
  return(disp_p_ci)
}

#' Calculate the sensitivity, specificity, positive/negative predictive
#' values and overall accuracy of the diagnostic assessments, with
#' associated 95% configdence intervals.
diagnostic_acc <- function(tp_tn_fp_fn, p_ci_fun) {
  res <- data.frame(
    reader = tp_tn_fp_fn$reader,
    sens = p_ci_fun(tp_tn_fp_fn$tp, tp_tn_fp_fn$tp + tp_tn_fp_fn$fn),
    spec = p_ci_fun(tp_tn_fp_fn$tn, tp_tn_fp_fn$tn + tp_tn_fp_fn$fp),
    ppv = p_ci_fun(tp_tn_fp_fn$tp, tp_tn_fp_fn$tp + tp_tn_fp_fn$fp),
    npv = p_ci_fun(tp_tn_fp_fn$tn, tp_tn_fp_fn$tn + tp_tn_fp_fn$fn),
    acc = p_ci_fun(
      tp_tn_fp_fn$tp + tp_tn_fp_fn$tn,
      tp_tn_fp_fn$tp + tp_tn_fp_fn$tn + tp_tn_fp_fn$fp + tp_tn_fp_fn$fn)
  )
  return(res)
}
```

2.3 Inter-reader variability

There are several methods for measuring the variability between raters/readers of diagnostic data (4). For our study, we use Fleiss' kappa (5), which provides a measure for the agreement or disagreement of m readers assigning n subjects into k categories.

Kappa can be interpreted using the categories proposed by Landis and Koch (6):

Kappa	Interpretation
<0.00	Poor
0.00-0.20	Slight
0.21-0.40	Fair
0.41-0.60	Moderate
0.61-0.80	Substantial
0.81-1.00	Almost perfect

```
#' Generate a matrix with reader classifications in the columns, subjects in
#' the rows, of the classification results from reading using 'class' data,
#' e.g.,
#' | 1_OCT      | 2_OCT      | ... |
#' | ---------- | ---------  | --- |
#' | Cancer     | Not cancer | ... |
#' | Not cancer | Not cancer | ... |
#'
#' NOTE that 'truth' is only used here for merging the reader results,
#' as the inter-reader variability is concerned with consistency between
#' readers, not their accuracy
cat_matrix <- function(truth, readers, class=c("OCT", "QME", "OCT_QME")) {
  m <- truth[, "ROI_id", drop = FALSE]
  idx <- 1
  for (r in readers) {
    tmp <- r[, c("ROI_id", class)]
    # Rename e.g., "OCT" of "Eng1" -> "1_OCT"
    # "OCT" of "Surg1" -> "2_OCT" etc
    colnames(tmp) <- c("ROI_id", paste(idx, class, sep = "_"))
    idx <- idx + 1
    # Merge the data to ensure every reader's classification of, e.g., ROIxyz
    # is in the same row
    m <- merge(m, tmp, by = "ROI_id")
  }
  # Remove the ROI_id column, it's now implicit, and also not a reader
  m <- m[, !(names(m) %in% "ROI_id")]
  return(m)
}
```

2.4 Significance of OCT vs. QME, and OCT vs. OCT+QME

In order to establish the statistical significance of the QME, and OCT+QME, results as compared to the OCT results, we use McNemar's test. McNemar's test is a widely used tool for comparing the performance of two diagnostic tests in terms of sensitivities and specificities (7).

```
#' Use MacNemar's test to find the p-values for the data in confusion matrix 1
#' vs confusion matrix 2
diagnostic_significance <- function(CM1, CM2) {
```

