STARD Checklist for FDA Mitotic Counting Study Using eeDAP

Brandon D. Gallas

19 April 2017

Contents

TITLE, ABSTRACT, AND KEYWORDS

1. Identify the article as a study of diagnostic accuracy (recommend MeSH heading " sensitivity and specificity")

TBD

INTRODUCTION

Research questions or aims

2. State the research questions or aims, such as estimating diagnostic accuracy or comparing accuracy between tests or across participant groups

The aims of this study are to characterize the precision of mitotic counting from pathologists evaluating pre-specified fields of view (FOVs) on whole slide images (WSIs) and the corresponding FOVs on the microscope. We will compare the precision of these counts (and their differences) to the precision when the FOVs are not pre-specified, when the pathologists select their own FOVs. We believe that pre-specifying FOVs will

lead to more precise measurements and comparisons of WSI and microscope evaluations. The results from this study will support the qualification of eeDAP [Gallas2014_J-Med-Img_v1p037501] as a Medical Device Development Tool.

METHODS: Participants

Study Population

3. Describe the study population: the inclusion and exclusion criteria and the settings and locations where the data were collected.

4. Describe participant recruitment: was this based on presenting symptoms, results from previous tests, or the fact that the participants had received the index tests or the reference standard?

5. Describe participant sampling: was this a consecutive series of participants defined by selection criteria in items 3 and 4? If not, specify how participants were further selected.

6. Describe data collection: was data collection planned before the index tests and reference standard were performed (prospective study) or after (retrospective study)?

We will select 4 slides that have been evaluated during a previous mitotic counting study (mcsStudy1). The (H&E) slides contain specimens from patients (dogs) diagnosed with canine oral melanoma. The 4 slides will be selected based on the reader-average counts from mcsStudy1. We will select slides where the reader-average count ranged from just a few to many mitotic counts.

The protocol for mcsStudy1 followed clinical practice: the pathologists were asked to count the number of mitoses in 10 consecutive non-overlapping high-powered fields (HPF) starting in an area of high mitotic activity (Smedley et al. 2011). In mcsStudy1, the counts were collected using WSI images and glass slides. In the digital mode, the evaluation FOVs of each pathologist were recorded. This data will guide us in selecting 40 FOVs per slide that are appropriate for counting mitotic figures.

The FOVs that we will use in our study will be 200 um x 200 um, which is equivalent to 800 x 800 pixels of a WSI with a scanning resolution of 0.25 um/pixel. This size was selected to fit within most digital displays without panning or zooming.

While extracting or outlining the FOVs digitally is not challenging, we need to use a reticle in the microscope eyepiece to ouline the evaluation area for that mode.

METHODS: Test Methods

7. Describe the reference standards and its rationale.

8. Describe technical specifications of material and methods involved, including how and when measurements were taken, or cite references for a) index test or b) reference test

9. Describe definition of and rationale for the units, cut-off points, or categories of the results of the a) index test and b) reference standard.

10. Describe the number, training and expertise of the persons executing and reading the a) index tests and b) reference standards.

11. Were the readers of the a) index test and b) reference standards blind (masked) to the results of the other test? Describe any other clinical information available to the readers.

There is no reference standard for mitotic counting.

The pathologists will use eeDAP for data collection (evaluation environment for digital and analog pathology). eeDAP is a software and hardware platform for designing and executing digital and analog pathology studies where evaluation FOVs in the digital image are registered to the real-time view on the microscope (Gallas et al. 2014). As such, a study can be designed where pathologists are asked to evaluate a pre-selected list of evaluation FOVs in Digital mode and Microscope real-time mode (MicroRT mode). eeDAP collects the pathologist evaluations while cycling through the list of FOVs.

Details on the equipment we will use will be added later.

The locations where pathologists will evaluate the WSI images and glass slides has not yet been fully determined. However, we plan to collect some data at the Pathology Informatics Summit, the National Cancer Institute, and the Memorial Sloan Kettering Cancer Center.

The data to be collected are counts per evaluation area.

We will recruit 12 pathologists and will collect the counts from at least 4 readers per FOV. If readers evaluate the same case in both modes, there will be at least 10 days in between to eliminate image recall. Each reader will evaluate their case load independently. Each reader will only be given images and slides; they will not be given any other clinical information. Each reader will be given moderate training for indentifying mitoses.

