Statistical Protocol: Clinical Comparison and Validation of Openly Available Deep Learning Methods for Automated Tumor Volume Delineation on PET-CT of Head and Neck Cancer

David Gergely Kovacs Petersen

05. September 2022

Contents

Context 1
R packages used in this protocol 1
Outcome measures 2
Implementation 2
Significance level and power 2
Handling of missing data 2
Selection of most promising deep learning method 2
Comparison to clinical performance 3
Lesion level agreement and Hausdorff distance 4
Appendix 1 5

Context

The procedure outlined in this protocol will be performed subsequently the selection and implementation of promising deep learning models identified in a litterature reivew.

This study of artificial intelligence methods is carried out in accordance with the CLAIM checklist. CLAIM is avalable at https://pubs.rsna.org/doi/epdf/10.1148/ryai.2020200029.

R packages used in this protocol

```
require(pwr)
```

```
## Loading required package: pwr
```

Documentation of pwr package is available here (https://cran.r-project.org/web/packages/pwr/pwr.pdf).

Outcome measures

Primary

Tumor delineation accuracy measured using the DICE-coefficient.

Secondary

1. Tumor delineation accuracy measured using Hausdorff distance.
2. Lesion-level detection accuracy using F1-score (harmonic mean of positive predictive value and sensitivity).

In all cases, the PET-CT scans were performed a few days prior to radiotherapy treatment.

Implementation

The used implementation of the DICE-coefficient, Hausdorff distance and lesion level detection metrics are available online (https://github.com/davidkvcs/rh-scripts-1/blob/master/rhscripts/metrics.py). For the Hausdorff distance the function hausdorff_distance_with_resampling will be used.

Significance level and power

Throughout our analyses a significance level of 0.05 and a power of 0.8 is be used.

Handling of missing data

Patients who do not have scans adhering to our standard PET-CT acquisition protocol are excluded. We manually evalute and correct all tumor delineations to ensure clinical segmentation quality in training, validation and test data. Cases where tumor delineation is not performed or incomplete, are excluded.

It can occur, that either the evaluated deep learning method or the participating physicians do not detect any PET-positive tumor volume in a scan. In this situation, the scan is evaluated as PET-negative, and there will be no delineated volume. This scenario will be handled in the following way:

If compared methods agree that a case is PET-negative, the following is registered: DICE-coefficient = 1. F1-score = 1. Hausdorff distance = 0.

If one method detects a volume, when the other does not, the following is registered: DICE-coefficient = 0. F1-score = 0. Hausdorff distance = NA (distance will be infinite).

Selection of most promising deep learning method

One-way analysis of variance (ANOVA) of model performance on the primary outcome measure will be used to analyse differences between methods. If no conclusions can be made based on ANOVA, the best method will be selected based on the secondary outcomes measures. If no conclusions can be made based on secondary outcome measures, the best method will be selected based on ease of implementation and prediction times.

Based on an early litterature review, we expect to choose seven methods for implementation, i.e. k=7 in the power calculation.

The implementation of pwr.anova.test is based on Cohen, J. (1988); Statistical power analysis for the behavioral sciences (2nd ed.); Hillsdale,NJ: Lawrence Erlbaum. An f-value at 0.1 corresponds to a small effect size according to Cohen 1988 ch. 8 section 8.2.3 "Small," "Medium," and "Large" f Values, page 284 in Cohen, J. (1977); Statistical power analysis for the behavioral sciences; Academic Press.

Based on previous experience we expect an error variance around 0.02, and we wish to be able to show differences in dice score of 0.025. This yields the following value for the effect size f:

```
round(sqrt((0.025^2)/0.02),2)
```

```
## [1] 0.18
```

I.e. an f-value of 0.18. Finally, we choose a more convervative value for f = 0.1 to increase the probability, that we can eventually distinguish competing methods, even if the variance of the tested methods turn out to be larger than expected.

Given these conditions, we solve for the population size n:

```
pwr.anova.test(f=0.1,k=7,sig.level=0.05,power = 0.8)
```

```
##
##      Balanced one-way analysis of variance power calculation
##
##               k = 7
##               n = 195.5339
##               f = 0.1
##       sig.level = 0.05
##           power = 0.8
##
## NOTE: n is number in each group
```

Hence we plan to include 196 patients to select the most promising method.

Comparison to clinical performance

A paired t-test will be used to compare deep learning and clinical performance

Based on experience we expect a standard devition in DICE-score measurements of 0.14. We would like to show a difference in DICE-score of 0.05. The study is paired and the alternative hypothesis is two-sided.

Given these conditions, we can solve for the population size n:

```
power.t.test(n = NULL, delta = 0.05, sd = 0.14, sig.level = 0.05,
             power = 0.8,
             type = c("paired"),
             alternative = c("two.sided"))
```

```
##
##      Paired t test power calculation
##
##               n = 63.48273
##           delta = 0.05
##              sd = 0.14
##       sig.level = 0.05
##           power = 0.8
##     alternative = two.sided
##
## NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs
```

Hence we include 64 patients to evaluate clinical performance.