```
  sens_p <- NULL
  spec_p <- NULL
  acc_p <- NULL
  for (r in CM1$reader) {
    # Construct the 2x2 contingency table for sensitivity, specificity,
    # and accuracy for both CM1 and CM2
    #
    # E.g.,
    # |    | CM1 | CM2 |
    # | tp |  15 |  24 |
    # | fn |   9 |   0 |
    #
    # Sensitivity: ratio between TP and FN
    sens_contin <- matrix(
      c(CM1[CM1$reader == r, "tp"], CM1[CM1$reader == r, "fn"],
        CM2[CM2$reader == r, "tp"], CM2[CM2$reader == r, "fn"]),
      nrow=2, ncol=2, dimnames=list(c("tp", "fn"), c("CM1", "CM2")))
    # Specificity: ratio between TN and FP
    spec_contin <- matrix(
      c(CM1[CM1$reader == r, "tn"], CM1[CM1$reader == r, "fp"],
        CM2[CM2$reader == r, "tn"], CM2[CM2$reader == r, "fp"]),
      nrow=2, ncol=2, dimnames=list(c("tn", "fp"), c("CM1", "CM2")))
    # Accuracy: Ratio between true and false
    cm1_true  <- CM1[CM1$reader == r, "tp"] + CM1[CM1$reader == r, "tn"]
    cm1_false <- CM1[CM1$reader == r, "fp"] + CM1[CM1$reader == r, "fn"]
    cm2_true  <- CM2[CM2$reader == r, "tp"] + CM2[CM2$reader == r, "tn"]
    cm2_false <- CM2[CM2$reader == r, "fp"] + CM2[CM2$reader == r, "fn"]
    acc_contin <- matrix( c(cm1_true, cm1_false, cm2_true, cm2_false),
      nrow=2, ncol=2, dimnames=list(c("t", "f"), c("CM1", "CM2")))
    # Calculate the corresponding p-values using Fischer's exact test
    sens_p <- append(sens_p, mcnemar.test(sens_contin)$p.value)
    spec_p <- append(spec_p, mcnemar.test(spec_contin)$p.value)
    acc_p  <- append(acc_p,  mcnemar.test(acc_contin)$p.value)
  }
  cm1_cm2_p_vals <- data.frame(
    reader = CM1$reader,
    sens_p = sens_p,
    spec_p = spec_p,
    acc_p  = acc_p
  )
  return(cm1_cm2_p_vals)
}

# Any field that fails the significance threshold of 0.05 is
# shown in bold text
p_val_sig_bold <- formatter(
  "span", x ~ ifelse(x <= 0.05, sprintf("%g", x), sprintf("**%g**", x))
)
```

3 Reader statistics

First, the number of positive/negative samples as reported from histopathology, along with an estimate of the percentage of involved (positive) samples.

```r
num_positive <- sum(Truth$Involved == "Positive")
num_negative <- sum(Truth$Involved == "Negative")
num_samples  <- num_positive + num_negative
samples <- data.frame(
  num_positive = num_positive,
  num_negative = num_negative,
  num_samples  = num_samples,
  prevalence = wald_p_ci(num_positive, num_samples)
)
kable(samples, row.names = FALSE)
```

num_positive	num_negative	num_samples	prevalence
24	130	154	15.6±5.7%

3.1 OCT only

Reader summary results using only the OCT images.

```r
oct <- tp_tn_fp_fn(Truth, readers, reader_names, "OCT")
```

3.1.1 Confidence intervals

```r
oct_acc <- diagnostic_acc(oct, p_ci_fun = wald_p_ci)
oct_res <- merge(oct, oct_acc, by = "reader", sort = FALSE)
kable(oct_res, row.names = FALSE)
```

reader	tp	tn	fp	fn	sens	spec	ppv	npv	acc
Eng 1	15	111	19	9	62.5±19.4%	85.4±6.1%	44.1±16.7%	92.5±4.7%	81.8±6.1%
Eng 2	17	113	17	7	70.8±18.2%	86.9±5.8%	50.0±16.8%	94.2±4.2%	84.4±5.7%
Surg 1	15	97	33	9	62.5±19.4%	74.6±7.5%	31.2±13.1%	91.5±5.3%	72.7±7.0%
Surg 2	16	94	36	8	66.7±18.9%	72.3±7.7%	30.8±12.5%	92.2±5.2%	71.4±7.1%
Path	16	109	21	8	66.7±18.9%	83.8±6.3%	43.2±16.0%	93.2±4.6%	81.2±6.2%
Res	19	97	33	5	79.2±16.2%	74.6±7.5%	36.5±13.1%	95.1±4.2%	75.3±6.8%
Sonog	18	98	32	6	75.0±17.3%	75.4±7.4%	36.0±13.3%	94.2±4.5%	75.3±6.8%
Aggregate	116	719	191	52	69.0±7.0%	79.0±2.6%	37.8±5.4%	93.3±1.8%	77.5±2.5%

3.1.2 Inter-reader variability