METHODS: Statistical Methods

12. Describe methods for calculating or comparing methods of diagnostic accuracy and the statistical methods used to quantify uncertainty (e.g. 95% CI)

13. Describe methods for calculating test reproducibility, if done.

In the following, we consider different agreement metrics and different types of agreement. We will refer to scores from two readers ("reader.A" and "reader.B") each evaluating the same case or pair of cases ("case.1", "case.2"). For each type of agreement, we must specify if the readers are different or the same (between-reader or within-reader agreement) **AND** if reader.A and reader.B read with different modalities or the same modality (between-modality or within-modality agreement). For the situation where we have two modalities (WSI vs. microscope), we have five types of agreements possible:

1) One reader reads with the microscope twice.
    - *Within-reader, within microscope*
2) One reader reads with WSI twice.
    - *Within-reader, within WSI*
3) One reader reads using both WSI and the microscope.
    - *Within-reader, between modalities*
4) The two readers are different and they both use the microscope.
    - *Between-readers, within the microscope.*
5) The two readers are different and they both use WSI.
    - *Between-readers, within WSI*
6) The two readers are different. One uses WSI and the other uses the microscope.
    - *Between-readers, between modalities*

It is worth mentioning that if the goal is to measure one of the three within-reader agreements, each reader will have to evaluate the same cases twice. Best practices then require separate sessions with washout between them to mitigate against correlations between the evaluations that arise from memory. Therefore, if we want to evaluate all six types of agreement for every reader, it takes four reading sessions (two in each modality). This might be practially burdensome. Instead, we will only plan to evaluate the three between-reader types of agreement; this plan only requires one reading session per reader. If we are able to schedule two sessions for a reader, we can also evaluate within-reader, between-modality agreement.

There is a hierarchy in the types of agreement, and they are listed accordingly. Assuming that the microscope evaluation is the reference, within-reader and within-microscope agreement (type 1) is a baseline for all others.

Additionally, the within-reader types of agreement (types 1, 2, and 3) are baselines for the corresponding between-reader types agreement (types 4, 5, and 6). Finally, between-reader agreement within the microscope (type 4) is a baseline for the other two between-reader types of agreement (types 5 and 6).

We are not yet ready to fully specify the study design, including the size. However, we plan to include multiple readers. We hope to have each reader read in both modalities, but in order to recruit more readers, we may not make this a requirement. In the previous study (mcsStudy1), we found that the agreement between readers and between modalities was the same as the agreement within readers and between modalities. So there might not be a statistical benefit to using the same readers in both modalities. Therefore, the primary endpoints of the study will be the between-reader types of agreement. We will make the following comparisons:

- Within WSI agreement (type 5) vs. Within microscope agreement (type 4)
- Between modality agreement (type 6) vs. Within microscope agreement (type 4)

For each type of between-reader agreement, we will calculate it for each pair of readers and then average over all pairs. We will calculate the standard errors of the averages accounting for reader and case variability (Gallas et al. 2016), https://github.com/DIDSR/eeDAP/blob/master/000_docs/Gallas2016_Proc-SPIE_v9787p0F.pdf

Rank-based concordance

Rank-based concordance is a measure of ordinal prediction. It is very similar to Spearman's rank-based correlation in what it measures. The two measures are highly correlated with each other, but the two measures are not calculated the same way.

Rank-based concordance is a probability based on the scores (counts) from two readers evaluating the same cases (Kim 1971). It is best understood as the average of basic observational outcomes defined as follows.

There are five possible rank-based outcomes:

1) **Concordance**: Both readers score the same case higher than the other case.
2) **Discordance**: The readers score the cases in the opposite order.
3) **Tie in A only**: Reader A gives both cases the same score. Reader B gives the cases different scores.
4) **Tie in B only**: Reader B gives both cases the same score. Reader A gives the cases different scores.
5) **TIe in both A and B**: Reader A gives both cases the same score and so does reader B.

So the rank-based concordance is the average over all pairs of cases that are concordant. If the two readers evaluate N cases, there are N(N-1)/2 unique pairs of cases.

There is a similar probability for discordance and the three different types of ties. As a probability, the rank-based concordance ranges from zero to one. If there are no ties, then 1.0 indicates a perfect ordinal relationship between the two readers, 0.5 indicates there is no relationship between the two readers, and 0.0 indicates a perfectly awful relationship between the two readers. If there are ties, then concordance can only be as good as one minus the probability of ties. Concordance can be put on the same scale as correlation (2 x Concordance - 1), however this result is not equal to Spearman's rank-based correlation.