These 64 patients are sampled as the last 64 scans in our recuitement period to increase the quality of the validation (See "Narrow validation", p 201, [1] A. Kleppe, O. J. Skrede, S. De Raedt, K. Liestøl, D. J. Kerr, and H. E. Danielsen, "Designing deep learning studies in cancer diagnostics," Nat. Rev. Cancer, vol. 21, no. 3, pp. 199–211, 2021, doi: 10.1038/s41568-020-00327-9.)

In the following text, nuclear medicine spcecialists are referred to as "physicians".

Delineating physicians are provided with the clinical indication for the scan prior to delineating the tumor volumes. Patients are anonymized, so the physicians cannot look them up in our clinical systems to gain additional information.

It is important to take into account, that the clinical delineations are performed by different physicians, resulting in a clinical interobserver variability. Hence, each patient is delineated according to standard clinical protocol by two randomly selected physicians selected from a group of 6 experienced physicians, who are used to performing tumour delineation as part of clinical routine. For each patient, it is randomly selected, which of the two physicians counts as the reference. The randomization is blocked, ensuring, that each physician performs an equal number of delineations to the extent possible. It is ensured throughout, that the same physcian does not see the same case twice. The randomization order and which randomly selected physician counts as the reference in each case is presented in detail in Appendix 1.

Each participating physician will act as reference (MD_ref) 10-11 times, distributed as follows:

```
table(d_rand$ref)
```

```
##
##  1  2  3  4  5  6
## 11 10 11 11 11 10
```

Each physician will be measured as the evaluted physician (MD_eva) 10-11 times, distributed as follows:

```
table(d_rand$md)
```

```
##
##  1  2  3  4  5  6
## 11 11 10 10 11 11
```

Physicians are allowed to confer with a radiologist or nuclear medicine specialist. It is ensured that the conferring physician has not previously seen the case in question.

Finally, deep learning (DL) is used to predict the PET-positive tumor volume in all cases. The DICE-coefficient is calculated for deep learning and the evaluated physician to the reference in each case (MD_eva vs. MD_ref and DL vs. MD_ref). The resulting values are compared using the paired t-test.

Lesion level agreement and Hausdorff distance

While the DICE-coefficient is the most commonly used similarity measure, it can fall short in detecting a lack of ability of a method to identify the entire malignant volume. For example, a method could miss a lymph node, which only represents a small percentage of reference volume. This situation could result in a good agreement based on the DICE-coefficient even though an important malignant lymph node was missed. To detect such issues, we use the secondary outcome metrics:

- Hausdorff distance
- F1-score (https://en.wikipedia.org/wiki/F-score)

The Hausdorff distance represents the distance between the reference and predicted volumes.

The F1-score expresses the quality of lesion detection within patients (harmonic mean of positive predictive value and sensitivity, i.e. whether we identify all ture positive volumes and whether we only detect true positive volumes).

These metrics will be reported in our analysis for completeness, and they will be used in cases, where we cannot draw conclusions based on the DICE-coefficient.

Appendix 1

Below is included the dataframe showing the randomization order and which random physicians will delineate each patient case. The order is randomized and it ensures, that each physician delinates approximately the same number of times, and that no physician sees the same patient twice.

```
colnames(d_rand) <- c("Patient number","MD_ref","MD_eva")
print(d_rand, row.names=FALSE)
```

```
##  Patient number MD_ref MD_eva
##               1      1      6
##               2      6      3
##               3      5      3
##               4      3      6
##               5      2      3
##               6      6      4
##               7      3      5
##               8      1      2
##               9      2      3
##              10      3      1
##              11      1      5
##              12      1      2
##              13      2      6
##              14      2      4
##              15      5      2
##              16      4      5
##              17      5      3
##              18      2      3
##              19      5      1
##              20      5      4
##              21      1      3
##              22      3      2
##              23      4      1
##              24      1      2
##              25      3      5
##              26      3      4
##              27      1      6
##              28      4      5
##              29      5      6
##              30      3      1
##              31      6      4
##               32      3      4
##               33      4      1
##               34      2      4
##               35      4      1
##               36      6      2
##               37      4      6
##               38      2      3
##               39      5      6
##               40      6      1
##               41      4      6
##               42      2      5
##               43      1      2
##               44      6      1
##               45      5      1
##               46      3      5
##               47      6      5
##               48      1      6
##               49      1      5
##               50      3      6
##               51      4      5
##               52      4      5
##               53      1      4
##               54      2      6
##               55      4      2
##               56      5      2
##               57      6      1
##               58      4      3
##               59      5      2
##               60      6      4
##               61      3      1
##               62      2      3
##               63      5      2
##               64      6      4
```