```r
oct_irr <- cat_matrix(Truth, readers, "OCT")
kappam.fleiss(oct_irr, detail = TRUE)
```

```
##  Fleiss' Kappa for m Raters
##
##  Subjects = 154
##    Raters = 7
##     Kappa = 0.487
##
##         z = 27.7
##   p-value = 0
##
##             Kappa      z p.value
```

```
## Cancer      0.487 27.688   0.000
## Not cancer  0.487 27.688   0.000
```

3.2 QME only

Summary data using only the elasticity (QME) results.

```
qme <- tp_tn_fp_fn(Truth, readers, reader_names, "QME")
```

3.2.1 Confidence intervals

```
qme_acc <- diagnostic_acc(qme, p_ci_fun = wald_p_ci)
qme_res <- merge(qme, qme_acc, by = "reader", sort = FALSE)
kable(qme_res, row.names = FALSE)
```

reader	tp	tn	fp	fn	sens	spec	ppv	npv	acc
Eng 1	24	127	3	0	100.0±0.0%	97.7±2.6%	88.9±11.9%	100.0±0.0%	98.1±2.2%
Eng 2	24	127	3	0	100.0±0.0%	97.7±2.6%	88.9±11.9%	100.0±0.0%	98.1±2.2%
Surg 1	18	115	15	6	75.0±17.3%	88.5±5.5%	54.5±17.0%	95.0±3.9%	86.4±5.4%
Surg 2	23	127	3	1	95.8±8.0%	97.7±2.6%	88.5±12.3%	99.2±1.5%	97.4±2.5%
Path	23	127	3	1	95.8±8.0%	97.7±2.6%	88.5±12.3%	99.2±1.5%	97.4±2.5%
Res	21	127	3	3	87.5±13.2%	97.7±2.6%	87.5±13.2%	97.7±2.6%	96.1±3.1%
Sonog	23	127	3	1	95.8±8.0%	97.7±2.6%	88.5±12.3%	99.2±1.5%	97.4±2.5%
Aggregate	156	877	33	12	92.9±3.9%	96.4±1.2%	82.5±5.4%	98.7±0.8%	95.8±1.2%

3.2.2 Inter-reader variability

```
qme_irr <- cat_matrix(Truth, readers, "QME")
kappam.fleiss(qme_irr, detail = TRUE)
```

```
##  Fleiss' Kappa for m Raters
##
##  Subjects = 154
##    Raters = 7
##     Kappa = 0.835
##
##         z = 47.5
##   p-value = 0
##
##             Kappa      z p.value
## Cancer      0.835 47.504   0.000
## Not cancer  0.835 47.504   0.000
```

3.2.3 Significance vs. OCT

```
oct_qme_p <- diagnostic_significance(oct, qme)
format_table(oct_qme_p, list(
  sens_p = p_val_sig_bold, spec_p = p_val_sig_bold, acc_p = p_val_sig_bold),
  format = "markdown")
```

reader	sens_p	spec_p	acc_p
Eng 1	0.0148061	8.33969e-19	7.60251e-20

reader	sens_p	spec_p	acc_p
Eng 2	0.00405714	1.053e-19	1.65512e-21
Surg 1	**0.123658**	2.77279e-11	1.02206e-11
Surg 2	0.011921	1.79763e-12	4.75187e-14
Path	0.011921	6.08462e-18	2.98619e-19
Res	0.00326372	1.94831e-13	1.32479e-15
Sonog	0.00296711	9.00829e-14	5.70263e-16
Aggregate	9.214e-13	1.49994e-97	4.14447e-108

3.2.4 Automated classification results

```
qme_auto <- tp_tn_fp_fn(Truth, auto_readers, auto_reader_names, "QME")
auto_acc <- diagnostic_acc(qme_auto, p_ci_fun = wald_p_ci)
auto_res <- merge(qme_auto, auto_acc, by = "reader", sort = FALSE)
kable(auto_res, row.names = FALSE)
```

reader	tp	tn	fp	fn	sens	spec	ppv	npv	acc
QME_Auto	24	127	3	0	100.0±0.0%	97.7±2.6%	88.9±11.9%	100.0±0.0%	98.1±2.2%

3.3 OCT + QME

Summary data using the combined OCT and elasticty (QME) results.