Differences between log-transformed counts: Mean and variance

The difference between counts for the same case is not a complicated quantity. However, it compliments concordance (or correlation) by characterizing the interchangeability of two sets of scores, which assumes they are given on a common scale. Our experience from mcsStudy1 leads us to believe that the WSI mode yields more counts. Therefore, calibration may be needed to equate counts from WSI and counts from the microscope.

Our experience from mcsStudy1 also taught us that the variability of the differences between counts appears to be proportional to the mean. This is common for measurements, especially counts. Consequently, instead

of differences between counts, we consider differences in the logarithms of the counts+1. We add one to the counts because the log of zero is undefined. Furthermore, since the difference log(x1) - log(x2) equals log(x1/x2), the difference plot is essentially showing the ratio between the scores, not a terribly complicated quantity.

Lastly, the distribution of the differences of the log-transformed counts+1 appeared to roughly follow a normal distribution. As such, these differences can be characterized by their mean and variance, or following Bland and Altman (Bland and Altman 1986), the mean and the limits of agreement (LA = 2 x standard deviation). We can invert the log to describe the mean and variance in terms of the ratio of the counts. Until we include some results from mcsStudy1, please refer to Fig 2 of Veta et al. to see Bland-Altman plots of the logarithm of mitotic count data (Veta et al. 2016).

RESULTS: Participants

14. Report when study was done, including beginning and ending dates of recruitment

15. Report clinical and demographic characteristics of the study population (e.g. age, sex, spectrum of presenting symptoms, co morbidity, current treatments, recruitment centres)

16. Report the number of participants satisfying the criteria for inclusion that did or did not undergo the index tests and/or the reference standard; describe why participants failed to receive either test.

RESULTS: Test Results

17. Report time interval from the index tests to the reference standard, and any treatment administered between.

18. Report distribution of severity of disease (define criteria) in those with the target condition; other diagnoses in participants without the target condition.

19. Report a cross tabulation of the results of the index tests (including indeterminate and missing results) by the results of the reference standard; for continuous results, the distribution of the test results by the results of the reference standard.

20. Report any adverse events form performing the index tests or the reference standard.

RESULTS: Estimates

21. Report estimates of diagnostic accuracy and measures of statistical uncertainty (e.g. 95% CI)

22. Report how indeterminate results, missing responses and outliers of the index tests were handled.

23. Report estimates of variability of diagnostic accuracy between subgroups of participants, readers or centres, if done.

24. Report estimates of test reproducibility, if done.

DISCUSSION

25. Discuss the clinical applicability of the study findings.

Bibliography

Bland, J M, and D G Altman. 1986. "Statistical Methods for Assessing Agreement Between Two Methods of Clinical Measurement." *Lancet* 1 (8476): 307–10.

Gallas, Brandon D., Amrita Anam, Weijie Chen, Adam Wunderlich, and Zhiwei Zhang. 2016. "MRMC Analysis of Agreement Studies." In *Proc. Spie*, 9787:97870F–97870F–12. doi:10.1117/12.2217074.

Gallas, Brandon D., Marios A. Gavrielides, Catherine Conway, Adam Ivansky, Tyler Keay, Wei-Chung Cheng, Jason Hipp, and Stephen M. Hewitt. 2014. "Evaluation Environment for Digital and Analog Pathology (EeDAP): A Platform for Validation Studies." *J Med Img* 1 (3): 037501. doi:10.1117/1.JMI.1.3.037501.

Kim, Jae-On. 1971. "Predictive Measures of Ordinal Association." *Am J Sociol* 76 (5): 891–907.

Smedley, R. C., W. L. Spangler, D. G. Esplin, B. E. Kitchell, P. J. Bergman, H-Y. Ho, I. L. Bergin, and M. Kiupel. 2011. "Prognostic Markers for Canine Melanocytic Neoplasms: A Comparative Review of the Literature and Goals for Future Investigation." *Vet Pathol* 48 (1). Diagnostic Center for Population; Animal Health, Michigan State University, East Lansing, Michigan, USA. Smedley@dcpah.msu.edu: 54–72. doi:10.1177/0300985810390717.

Veta, Mitko, Paul J van Diest, Mehdi Jiwa, Shaimaa Al-Janabi, and Josien PW Pluim. 2016. "Mitosis Counting in Breast Cancer: Object-Level Interobserver Agreement and Comparison to an Automatic Method." *PloS One* 11 (8). Public Library of Science: e0161286.