```
oct_qme <- tp_tn_fp_fn(Truth, readers, reader_names, "OCT_QME")
```

3.3.1 Confidence intervals

```
oct_qme_acc <- diagnostic_acc(oct_qme, p_ci_fun = wald_p_ci)
oct_qme_res <- merge(oct_qme, oct_qme_acc, by = "reader", sort = FALSE)
kable(oct_qme_res, row.names = FALSE)
```

reader	tp	tn	fp	fn	sens	spec	ppv	npv	acc
Eng 1	21	130	0	3	87.5±13.2%	100.0±0.0%	100.0±0.0%	97.7±2.5%	98.1±2.2%
Eng 2	21	130	0	3	87.5±13.2%	100.0±0.0%	100.0±0.0%	97.7±2.5%	98.1±2.2%
Surg 1	17	126	4	7	70.8±18.2%	96.9±3.0%	81.0±16.8%	94.7±3.8%	92.9±4.1%
Surg 2	23	130	0	1	95.8±8.0%	100.0±0.0%	100.0±0.0%	99.2±1.5%	99.4±1.3%
Path	17	130	0	7	70.8±18.2%	100.0±0.0%	100.0±0.0%	94.9±3.7%	95.5±3.3%
Res	16	129	1	8	66.7±18.9%	99.2±1.5%	94.1±11.2%	94.2±3.9%	94.2±3.7%
Sonog	20	130	0	4	83.3±14.9%	100.0±0.0%	100.0±0.0%	97.0±2.9%	97.4±2.5%
Aggregate	135	905	5	33	80.4±6.0%	99.5±0.5%	96.4±3.1%	96.5±1.2%	96.5±1.1%

3.3.2 Inter-reader variability

```
oct_qme_irr <- cat_matrix(Truth, readers, "OCT_QME")
kappam.fleiss(oct_qme_irr, detail = TRUE)
```

```
##  Fleiss' Kappa for m Raters
##
##  Subjects = 154
##    Raters = 7
```

```
##     Kappa = 0.814
##
##         z = 46.3
##   p-value = 0
##
##             Kappa      z p.value
## Cancer      0.814 46.287   0.000
## Not cancer  0.814 46.287   0.000
```

3.3.3 Significance vs. OCT

```
oct_oct_qme_p <- diagnostic_significance(oct, oct_qme)
format_table(oct_oct_qme_p, list(
  sens_p = p_val_sig_bold, spec_p = p_val_sig_bold, acc_p = p_val_sig_bold),
  format = "markdown")
```

reader	sens_p	spec_p	acc_p
Eng 1	0.0446097	2.03162e-19	7.60251e-20
Eng 2	0.0140193	2.52078e-20	1.65512e-21
Surg 1	**0.169811**	2.96285e-13	1.95049e-13
Surg 2	0.011921	5.26802e-13	1.41847e-14
Path	**0.109599**	1.51011e-18	1.15273e-18
Res	0.0290963	8.39952e-14	4.66069e-15
Sonog	0.0107874	2.5164e-14	5.70263e-16
Aggregate	2.01698e-09	7.01741e-103	2.06542e-109

References

1. Agresti A, Coull BA. Approximate is better than "exact" for interval estimation of binomial proportions. The American Statistician. 1998;52:119–26.

2. Brown LD, Cai TT, DasGupta A. Interval estimation for a binomial proportion. Statistical Science. 2001;16:101–33.

3. Hess AS, Shardell M, Johnson JK, Thom KA, Strassle P, Netzer G, et al. Methods and recommendations for evaluating and reporting a new diagnostic test. European Journal of Clinical Microbiology & Infectious Diseases. 2012;31:2111–6.

4. Gisev N, Bell JS, Chen TF. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. Research in Social and Administrative Pharmacy. 2013;9:330–8.

5. Fleiss JL. Measuring nominal scale agreement among many raters. Psychological Bulletin. 1971;76:378–82.

6. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33:159–74.

7. Kim S, Lee W. Does McNemar's test compare the sensitivities and specificities of two diagnostic tests? Statistical Methods in Medical Research. 2017;26:142–54.